Fusion of Scores in a Detection Context Based on Alpha Integration

Antonio Soriano
ansoto@upvnet.upv.es
Instituto de Telecomunicaciones y Aplicaciones Multimedia
Universitat Politècnica de València
Camino de Vera S/n
46530 Valencia
Spain

Luis Vergara
lvergara@dcom.upv.es
Instituto de Telecomunicaciones y Aplicaciones Multimedia
Universitat Politècnica de València
Camino de Vera S/n
46530 Valencia
Spain

Bouziane Ahmed
bouah@doctor.upv.es
Department of Electrical Engineering
University of Mostaganem
27000 Mostaganem
Algeria

Addisson Salazar
asalazar@dcom.upv.es
Instituto de Telecomunicaciones y Aplicaciones Multimedia
Universitat Politècnica de València
Camino de Vera S/n
46530 Valencia
Spain

Abstract

We present a new method for fusing scores corresponding to different detectors (two hypotheses case). It is based on alpha integration, which we have adapted to the detection context. Three optimization methods are presented: least mean-square error, maximization of the area under the ROC curve and minimization of the probability of error. Gradient algorithms are proposed for the three methods.

Different experiments with simulated and real data are included in the paper. Simulated data consider the two-detector case to illustrate the different factors influencing alpha integration and to demonstrate the improvements obtained by score fusion, with respect to the individual detector performance. Two real data cases have been considered. In the first one, multimodal biometric data have been processed. This case is representative of scenarios in which probability of detection is to be maximized for a given probability of false alarm. The second case is the automatic analysis of electroencephalogram and electrocardiogram records with the aim of reproducing the medical expert detections of arousals during sleeping. This case is representative of scenarios in which probability of error is to be minimized. The general superior performance of alpha integration verifies the interest of optimizing the fusing parameters.

1 Introduction

There are many scenarios where multiple detectors are to be fused to improve their individual performance (Khaleghi, Khamis, Karray & Razavi, 2013; Atrey, Hossain, El Saddik & Kankanhalli 2010; Yuksel, Wilson & Gader, 2012; Kittler, Hatef, Duin & Matas 1998). In general, the input to one single detector is a vector of measures (observation or feature vector) which are processed to obtain a scalar statistic to be compared with a threshold, thus obtaining a binary decision. Then, fusion of detectors can be made at three different levels: measures, statistics or decisions. Finding optimum fusion functions becomes simpler as we go from

measures to decisions level, but a price is paid in loss of information. Therefore, fusion at the statistics (intermediate) level becomes a reasonable compromise. On one hand, the number of variables to be fused is reduced to the number of available detectors, on the other hand it avoids the loss of information after thresholding. Usually the statistic is called "score". Depending on the application, the score is normalized in a given range or not. Different normalization techniques exist (Jain, Nandakumar & Ross, 2005), which are especially interesting in the case that heterogeneous detectors are to be fused. Normalized scores between 0 and 1 may be though as estimates of the *a posteriori* probability assigned by the detector to one of the two hypothesis, if they are properly calibrated (Zadrozny & Elkan, 2002).

In this paper we concentrate on the fusion of scores for detection purposes. Moreover, we will make use of α-integration. This later was proposed to integrate stochastic models in (Amari, 2007). The particular case of integrating Gaussian Mixtures was considered in (Wu, 2009). Noteworthy, α-integration can be used to fuse or combine any finite set of d numbers $m_i \;\; i = 1 \ldots d \,,\; (m_i \geq 0)$ in the form

$$m_\alpha = f_\alpha^{-1}\left(\sum_{i=1}^{d} w_i \cdot f_\alpha(m_i)\right), \qquad f_\alpha(m) = \begin{cases} m^{\frac{1-\alpha}{2}} & ,\alpha \neq 1 \\ log(x) & ,\alpha = 1 \end{cases}$$

$$w_i \geq 0, \quad \textstyle\sum_{i=1}^{d} w_i = 1 \quad . \qquad (1)$$

It has been demonstrated in (Amari, 2007), that if m_i and m_α are respectively associated to probability density functions $m_\alpha(x)$ and $m_i(x)$ of some random variable x, then $m_\alpha(x)$ is the probability density minimi7zing the cost function

$$\mathcal{J}\big(m_\alpha(x)\big) = \sum_{i=1}^{d} w_i \cdot D\langle m_i(x) | m_\alpha(x)\rangle, \tag{2}$$

where $D\langle m_i(x)|m_\alpha(x)\rangle$ is the α-divergence (Amari, 2007; Wu, 2009) between the two probability densities. Particular simple cases of fusion rules are obtained for particular selections of the parameter α. Thus, assuming that $w_i = 1/d$, we see that $\alpha = -1, 1, 3$ respectively renders the arithmetic mean, the geometric mean and the harmonic mean. Similarly, $\alpha = \infty\,/-\infty$ is equivalent to compute the minimum/maximum. Notice that (2) can be applied to the approximation of every positive function $m_\alpha(x)$ from a set of d positive functions $m_i(x)\ i = 1 \ldots d$.

In (Choi, Choi, Katake & Choe, 2010) (Choi, Choi & Choe, 2013) the authors present gradient descent algorithms to estimate both the α parameter and the coefficients $\boldsymbol{w} = [w_1 \ldots w_d]^{\boldsymbol{T}}$ minimizing the mean square error (MSE) of the approximation achieved by α-integration in some target values $t_j\ j = 1 \ldots N$

$$\mathcal{E}(\alpha, \boldsymbol{w}) = \frac{1}{N}\sum_{j=1}^{N}\Big(t_j - m_\alpha\big(x_j\big)\Big)^2 \quad . \tag{3}$$

Expressions for the gradients are obtained and convergence is experimentally tested in some simulated data.

The α-integration can be readily adapted to the fusion of scores in a detection context. Several detectors will produce several scores which can be fused using α-integration to obtain a unique (fused) score. In this paper we propose three methods for estimating the fusing parameters (α and $w_1 \dots w_d$) given a set of labelled training data. The first one is appropriate in case of working with normalized scores and is a direct adaptation of the least mean-square error (LMSE) criterion in equation (3), to the detection problem. The possible unbalanced number of labeled data between both hypotheses, and the different cost incurred by every type of erroneous decision (detection miss or false alarm) are accounted by some simple modification of the cost function. A second method is proposed based on the maximization of the area under the ROC curve (AUCmax). This is a cost function well suited for the detection framework, and allows both normalized and non-normalized scores. These two methods are appropriate in applications where the probability of detection is to be maximized for a given probability of false alarm. However there are scenarios where minimizing the probability of error is more convenient. Hence we propose a third method (MPE) where the α-integration parameters are estimated so that the probability of taken wrong decisions is minimized. This method requires that the scores are normalized. Gradients algorithms are devised for the three methods.

The next section is devoted to the LMSE approach. Then, AUCmax is considered in Section 3. Some experiments with LMSE and AUCmax criteria based on simulations are presented in Section 4 with the aim of illustrating the concept and

the interest of the new methods of α-integration. Section 5 presents the application of LMSE and AUCmax to α-integration in biometric data. Finally, it is considered the MPE method in Section 6, which is applied in a medical diagnosis problem: automatic detection of arousals during sleeping. Minimizing the wrong detections (relative to a medical expert) is the essential objective in this applications. Conclusions ends the paper.

2 Estimating the α-integration parameters by LMSE criterion

In a detection scenario we must decide between two hypotheses H_0 and H_1. Let us assume that we have *d* different detectors working on the same hypotheses and that everyone contributes with a score s_i, in a manner that higher values of the score play in favor of selecting H_1 and viceversa. Let us also assume that the scores are normalized so that $0 \leq s_i \leq 1$. Apart from this, there are no other constraints. Thus, the specific way in which every detector computes its score is of no concern here. Similarly the detectors may share the same input of observations or have totally different inputs, they may be statistically independent or not, and so on.

What we want is a unique score s_α. Considering (1), the α-integration solution is given by

$$s_\alpha(\boldsymbol{s} = [s_1 \dots s_d]^T) = \begin{cases} \left(\sum_{i=1}^{d} w_i \cdot s_i^{\frac{1-\alpha}{2}}\right)^{\frac{2}{1-\alpha}} & , \quad \alpha \neq 1 \\ exp\left(\sum_{i=1}^{d} w_i \cdot log(s_i)\right) & , \quad \alpha = 1 \end{cases} \quad . \qquad (4)$$

Let us assume that sequences of labelled scores are available, i.e., we have a set of couples $\{ \boldsymbol{s}^j, y^j \}$ $j = 1 \dots N$ where $\boldsymbol{s}^j = \left[s_1^j \dots s_i^j \dots s_d^j\right]^{\boldsymbol{T}}$ is the vector of scores provided by the detectors, and y^j is the corresponding known binary decision ($y^j = 1$ if H_1 is true and $y^j = 0$ if H_0 is true). We may use this set to learn the parameters by minimizing a cost function as indicated in (3), which now becomes

$$\mathcal{E}(\alpha, \boldsymbol{w}) = \frac{1}{N} \sum_{j=1}^{N} \left(y^j - s_\alpha(\boldsymbol{s}^j)\right)^2 \qquad (5)$$

We see that by minimizing the cost function (5) we are trying to approximate the fused score to 1 when the true hypothesis is H_1 and to 0 when the true hypothesis is H_0 .

In many detection scenarios there is a significant unbalance between the sizes of the subsets of the training set $\{ \boldsymbol{s}^j, y^j \}$ $j = 1 \dots N$ corresponding to H_1 and H_0 .This is the case in novelty detection (Pimentel, Clifton, Clifton & Tarassenko 2014) or detection of signals in a noise background (Soriano, Vergara, Moragues,

& Miralles 2014). . In those cases minimization of (5) will be "blind" to H_1 . To account for this problem we propose a modification of the cost function (5). Let us call N_1 and N_0 the sizes of the subsets corresponding, respectively, to H_1 and H_0 , hence $N = N_1 + N_0$. Instead of minimizing the overall mean square error, we compute separately the mean square errors corresponding to H_1 and H_0 . Then the mean of both values is to be minimized. Taking advantage of the binary value of y^j, the new cost function can be expressed in the form

$$\mathcal{E}(\alpha, \boldsymbol{w}) = \frac{1}{2}\left\{\frac{1}{N_1}\sum_{j=1}^{N}\left(y^j - s_\alpha(\boldsymbol{s}^j)\right)^2 y^j + \frac{1}{N_0}\sum_{j=1}^{N}\left(y^j - s_\alpha(\boldsymbol{s}^j)\right)^2 (1-y^j)\right\} =$$

$$= \frac{1}{2}\sum_{j=1}^{N}\left(y^j - s_\alpha(\boldsymbol{s}^j)\right)^2 \left(\frac{y^j}{N_1} + \frac{(1-y^j)}{N_0}\right) \qquad (6)$$

In this manner the contributions to the error are normalized with respect to the size of the training subsets.

Similarly, we can consider the possibility of weighting the contributions of the different types of errors. These can be of two types: decide H_0 when the true hypothesis is H_1 (detection miss) or decide H_1 when the true hypothesis is H_0 (false alarm). A simple modification of (6) can consider this option

$$\mathcal{E}(\alpha, \beta, \boldsymbol{w}) = \frac{1}{2}\sum_{j=1}^{N}\left(y^j - s_\alpha(\boldsymbol{s}^j)\right)^2 \cdot \left(\beta\frac{y^j}{N_1} + (1-\beta)\frac{(1-y^j)}{N_0}\right) \quad 0 \le \beta \le 1\,. \qquad (7)$$

Notice that the modification of the new LMSE cost function of equation (7) implies a different weighting for every training sample contribution to the MSE as computed in (5). Also notice that N_0 and N_1 are the number of available training samples of each class, and β is a value fitted by the user depending on the importance given to every type of error. However, α and $\boldsymbol{w}$ can be estimated to minimize (7). In the following we present gradient algorithms to estimate the optimum value of α and $\boldsymbol{w}$.

We have to compute the derivatives of the error cost function in (7) with respect to α and w_i $i = 1..d$. Let us define

$$c_\beta^j = \left(\frac{y^j}{N_1}\beta + \frac{(1-y^j)}{N_0}(1-\beta)\right) \quad . \tag{8}$$

Then

$$\frac{\partial \varepsilon}{\partial \alpha} = -\sum_{j=1}^{N}\left(y^j - s_\alpha(\boldsymbol{s}^j)\right)\frac{\partial s_\alpha(\boldsymbol{s}^j)}{\partial \alpha}c_\beta^j \quad , \tag{9a}$$

$$\frac{\partial s_\alpha(\boldsymbol{s}^j)}{\partial \alpha} = \frac{2s_\alpha}{1-\alpha}\left(\frac{log\left(\sum_i w_i \cdot f_\alpha\left(s_i^j\right)\right)}{1-\alpha} + \frac{\sum_i w_i \cdot \frac{\partial f_\alpha\left(s_i^j\right)}{\partial \alpha}}{\sum_i w_i \cdot f_\alpha\left(s_i^j\right)}\right) \quad , \tag{9b}$$

$$\frac{\partial f_\alpha\left(s_i^j\right)}{\partial \alpha} = -\frac{1}{2}log\left(s_i^j\right)\cdot s_i^{\frac{1-\alpha}{2}} \quad . \tag{9c}$$

Moreover

$$\frac{\partial \varepsilon}{\partial w_i} = -2\sum_{j=1}^{T}\left(y^j - s_\alpha(\boldsymbol{s}^j)\right)\frac{\partial s_\alpha(\boldsymbol{s}^j)}{\partial w_i}c_\beta^j \quad , \tag{10a}$$

$$\frac{\partial s_\alpha(\boldsymbol{s}^j)}{\partial w_i} = \begin{cases} \frac{2}{1-\alpha}\left(\frac{s_\alpha \cdot f_\alpha(s_i^j)}{\sum_k w_k \cdot f_\alpha(s_k^j)}\right) & , \alpha \neq 1 \\ s_\alpha(\boldsymbol{s}^j) \cdot log(s_i^j) & , \alpha = 1 \end{cases} \quad . \qquad (10b)$$

Hence the corresponding gradient algorithms will be:

$$\alpha(l+1) = \alpha(l) - \eta_\alpha \frac{\partial \varepsilon}{\partial \alpha}(l) \quad , \qquad (11)$$

$$\boldsymbol{w}(l+1) = \boldsymbol{w}(l) - \eta_w \frac{\partial \varepsilon}{\partial \boldsymbol{w}}(l) \quad . \qquad (12)$$

Where $\frac{\partial \varepsilon}{\partial \alpha}(l)$ and $\frac{\partial \varepsilon}{\partial \boldsymbol{w}}(l)$ are obtained by respectively using (9) and (10), substituting α by $\alpha(l)$ and w_i by $w_i(l)$ where necessary. The values η_α and η_w are the learning rates constants which control the speed of convergence. In all the experiments included in this paper, these values have been fitted using similar values to those one recommended in (Choi, Choi, Katake & Choe, 2010) (Choi, Choi & Choe, 2013). Small variations around those values of η_α and η_w influenced the converging speed, but the final estimates remained the same.

3. Estimating the α-integration parameters by AUCmax

LMSE criterion minimizes the MSE, where the error is defined as the (weighted) difference between the final (integrated score) and the target value (1 for H_1, 0 for H_0). This seems a priori a reasonable criterion to obtain a good detector, but by no means implies that the probability of detection is maximized for a given probability of false alarm. Ultimately, the detector performance depends on the statistical distribution of the integrated scores under every hypothesis. This

suggests the convenience of a new criterion which could directly incorporate the detector performance.

Different figures of merit have been proposed to evaluate the detector performance (Parker, 2013). Among them the AUC is the most popular. Moreover, AUC has two advantages in comparison with MSE:

-we can optimize the fusion in specific intervals of the probability of false alarm depending on the application requirements.

-scores of the labeled training set are not required to be normalized between 0 and 1.

A ROC curve represents the probability of detection P_d as a function of the probability of false alarm P_f , let us represents this curve by the function $P_d(P_f)$. We can compute the area associated to that function in a given interval (γ_1, γ_2) of the independent variable P_f , by integrating $P_d(P_f)$, the result of the integral will be the AUC corresponding to that interval. Let us define a normalized AUC in a given interval:

$$nAUC_{\gamma_1}^{\gamma_2} = \frac{1}{\gamma_2 - \gamma_1} \int_{\gamma_1}^{\gamma_2} P_d(P_f) dP_f \tag{13}$$

Where $0 \leq \gamma_1 < 1$ and $\gamma_1 < \gamma_2 \leq 1$ limit the interval of interest where the normalized AUC is to be computed.

We must find the parameter set $v^* = \{\alpha,\ \boldsymbol{w} = [w_1 \dots w_d]^{\boldsymbol{T}}\}$ such that

$$v^* = \underset{v=\{\alpha,\boldsymbol{w}\}}{\operatorname{argmax}}\left(nAUC_{\gamma_1}^{\gamma_2}\right) \tag{14}$$

Under the constraints

$$0 \leq w_i \leq 1\,, \quad \sum_{i=1}^{d} w_i = 1 \tag{15}$$

In the following we propose a new method to solve this optimization problem. The cost function $\left(nAUC_{\gamma_1}^{\gamma_2}\right)$ will be obtained by means of the empirical non-parametric method recently proposed in (Narasimhan & Agarwal, 2013) for measuring the partial AUC, which was presented as an improvement of the one in (Dodd & Pepe 2003).

The training set $S = \{S_1, S_0\}$, consisting of N instances of score vectors $\boldsymbol{s} = [s_1 \dots s_d]^T$ and the corresponding fused score s_α, can be divided into two subsets corresponding to each hypothesis, H_1 or H_0:

$$\begin{aligned} S_1 &= \left\{\left\{\boldsymbol{s}_i \rightarrow s_{\alpha_i}^{H_1}\right\} \in H_1, \quad i = 1,..,N_1\right\} \\ S_0 &= \left\{\left\{\boldsymbol{s}_j \rightarrow s_{\alpha_j}^{H_0}\right\} \in H_0, \quad j = 1,..,N_0\right\} \end{aligned} \tag{16}$$

Let us name S_0^* the S_0 subset sorted in descending order of fused scores $s_{\alpha_j}^{H_0}$:

$$S_0^* = \left\{s_{\alpha_j}^{*H_0} > s_{\alpha_{j+1}}^{*H_0} \in H_0, j = 1,..,N_0\right\} \tag{17}$$

We can evaluate the normalized area $nAUC_{\gamma_1}^{\gamma_2}$ by numerical integration. This can be made by uniformly sampling the ROC curve, adding all the sample values and normalizing by the total number of samples. To define the sampling points, we take into account that the test is implemented by comparing score s_α with a threshold *t*. Therefore, every threshold establishes one point $P_d^t\left(P_f^t\right)$ of the ROC curve. We select consecutive values $s_{\alpha_j}^{*H_0}$ of the set S_0^* in a given interval as thresholds. For every threshold $t_j = s_{\alpha_j}^{*H_0}$ we count the number of values in S_1 which are above the threshold, this number divided by N_1 is an empirical estimate $P_d^{t_j}\left(P_f^{t_j}\right)$ for that threshold. Summing all values $P_d^{t_j}$ so obtained and dividing by the total number of summed values we obtain an empirical estimate of the normalized AUC in a given interval. The selected interval of thresholds in S_0^* must be in concordance with the P_f interval (γ_1, γ_2). But notice that as the elements in S_0^* are sorted in descending order, the thresholds $t_j = s_{\alpha_j}^{*H_0}$ correspond to empirical values $P_f^{t_j} = \frac{j}{N_0}$. Then the selected interval in S_0^* must be $\left(s_{\alpha_{j_{\gamma_1}}}^{*H_0}, s_{\alpha_{j_{\gamma_2}}}^{*H_0}\right)$ where $j_{\gamma_1} = \lceil N_0\gamma_1 \rceil$ is defined as the next higher whole number of value $N_0\gamma_1$ and $j_{\gamma_2} = \lfloor N_0\gamma_2 \rfloor$ as the next lesser whole number of value $N_0\gamma_2$. This leads to the empirical normalized AUC estimator of equations (18a) and (18b) below. We consider separately in (18a) the case in which the limits of the interval for integration are so close that, after truncations, the order of the limits is inverted, i.e., $j_{\gamma_1} > j_{\gamma_2}$. It that case only one sample $P_d^{t_j}$ for $t_j = s_{\alpha_{j_{\gamma_1}}}^{*H_0}$ is obtained for estimating the normalized AUC. The other cases, when $j_{\gamma_1} \leq j_{\gamma_2}$, are all included

in equation (18b). Notice that the truncation effects in j_{γ_1} and j_{γ_2} are compensated by the term a_1.

- If $j_{\gamma_1} > j_{\gamma_2}$:

$$n\widehat{AUC}_{\gamma_1}^{\gamma_2} = \frac{1}{N_1} \cdot \sum_{i=1}^{N_1} \mathbb{I}\left(s_{\alpha_i}^{H_1} > s_{\alpha_{j_{\gamma_1}}}^{*H_0}\right) \tag{18a}$$

- If $j_{\gamma_1} \leq j_{\gamma_2}$:

$$n\widehat{AUC}_{\gamma_1}^{\gamma_2} = \frac{1}{N_1 N_0 \cdot (\gamma_2 - \gamma_1)} \cdot \sum_{i=1}^{N_1} (a_1 + a_2)$$

$$a_1 = \left(j_{\gamma_1} - N_0\gamma_1\right) \cdot \mathbb{I}\left(s_{\alpha_i}^{H_1} > s_{\alpha_{j_{\gamma_1}}}^{*H_0}\right) + \left(N_0\gamma_2 - j_{\gamma_2}\right) \cdot \mathbb{I}\left(s_{\alpha_i}^{H_1} > s_{\alpha_{j_{\gamma_2}+1}}^{*H_0}\right) \tag{18b}$$

$$a_2 = \sum_{j=j_{\gamma_1}+1}^{j_{\gamma_2}} \mathbb{I}\left(s_{\alpha_i}^{H_1} > s_{\alpha_j}^{*H_0}\right)$$

$\mathbb{I}(\cdot)$ is a logic function which returns '1' when the relation evaluated is true and '0' otherwise. Defining the new variable $\varepsilon_{ij} = s_{\alpha_i}^{H_1} - s_{\alpha_j}^{*H_0} \in [0,1]$ a unit step function $\mathcal{U}(\cdot)$ can be used instead of the logic function $\mathbb{I}(\cdot)$:

$$\mathbb{I}\left(s_{\alpha_i}^{H_1} > s_{\alpha_j}^{*H_0}\right) = \mathcal{U}\left(\varepsilon_{ij} = s_{\alpha_i}^{H_1} - s_{\alpha_j}^{*H_0}\right), \qquad \mathcal{U}(x) = \begin{cases} 1 & , x \geq 0 \\ 0 & , x < 0 \end{cases} \tag{19}$$

In order to transform the empirical normalized AUC into a differentiable function, a continuous approximation of the unit step function must be carried out. A natural choice is the sigmoid function (Herschtal & Raskutti 2004):

$$\mathcal{U}\left(\varepsilon_{ij} = s_{\alpha_i}^{H_1} - s_{\alpha_j}^{*H_0}\right) \approx \theta_\delta(\varepsilon_{ij}) = \frac{1}{1 + e^{-\delta \cdot \varepsilon_{ij}}} \tag{20}$$

As it can be observed in figure 1, the sigmoid function $\theta_\delta(\cdot)$ may approximate the unit step function, with arbitrarily small approximation error, by selecting a large enough δ value.

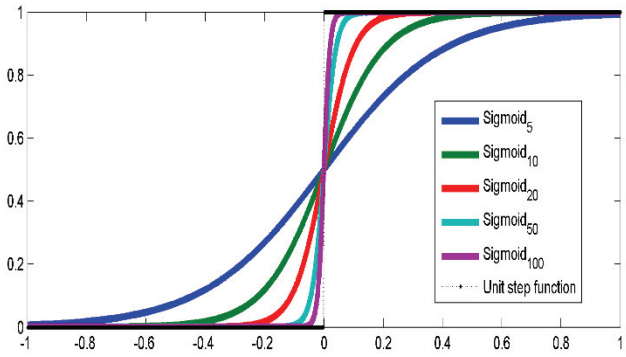

Figure 1. Approximating a unit step function using a sigmoid function.

Using the sigmoid function in (18) a constrained nonlinear minimization problem is stated:

$$\begin{aligned} &v^* = \underset{v=\{\alpha,\boldsymbol{w}\}}{\operatorname{argmin}}\left(g(\alpha,\boldsymbol{w}) = 1 - n\widehat{AUC}_{\gamma_1}^{\gamma_2}\right), \\ &0 \leq w_i \leq 1\,, \quad \sum_{i=1}^{d} w_i = 1 \end{aligned} \tag{21}$$

To solve this optimization problem an interior point algorithm can be used (Byrd, Hribar & Nocedal, 1999; Waltz, Morales, Nocedal, & Orban 2006).

The gradient of the objective function $\Delta g(\alpha, \boldsymbol{w}) = \left(\frac{\partial}{\partial \alpha} g(\alpha, \boldsymbol{w}),\ \frac{\partial}{\partial \boldsymbol{w}} g(\alpha, \boldsymbol{w})\right)$ can be obtained to improve the interior point algorithm due to the using of the differentiable sigmoid function in expressions (18a) and (18b). Differentiating the $\widehat{nAUC}_{\gamma_1}^{\gamma_2}$ estimator with respect to a generic parameter ν:

- If $j_{\gamma_1} > j_{\gamma_2}$:

$$\frac{\partial}{\partial \nu}\left(\widehat{nAUC}_{\gamma_1}^{\gamma_2}\right) = \frac{1}{N_1 N_0 \cdot (\gamma_2 - \gamma_1)} \cdot \sum_{i=1}^{N_1} N_0(\gamma_2 - \gamma_1) \cdot \frac{\partial \theta}{\partial \nu}\left(s_{\alpha_i}^{H_1} - s_{\alpha_{j_{\gamma_1}}}^{*H_0}\right) \qquad (22a)$$

- If If $j_{\gamma_1} \leq j_{\gamma_2}$:

$$\frac{\partial}{\partial \nu}\left(\widehat{nAUC}_{\gamma_1}^{\gamma_2}\right) = \frac{1}{N_1 N_0 \cdot (\gamma_2 - \gamma_1)} \cdot \sum_{i=1}^{N_1}\left(\frac{\partial a_1}{\partial \nu} + \frac{\partial a_2}{\partial \nu}\right)$$

$$\frac{\partial a_1}{\partial \nu} = \left(\left(j_{\gamma_1} - N_0\gamma_1\right)\frac{\partial \theta}{\partial \nu}\left(s_{\alpha_i}^{H_1} - s_{\alpha_{j_{\gamma_1}}}^{*H_0}\right)\right. \qquad (22b)$$

$$\left. + \left(N_0\gamma_2 - j_{\gamma_2}\right)\frac{\partial \theta}{\partial \nu}\left(s_{\alpha_i}^{H_1} - s_{\alpha_{j_{\gamma_2}+1}}^{*H_0}\right)\right)$$

$$\frac{\partial a_2}{\partial \nu} = \sum_{j=j_{\gamma_1}+1}^{j_{\gamma_2}} \frac{\partial \theta}{\partial \nu}\left(s_{\alpha} - s_{\alpha_j}^{*H_0}\right)$$

Continuing with the differentiation chain, the partial derivative of the sigmoid function must be obtained:

$$\frac{\partial\theta(\varepsilon_{ij})}{\partial v} = \delta \frac{e^{-\delta\cdot\varepsilon_{ij}}}{\left(1+e^{-\delta\cdot\varepsilon_{ij}}\right)^2}\frac{\partial\varepsilon_{ij}}{\partial v} \tag{23}$$

The partial derivative of the variable $\varepsilon_{ij} = s_{\alpha_i}^{H_1} - s_{\alpha_j}^{*H_0}$ depends on the partial derivative of the fused score $\frac{\partial s_\alpha}{\partial v}$, which is known (equations (9b) and (10b)).

$$\frac{\partial\varepsilon_{ij}}{\partial v} = \frac{\partial s_\alpha}{\partial v}\left(\boldsymbol{s}_i^{H_1}\right) - \frac{\partial s_\alpha}{\partial v}\left(\boldsymbol{s}_j^{*H_0}\right) \tag{24}$$

Substituting the generic parameter v by the parameters α and $w_i$in the differentiation chain of equations (22), (23), (24), (9b), (10b), the gradient of the objective function can be obtained straightforwardly.

4. Experiments with simulated data

We have performed a number of simulations with the aim of illustrating the different factors influencing α-integration for the fusion of detectors, as well as the specific interest of the proposed modifications. We have considered the fusion of two detectors (d=2). Every detector provides one score $s_i \;\; i = 1,2$ which is modelled as a random variable uniformly distributed in a given interval which depends on the true hypothesis $H_k \;\; k = 0,1$. Let us respectively call l_{ik}^l and l_{ik}^u to the lower and upper limits of the intervals corresponding to the uniform distribution of the scores provided by detector i under hypothesis H_k .

We show in figures 2 to 10 the results of 9 experiments. In experiments 1 to 6, the LMSE gradient algorithm was used to estimate the optimum value of α and/or w_1, w_2. However, in experiments 7 to 9, the AUCmax was considered.

Every figure is formed by 6 subfigures showing (from left to right and from top to bottom):

-The 2-D distribution of the training set of scores.

-The curves of convergence of the parameter α and/or the coefficients w_1 and w_2 corresponding to the gradient algorithm of equations (11a) and (11b).

-The ROC curves of the three detectors (two individual detectors and the fused one) representing the probability of detection P_d in terms of the probability of false alarm P_f.

-The 2-D contour curves defining the decision regions of the α-integrated detector.

-The uniform distributions of the scores s_1 and s_2 corresponding to every individual detector.

-The final distributions of the score s_α obtained after α-integration

In all the experiments, the training (estimation of the optimum value of α and/or w_1, w_2) was made by using N=5000 labelled scores. The evaluation performance (ROC curves and fused score distributions) was obtained from a set of 10000 scores. Other experiments were made by using different training and evaluation sizes, but the general conclusions remained the same.

Figure 2 corresponds to the experiment 1. As we see the parameter α is learned by means of the gradient algorithm (11) and converges to a final value (0.6) after only some 15 iterations. The sizes of the training sets are the same for both hypotheses $N_0 = N_1 = 0.5N$. The weigthing coefficients are not estimated but fitted to the same value ($w_1 = w_2 = 0.5$). The parameter $\beta = 0.5$, i.e., no preference is given *a priori* to any hypothesis. The limits of the uniform distributions of the individual scores are $l_{10}^l = l_{20}^l = 0$, $l_{10}^u = l_{20}^u = 0.8$, $l_{11}^l = l_{21}^l = 0.2$, $l_{11}^u = l_{21}^u = 0.8$ (l_{dh}^x, $x = \{l: lower, u: upper\}, d = \{1: detector\ 1, 2: detector\ 2\}\ and\ h = \{0: H_0, 1: H_1\}$). This implies a large overlap between both hypotheses when working separately with the individual detectors. However the distributions of the integrated score s_α are no longer uniform, showing the better separation between hypotheses achieved after α-integration. This can also be observed by looking to the ROC curves.

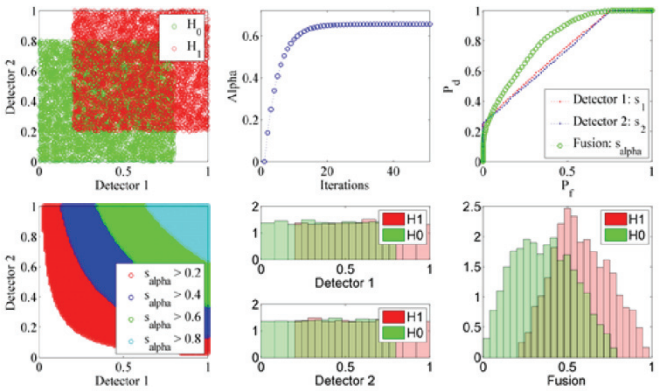

Figure 2. Experiment 1: $N_0 = N_1 = 0.5N$, $w_1 = w_2 = 0.5$, $\beta = 0.5$, $l_{10}^l = l_{20}^l = 0$, $l_{10}^u = l_{20}^u = 0.8$, $l_{11}^l = l_{21}^l = 0.2$, $l_{11}^u = l_{21}^u = 1$.

The experiment 2 (figure 3) illustrates the interest of the modification included in (6) to account for the possible different sizes of the training sets under every hypothesis. Thus in figure 2 the sizes of the training set under every hypothesis are very different ($N_0 = 0.2N \;\; N_1 = 0.8N$). The rest of the parameters are the same than those of the first experiment. We can see that the parameter α converges to the same value of experiment 1, hence the performance of the detector after fusion should be the same. This is verified by observing that the ROC curves, the 2D contours and the distribution of the integrated score s_α are practically the same in both experiments.

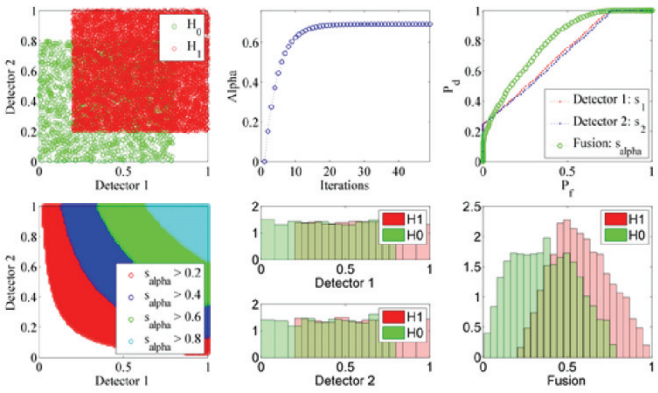

Figure 3. Experiment 2: $N_0 = 0.2N \;\; N_1 = 0.8N, \;\; w_1 = w_2 = 0.5, \;\; \beta = 0.5,$ $l_{10}^l = l_{20}^l = 0, \; l_{10}^u = l_{20}^u = 0.8, \; l_{11}^l = l_{21}^l = 0.2, \; l_{11}^u = l_{21}^u = 1.$

The next two experiments illustrate how parameter β may be used to bias the α-integrated detector towards one of the two hypotheses. Thus is figure 4 we show the same case than in figure 1, except that now $\beta = 0.9$, so that the contribution to the global error due to deciding H_0 when the true hypothesis is H_1 is much more significant than vice versa. We see in figure 4 that α converges to $-\infty$, i.e.

the α-integrated detector tends towards computing the maximum of the two individual scores, which clearly bias the decisions in favor of H_1 . This bias can also be observed in the form adopted by the 2D contour curves defining the decision regions, and in the resulting distributions of s_α . Finally we see in the ROC curves that for a probability of false alarm greater than ~0.6 the individual detectors have grater probability of detection than the α-integrated detector

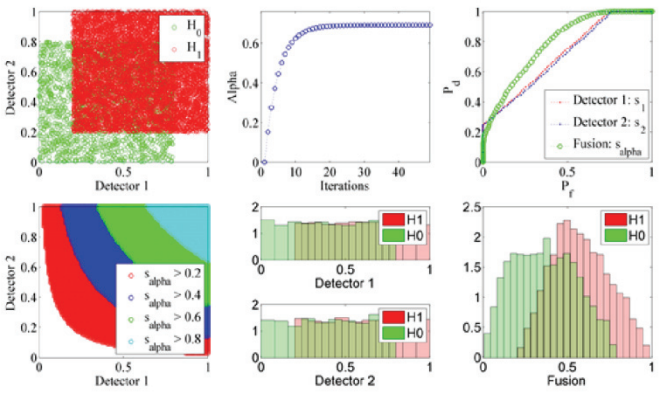

Figure 4. Experiment 3: $N_0 = N_1 = 0.5N$, $w_1 = w_2 = 0.5$, $\beta = 0.9$, $l_{10}^l = l_{20}^l = 0$, $l_{10}^u = l_{20}^u = 0.8$, $l_{11}^l = l_{21}^l = 0.2$, $l_{11}^u = l_{21}^u = 1$.

Experiment 4 is similar to experiment 3, but now $\beta = 0.1$, so that the fusion of detectors is biased in favor of H_0. We see in figure 5 that α converges to ∞, i.e. the fusion tends towards computing the minimum of the two individual scores . The 2D contour curves, and the resulting distributions of s_α are modified accordingly. Finally we see in the ROC curves that for a probability of false alarm

less than ~0.1 the individual detectors have greater probability of detection than the fused detector.

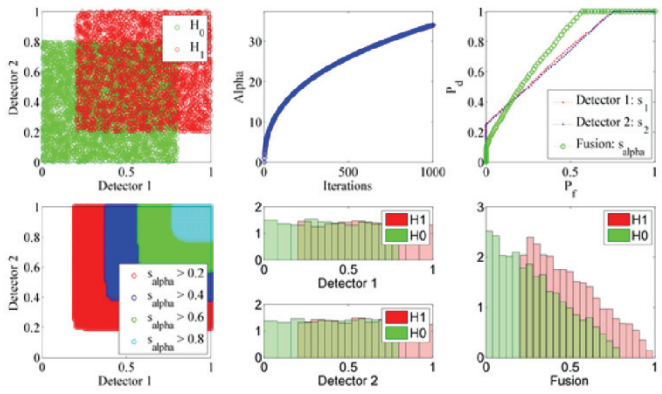

Figure 5. Experiment 4: $N_0 = N_1 = 0.5N$, $w_1 = w_2 = 0.5$, $\beta = 0.1$, $l_{10}^l = l_{20}^l = 0$, $l_{10}^u = l_{20}^u = 0.8$, $l_{11}^l = l_{21}^l = 0.2$, $l_{11}^u = l_{21}^u = 1$.

The next experiment illustrate that an optimum linear combiner (weighted arithmetic mean) of the individual scores is a particular constrained case of α-integration. Notice in (4) that if $\alpha = -1$ then $s_\alpha = \sum_{i=1}^{d} w_i \cdot s_i$.

We show in figure 6 the results of the experiment 5, which is the same case of experiment 1, except that $\alpha = -1$ and the weighting coefficients are learned by the gradient algorithm (11b). As both individual detectors produce the same score distribution (both detectors perform the same), the gradient algorithm converges to $w_1 = w_2 = 0.5$. Notice that the contour curves are now straight lines in concordance with (12). This implies some suboptimality with respect to the experiment 1, where the optimum α was learned by the gradient algorithm, and was different from -1. Suboptimality can be appreciated too by comparing the

ROC curve of the α-integrated detector in figure 6 with the corresponding of figure 2.

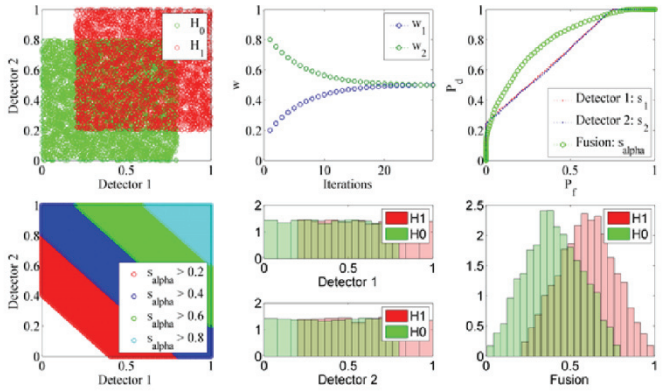

Figure 6. Experiment 5: $N_0 = N_1 = 0.5N$ $,\alpha = -1,$ $\beta = 0.5,$ $l_{10}^l = l_{20}^l = 0,$ $l_{10}^u = l_{20}^u = 0.8,$ $l_{11}^l = l_{21}^l = 0.2,$ $l_{11}^u = l_{21}^u = 1.$

The experiment 6 illustrates the case of combining two detectors having different performances. We have modified the experiment 5, so that the uniform distribution of the scores of the detector 1 under H_1 is narrowed (between 0.4 and 1). This implies that the detector 1 performs better than the detector 2 under H_1. Then we see in figure 7 that the gradient algorithm converges to weights such that $w_1 > w_2$, so that the detector 1 has more influence in the final α-integrated detector. This produced a rotation of the 2D contours to accommodate a bias towards the decisions of the detector 1.

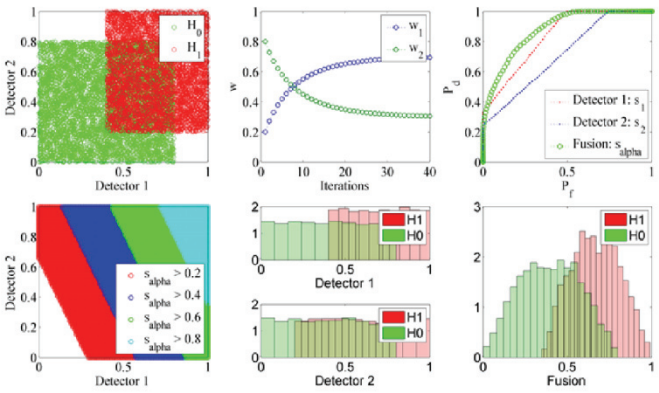

Figure 7. Experiment 6: $N_0 = N_1 = 0.5N$, $\alpha = -1$, $\beta = 0.5$, $l_{10}^l = l_{20}^l = 0$, $l_{10}^u = l_{20}^u = 0.8$, $l_{11}^l = 0.4$ $l_{21}^l = 0.2$, $l_{11}^u = l_{21}^u = 1$.

In the next three experiments we are going to use the parameter estimation method based on AUCmax. Using this method we can select the interval of probabilities of false alarm in which we want to obtain the best results. These experiments are like experiment 1, where two equal detectors are fused by means of α integration, but now the new training method based on the AUCmax is used. Firstly (experiment 7) we have estimated all the parameters to maximize $nAUC_0^1$. The results are represented in figure 8. In this case we can see how the weighting parameters obtained are the same for each detector $w_1 = w_2 = 0.5$ and the estimated α parameter converges to a value so that the whole AUC of the ROC curve obtained after fusion, is maximized. Notice that the ROC curves are quite similar to the ones in Figure 1.

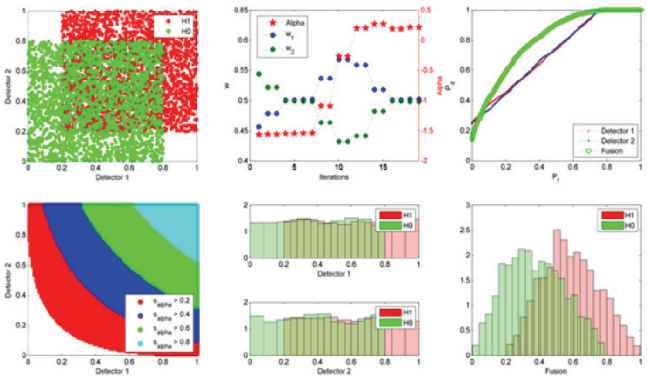

Figure 8. Experiment 7: $N_0 = N_1 = 0.5N$, $w_1 = w_2 = 0.5$, $\beta = 0.5$, $l_{10}^l = l_{20}^l = 0$, $l_{10}^u = l_{20}^u = 0.8$, $l_{11}^l = l_{21}^l = 0.2$, $l_{11}^u = l_{21}^u = 1$. Parameters are optimized by AUCmax ($nAUC_0^1$).

In the two final experiments we have changed the P_f interval of the ROC curves in which we want to maximize the AUC. Thus, in experiment 8 (figure 9) $nAUC_0^{0.1}$ is maximized, and in experiment 9 (figure 10) $nAUC_{0.6}^1$ is maximized.

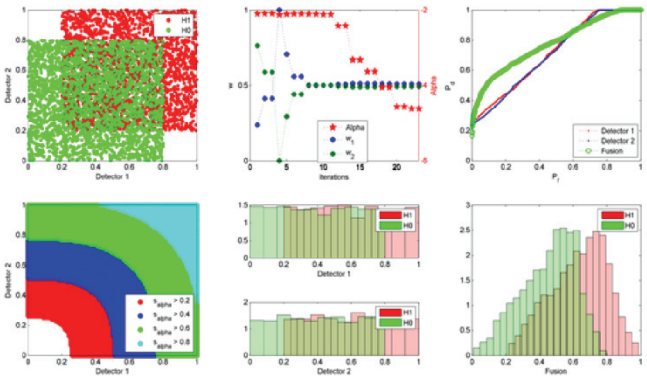

Figure 9. Experiment 8: $N_0 = N_1 = 0.5N$, $w_1 = w_2 = 0.5$, $\beta = 0.5$, $l_{10}^l = l_{20}^l = 0$, $l_{10}^u = l_{20}^u = 0.8$, $l_{11}^l = l_{21}^l = 0.2$, $l_{11}^u = l_{21}^u = 1$. Parameters are optimized by AUCmax ($nAUC_0^{0.1}$).

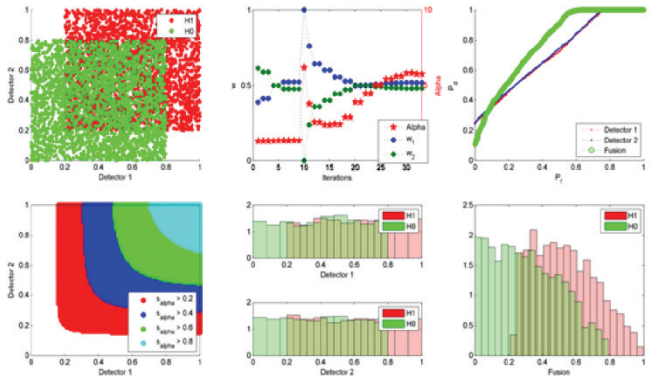

Figure 10. Experiment 9: $N_0 = N_1 = 0.5N$, $w_1 = w_2 = 0.5$, $\beta = 0.5$, $l_{10}^l = l_{20}^l = 0$, $l_{10}^u = l_{20}^u = 0.8$, $l_{11}^l = l_{21}^l = 0.2$, $l_{11}^u = l_{21}^u = 1$. Parameters are optimized by AUCmax ($nAUC_{0.6}^1$).

In these two cases, due to the same behavior of both detectors the estimated weighting parameters are equal, but α, which controls the shape of the separation frontiers, converges to a value which allows a better probability of detection after fusion, in the specified false alarm intervals of the ROC curves.

5. Application of α-integration in biometrics score fusion

Biometrics refers to the automatic identification of an individual based on his/her physiological traits (Jain, Ross & Prabhakar, 2004). The performance of a biometric system can be measured by reporting its false accept rate (FAR), equivalent to the concept of probability of false alarm P_f considered so far, and false reject rate (FRR), equivalent to the concept of 1- P_d . These systems are subject to low FAR (usually less than 0.1%).

Biometric systems based on a single source of information (unimodal systems) suffer from limitations like the lack of uniqueness, non-universality and noisy data (Jain & Ross, 2004) and hence, may not be able to achieve the desired performance requirements of real-world applications. In contrast, multimodal biometric systems combine information from its component modalities to arrive at a decision (Ross & Jain, 2003). Multimodal biometric authentication requires fusing information of different modalities like fingerprint, face, iris, retina, voice, ... Several studies (Toh, Jiang, & Yau, 2004; Wang, Tan, & Jain 2003) have demonstrated that by consolidating information from multiple sources, better performance can be achieved compared to the individual unimodal systems.

In a multimodal biometric system, integration can be done at (i) feature level, (ii) matching score level, or (iii) decision level. Matching score level fusion is commonly preferred because matching scores are easily available and contain sufficient information to distinguish between a genuine and an impostor case. Given a number of biometric systems, one can generate matching scores for a pre-specified number of users even with-out knowing the underlying feature extraction and matching algorithms of each biometric system. Thus, combining information contained in the matching scores seems both feasible and practical (Dass, Nandakumar & Jain, 2004).

In this paper we have tested the use of α-integration to fuse the matching scores in a multimodal biometric system. In particular we have used the public database *Biometric Scores Set - Release 1* (BSSR1) (N. US Department, 2013). BSSR1 is a set of raw output similarity scores from two face recognition systems and one

fingerprint system, operating on frontal faces, and left and right index live-scan fingerprints, respectively. The data are intended to permit interested parties to investigate a range of outstanding statistical problems related to biometrics. BSSR1 contains three partitions, which content is described in Table I.

PARTITION	*Number of individuals*	*Number of detectors*	*Scores available by detector*
1	$3x10^3$	**4** 2 measures of 2 face matchers	Total: $9x10^6$ Genuine: $3x10^3$
2	$6x10^3$	**2** 1 measure of right and 1 measure of left index fingerprint of 1 fingerprint matcher	Total: $36x10^6$ Genuine: $6x10^3$
3	517	**4** 1 measure of 2 face matchers, 1 measure of right and 1 measure of left index fingerprint of 1 fingerprint matcher	Total: 517^2 Genuine: 517

Table I. Description of the BSSR1partition content.

Many possible experiments may be devised from these three partitions. We have selected four experiments whose results are respectively shown in Tables II to V. In all the experiments we have obtained the GARs corresponding to three different FARs, for different methods of score fusion. The shown GAR values are the average of 30 iterations. In every iteration the available score set of the corresponding BSSR1 partition have been randomly divided into two halves. The first half has been used for training and the second for evaluation.

Thus in Table II, α-integration based on LMSE and on AUCmax criteria, are compared with simpler rules. The partition 1 was considered, and the scores were normalized between 0 and 1, as this is a requirement for α-integration based on LMSE. Normalization was made by computing the *a posteriori* probability of every hypotheses given the score, i.e.

$$s_{norm} = P(H_1|s) = \frac{f(s|H_1)P(H_1)}{f(s|H_1)P(H_1)+f(s|H_0)P(H_0)} \quad (25)$$

Where s_{norm} and s are, respectively, the scores after and before the normalization, $f(s|H_k)$ is the probability density of s conditioned to hypothesis H_k and $P(H_k)$ is the a priori probability of hypothesis H_k. These probabilities were estimated from the percentages of instances of H_k inside the training set of scores. Moreover, $f(s|H_k)$ has been estimated using nonparametric Gaussian kernel methods. Other methods of normalization are possible (Jain, Nandakumar & Ross, 2005), but its influence on the results is out of the scope of this work.

	FAR **0,001%**	***FAR*** **0,01%**	***FAR*** **0,1%**
Arithmetic mean	*97.859*	*98.823*	*99.510*
Geometric mean	*96.229*	*98.691*	*96.609*
Min	*72.305*	*79.816*	*85.724*
Max	*97.424*	*98.622*	*99.426*
α-integration (LMSE)	*83.767*	*97.019*	*98.693*
α-integration (AUCmax, $nAUC_0^{10^{-4}}$)	***98.851***	***99.135***	***99.601***

Table II. Experiment 1. GAR (%) corresponding to different methods applied to partition 1 of BSSR1. Scores were normalized by using (25).

As we can see in Table II, the best results are obtained with α-integration (AUCmax). Tuning the maximization of AUC in an interval of the ROC curve $P_f \in [0,10^{-4}]$, is important in this experiment if we compare with the results obtained by α-integration (LMSE). In fact, notice that in some cases of Table II, α-integration (LMSE) performs even worse than other simple rules. This is because no direct maximization of the GAR is made by α-integration (LMSE) and reinforces the interest of the new proposed criterion AUCmax.

In experiments 2, 3 and 4, we have considered the original scores without normalization, hence the α-integration (LMSE), was not applied. Each experiment corresponds to a different partition. Thus we show in Tables III, IV and V the results obtained with partitions 1, 2 and 3, respectively. We can see in all cases the superior performance of fusion based on α-integration (AUCmax), thus showing the interest of optimizing the fusing parameters

	$\boldsymbol{FAR}$ $\mathbf{0,001\%}$	$\boldsymbol{FAR}$ $\mathbf{0,01\%}$	$\boldsymbol{FAR}$ $\mathbf{0,1\%}$
Arithmetic mean	*92.990*	*93.901*	*96.172*
Geometric mean	*90.799*	*92.864*	*95.404*
Min	*57.969*	*73.896*	*84.135*
Max	*87.161*	*90.223*	*93.436*
α-integration (AUCmax, $nAUC_0^{10^{-4}}$)	***98.093***	***99.417***	***99.611***

Table III. Experiment 2. GAR (%) corresponding to different methods applied to partition 1 of BSSR1. Scores are not normalized.

	FAR $0,001\%$	FAR $0,01\%$	FAR $0,1\%$
Arithmetic mean	*88.393*	*91.170*	*93.895*
Geometric mean	*85.410*	*89.007*	*92.304*
Min	*75.546*	*79.740*	*84.425*
Max	*86.570*	*90.298*	*93.311*
α-integration (AUCmax, $nAUC_0^{10^{-4}}$)	***88.542***	***91.409***	***94.011***

Table IV. Experiment 3. GAR (%) corresponding to different methods applied to partition 2 of BSSR1. Scores are not normalized.

	FAR $0,001\%$	FAR $0,01\%$	FAR $0,1\%$
Arithmetic mean	*50.752*	*65.018*	*77.320*
Geometric mean	*65.135*	*74.904*	*83.998*
Min	*59.807*	*71.365*	*81.538*
Max	*49.176*	*63.914*	*76.416*
α-integration (AUCmax, $nAUC_0^{10^{-4}}$)	***66.799***	***75.971***	***84.995***

Table V. Experiment 4. GAR (%) corresponding to different methods applied to partition 3 of BSSR1. Scores are not normalized.

6. Estimating the α-integration parameters by MPE: an application in medical diagnosis

So far we have considered that the ROC curve of the integrated detector is the essential element to be optimized by α-integration. This is implicitly done with the LMSE criterion by trying to obtain integrated scores as close as possible to 1 when the true hypothesis is H_1 or to 0 when H_0 is in force. On the other hand ROC curves are explicitly optimized by AUCmax. This approach is appropriate in

those detection problems where having control of the probability of false alarm P_f is a crucial aspect. However, there are applications where it is better to minimize the probability of error P_e, i.e., the probability of selecting a wrong hypothesis. This is a typical criterion in digital transmission, where an error happens whenever a symbol "1" is decided in reception but the emitted symbol was "0" or viceversa. Thus P_e becomes the essential figure of merit of a digital communication system performance. There are other areas where minimizing the P_e of a detector is the appropriate optimization goal. One of them is automatic medical diagnosis. Very often long biosignal records, e.g., Electrocardiogram (ECG) and Electroencefalogram (EEG) recordings, must be visually analyzed by the medical expert to detect the possible presence of some predefined events in the signals. The amount and sequencing of these events may help in the diagnosis of pathologies. This task can be eased and dramatically accelerated by replacing the expert by an automatic detector. In this type of problem the goal is to reproduce the detections of the expert, which are considered as correct detections, as much as possible. Hence minimizing P_e is the best option. In this section we are going to show the results obtained by α-integration in the implementation of an automatic detector which integrates two scores corresponding to different modalities (EEG and ECG). Before that, let us propose in the following a new method for estimating the α-integration parameters, which is more appropriate for this kind of scenarios. We will call it Minimum Probability of Error (MPE) criterion.

As in section 2, let us assume that we have a training set of couples $\{ \boldsymbol{s}^j, y^j \}$ $j = 1 \ldots N$, where $\boldsymbol{s}^j = \left[s_1^j \ldots s_i^j \ldots s_d^j \right]^{\boldsymbol{T}}$ is the vector of scores provided by the detectors, and y^j is the corresponding known binary decision ($y^j = 1$ if H_1 is true and $y^j = 0$ if H_0 is true). Minimization of the P_e corresponding to the foregoing set is equivalent to maximization of the probability of taking correct decisions across the whole set of couples $\{ \boldsymbol{s}^j, y^j \}$ $j = 1 \ldots N$. Let us call P_c^j to the probability of taking a correct decision y^j from the fused score $s_\alpha(\boldsymbol{s}^j)$; it can be expressed in the form

$$P_c^j = P\left(y^j = 1/s_\alpha(\boldsymbol{s}^j)\right)^{y_j} P\left(y^j = 0/s_\alpha(\boldsymbol{s}^j)\right)^{1-y_j} \quad . \tag{24}$$

We assume that the scores to be fused are normalized and calibrated (Jain, Nandakumar & Ross, 2005), (Zadrozny & Elkan, 2002), so that we can consider that $P(y^j = 1/s_i^j) = s_i^j \quad i = 1 \ldots d$. Therefore, after α-integration we have that $P\left(y^j = 1/s_\alpha(\boldsymbol{s}^j)\right) = s_\alpha(\boldsymbol{s}^j)$. Then, substituting in (24)

$$P_c^j = s_\alpha(\boldsymbol{s}^j)^{y_j} \left(1 - s_\alpha(\boldsymbol{s}^j)\right)^{1-y_j} \quad . \tag{25}$$

Let us call P_c to the probability of taking correct decisions across the whole set of couples $\{ \boldsymbol{s}^j, y^j \}$ $j = 1 \ldots N$. If the measurements are independent for different values of the index j, we can write:

$$P_c = \prod_{j=1}^{N} P_c^j = \prod_{j=1}^{N} s_\alpha(\boldsymbol{s}^j)^{y_j} \left(1 - s_\alpha(\boldsymbol{s}^j)\right)^{1-y_j} \quad . \tag{26}$$

Finally, taking logarithms in (26) and changing the sign, we define the cost function to be minimized with the MPE criterion

$$-lnP_c = -\sum_{j=1}^{N} y_j ln\left(s_\alpha(\boldsymbol{s}^j)\right) + (1-y_j) ln\left(1-s_\alpha(\boldsymbol{s}^j)\right) \quad . \qquad (27)$$

Minimization can be done also by a gradient algorithm. Let us compute the required derivatives

$$\frac{\partial(-lnP_c)}{\partial\alpha} = -\sum_{j=1}^{N}\left(\frac{y^j}{s_\alpha(\boldsymbol{s}^j)} - \frac{1-y^j}{1-s_\alpha(\boldsymbol{s}^j)}\right)\frac{\partial s_\alpha(\boldsymbol{s}^j)}{\partial\alpha} \quad , \qquad (28a)$$

$$\frac{\partial(-lnP_c)}{\partial w_i} = -\sum_{j=1}^{N}\left(\frac{y^j}{s_\alpha(\boldsymbol{s}^j)} - \frac{1-y^j}{1-s_\alpha(\boldsymbol{s}^j)}\right)\frac{\partial s_\alpha(\boldsymbol{s}^j)}{\partial w_i} \quad . \qquad (28b)$$

Where $\frac{\partial s_\alpha(\boldsymbol{s}^j)}{\partial\alpha}$ can be computed using equations (9b) and (9c), and $\frac{\partial s_\alpha(\boldsymbol{s}^j)}{\partial w_i}$ can be computed using (10b). Hence a gradient algorithm like the one in equations (11) and (12) can be implemented using these new derivatives to obtain the MPE parameters.

With these estimated parameters, given any vector $\boldsymbol{s} = [s_1 \dots s_i \dots s_d]^T$ of individuals scores, we are able to compute the integrated score $s_\alpha(\boldsymbol{s})$. Then we have to implement the test which takes the final decision in a form which is to be consistent with the essential objective of minimizing P_e . It is well known in detection theory (Hippenstiel, 2002) that the optimum detector which minimize P_e from (in this case) the observation $s_\alpha(\boldsymbol{s})$ is obtained by the test

$$P\left(y = 1/s_\alpha(\boldsymbol{s})\right) \underset{H_0}{\overset{H_1}{\gtrless}} P\left(y = 0/s_\alpha(\boldsymbol{s})\right) \quad . \quad (29)$$

But $P\left(y = 0/s_\alpha(\boldsymbol{s})\right) = 1 - P\left(y = 1/s_\alpha(\boldsymbol{s})\right)$ so (29) is equivalent to

$$P\left(y = 1/s_\alpha(\boldsymbol{s})\right) \underset{H_0}{\overset{H_1}{\gtrless}} \frac{1}{2} \quad . \quad (30)$$

But we have assumed that $P\left(y = 1/s_\alpha(\boldsymbol{s})\right) = s_\alpha(\boldsymbol{s})$, so the MPE test will simply be

$$s_\alpha(\boldsymbol{s}) \underset{H_2}{\overset{H_1}{\gtrless}} \frac{1}{2} \quad . \quad (31)$$

We have considered the MPE criterion in the estimation of the α-integration parameters in an application of medical diagnosis. The problem belongs to the area of computer-assisted sleep staging (Agarwal & Gotman, 2001). In particular we want to build an automatic detector of arousals during sleeping, as their frequency of appearance are related with the presence of apnea and epilepsy. Normally, arousals are detected by a medical expert from a visual inspection of the so called polysomnograms (PSG), which are a set of EEGs obtained from the patient while sleeping. This manual task is tedious and susceptible to errors after a long period of analysis. Then, it is proposed in (Salazar, Vergara & Miralles, 2010) and automatic technique which, extracting four features from the PSG signals, generates automatic detections of arousals every epoch of 30 seconds. The

method consists in a Bayesian classifier which assumes a Hidden Markov Model for the evolution of the sleeping stages and a Non Gaussian Mixture model for the multivariate probability density in the feature space (please see (Salazar, Vergara & Miralles, 2010) for more specific details).

Here we want to verify the possible improvement in the detection of arousals by combined use of EEG and ECG information. From the ECG records and after some standard signal processing (Kaufmann, T., Sütterlin, S., Schulz, S.M., & Vögele, C., 2011), the heart beats (R-peaks) are extracted. Then the sequence of RR intervals between consecutive R-peaks is formed. This is termed the Heart Rate Variability (HRV) signal, which has been extensively used for health monitoring -see for example (Bouziane, Yagoubi, Vergara, & Salazar, 2015) and references there in-. Three features are extracted in every 30 seconds epoch. Two of them are time domain features: the mean and the standard deviation of the RR intervals. The third feature is the quotient between the low- frequency (LF) (0.04-0.15Hz) and the high-frequency (HF) (0.15-0.4Hz) powers, obtained from the Power Spectral Density (PSD) of the RR-sequence.

For the experiment we had four subjects. EEG and ECG signals were synchronously recorded for every subject during sleeping. Every recording session lasted some 7,5 hours (some 900 epochs of 30 seconds). A medical expert generated a binary decision for every epoch (presence or not presence of arousal), which will be the target decision or ground truth. For every subject, we used the first half of his recording session for training and the second half for testing the detectors performance. Using the methods described in (Salazar, Vergara &

Miralles, 2010), a score is generated from the EEG information for every epoch. Similarly, a score is obtained from the ECG features described in the foregoing paragraph, using a Support Vector Machine (SVM) classifier. Finally both scores are α-integrated.

The goal is to reproduce the manual detections given by the expert as much as possible, then every discrepancy with the expert will be considered an error and the probability of error is to be minimized. Hence, the decisions corresponding to the EEG and ECG modalities are obtained by respectively introducing the EEG score and the ECG score in the test of equation (31). On the other hand, we have used the MPE criterion for the estimation of the α-integration parameters, and the α-integrated score is also considered in the test (31) to generate decisions.

The left side of Table VI shows the results in terms of percentage of decisions which coincide with the expert decisions for the three possible automatic cases: isolated scores obtained from the EEG signals, isolated scores obtained from the EEG signals and scores derived from α-integration of both. The corresponding α-integration parameters estimated with the MPE criterion, are indicated in the right side of Table VI. We see that improvements after α-integration appear in subjects 1, 2 and 3, meanwhile the percentage corresponding to subject 4 is the same one obtained with isolated ECG scores. The very large value of α corresponding to subject 4 confirms that the minimum score is selected which seems to correspond to the ECG score in this case. Moreover, the weights are clearly unbalanced in favor of the ECG score, in subject 4. In any case, notice that α-integration yields a

performance which is as least as good as the best individual performance. Thus, even in the case of no improvement, α-integration is able to "select" the best automatic detector between the two available.

	EEG (%)	ECG (%)	α–int (%)		α	w_{EEG}	w_{ECG}
Subject 1	78.60	80.55	**84.70**		10.95	0.5053	0.4947
Subject 2	77.39	74.37	**77.51**		17.02	0.5552	0.4448
Subject 3	89.13	90.48	**91.72**		10.15	0.4306	0.5786
Subject 4	80.45	**93.93**	**93.93**		96.02	0.2009	0.7991

Table VI. Left side: Percentage of decisions coincident with the expert decisions corresponding to EEG scores, ECG scores and α-integrated scores. Right side: estimated α-integration parameters with the MPE criterion.

7. Conclusions

We have presented a new method for the fusion of scores obtained from different detectors based on α-integration. It is a generalization of simpler rules which allows optimum fitting of the parameters and find rationale in the optimum integration of stochastic models. Three optimality criteria have been considered: LMSE, AUCmax and MPE. While the first two relates implicitly or explicitly in optimizing the ROC curves , i.e., maximizing probability of detection for a given probability of false alarm, the last one focus in minimizing the probability of error.

We have proposed new gradient algorithms for the three criteria. In the LMSE case, we have adapted to the detection context a gradient algorithm previously proposed in the general framework of α-integration. Some variations have been included to account for unbalanced distribution of the training data sizes and relative significance of every type of error in the global MSE. Regarding AUCmax, a new algorithm has been proposed based on transforming an empirical nonparametric measure of AUC in a differentiable function. A key advantage of AUCmax with respect to LMSE is that it allows tuned optimization in selected intervals of the ROC curves. In MPE a new cost function is defined which is the negative of the log probability of correct answers.

We have included different experiments with simulated data with the aim of illustrating the different factors influencing α-integration with both LMSE and AUCmax. It has been shown that the fusion of two-detector scores, leads to significant improvements of the ROC curves.

Finally, two real data cases have been considered. The first one corresponds to the fusion of scores in multimodal biometric data. In this application the goal is to have the maximum genuine acceptation ratio (equivalent to probability of detection) for a given (rather small) false alarm ratio, hence both LMSE and AUCmax have been considered. Different experiments have been done with different data sets, showing the superior performance of α-integration with respect to simpler rules, which not allow the optimization of the fusing parameters. It has been demonstrated also the interest of the tuning capability of AUCmax to a selected range of probabilities of false alarm.

The second real data case is in the area of automatic analysis of medical records to reproduce the manual decisions taken by the medical expert, so the best criterion is MPE. We have presented the theoretical analysis, including gradient computations, of α-integration based on MPE. The method has been applied in the fusion of two scores, respectively obtained from EEG and ECG records. The problem was the automatic detection of arousals during sleeping, which is currently done manually by the medical expert. Experiments in four subjects have illustrated the potential interest of MPE α-integration in these kind of problems.

Acknowledgements

This work has been supported by Generalitat Valenciana under grants PROMETEOII 2014-032, ISIC2012-006 and by Spanish administration under grant TEC2014-58438-R.

References

Agarwal, R., Gotman, J. (2001). Computer-assisted sleep staging, *IEEE Trans. On Biomedical Engineering*, *48* (12), 1412-1423.

Amari, S. (2007). Integration of stochastic models by minimizing α-divergence. *Neural Computation*, *19*, 2780-2796.

Atrey, P., Hossain, M., El Saddik, A., & Kankanhalli M. (2010). Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, *16*, 345-379.

Bouziane, A., Yagoubi, B., Vergara, L., & Salazar, A. (2015). The ANS Sympathovagal Balance Using a Hybrid Method Based on the Wavelet Packet

and the KS-Segmentation Algorithm. In *Proc. Int'l. Conf. Circuits, Systems, Signals and Telecomm., (CSST)* (pp. 75-83) WSEAS press.

Byrd R. H., Hribar M. E., & Nocedal J. (1999). An interior point algorithm for large scale nonlinear programming, *SIAM J. Optim.,9* (4), 877-900.

Choi, H., Choi, S., Katake, A., & Choe, Y. (2010). Learning α-integration with partially labeled data. In *Proc. IEEE Int'l. Conf. Acoustics, Speech, and Signal Processing (ICASSP)* (pp. 2058-2061) Piscataway, NJ: IEEE.

Choi, H., Choi, S., & Choe, Y. (2013). Parameter Learning for Alpha Integration, *Neural Computation, 25*, 1585–1604.

Dass, S. C., Nandakumar, K., & Jain, A. K. (2004) A principled approach to score level fusion in multimodal biometric systems. *In Proceedings of Fifth International Conference on AVBPA* (pp. 1049–1058). Berlin: Springer.

Dodd, L. E., & Pepe, M. S. (2003). Partial AUC estimation and regression, *Biometrics*, *59* (3), 614-623.

Herschtal, A., & Raskutti, B. (2004). Optimising area under the ROC curve using gradient descent. In *Proc. 21st International Conference on Machine Learning (ICML*) (pp. 49-56). New York: ACM.

Hippenstiel, R.D. (2002). *Detection Theory: Application and Digital Signal Processing*. Boca Raton, Florida: CRC Press.

Jain, A., Nandakumar, K, & Ross, A. (2005). Score normalization in multimodal biometric systems. *Pattern Recognition, 38*, 2270-2285.

Jain, A. K., & Ross, A. (2004). Multibiometric Systems. *Communications of the ACM, Special Issue on Multimodal Interfaces*, *47*, 34-40.

Jain, A. K., Ross, A., & Prabhakar, S. (2004). An Introduction to Biometric Recognition. *IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on Image- and Video-Based Biometrics*, 14, 4-20.

Kaufmann, T., Sütterlin, S., Schulz, S.M., & Vögele, C., (2011). ARTiiFACT: a tool for heart rate artifact processing and heart rate variability analysis. *Behavior Research Methods* 43 (4), 1161–1170.

Khaleghi, B., Khamis, A., Karray, F. O., & Razavi, S.N. (2013). Multisensor data fusion: A review of the state-of-the-art. *Information Fusion* ,*14* (1), 28-44.

Kittler, J., Hatef, M., Duin, R. P., & Matas, J. (1998). On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20* (3), 226–239.

N. US Department of Commerce, Biometric Scores Set. *[On line]: http://www.nist.gov/itl/iad/ig/biometricscores.cfm.* [Accessed: 24-jun-2013].

Narasimhan, H., & Agarwal, S. (2013). A Structural {SVM} Based Approach for Optimizing Partial AUC. In *Proc. 30th International Conference on Machine Learning (ICML)* (pp 516-524), JMRL W&CP.

Parker, Ch. (2013). On measuring the performance of binary classifiers. *Knowledge and Information Systems*, *35* (1) , 131-152.

Pimentel, M.A.F., Clifton, D. A., Clifton, L., & Tarassenko, L. (2014). A review of novelty detection, *Signal Processing, 99*, 215-249.

Ross, A, Jain, & A. K. (2003). Information Fusion in Biometrics. *Pattern Recognition Letters*, *Special Issue on Multimodal Biometrics, 24*, 2115–2125.

Salazar A., Vergara, L, & Miralles, R. (2010). On including sequential dependencies in ICA mixtures models. *Signal Processing*, *90*, 2314-2318.

Soriano, A., Vergara, L., Moragues, J., & Miralles, R. (2014). Unknown signal detection by one-class detector based on Gaussian copula, *Signal Processing, 96*, 315–320.

Toh K.A., Jiang, X., & Yau, W.Y. (2004). Exploiting Global and Local Decisions for Multi-modal Biometrics Verification. *IEEE Transactions on Signal Processing, 52*, 3059–3072.

Waltz, R. A., Morales, J. L., Nocedal, J., & Orban, D. (2006). An interior algorithm for nonlinear optimization that combines line search and trust region steps. *Math. Program., 107* (3) , 391-408.

Wang Y., Tan T., & Jain, A. K. (2003). Combining Face and Iris Biometrics for Identity Verification. *In Proceedings of Fourth International Conference on AVBPA* (pp 805-813). Berlin: Springer.

Wu, D. (2009). Parameter Estimation for α-GMM Based on Maximum Likelihood Criterion. *Neural Computation, 21*, 1776–1795.

Yuksel, S. E., Wilson, J. N., & Gader, P. D. (2012). Twenty Years of Mixture of Experts. *IEEE Trans. On Neural Networks and Learning Systems, 23* (4), 1177-1193.

Zadrozny, B., & Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In *Proc. 8th ACM SIGKDD International Conference on Knowledge Discovery and Data mining* (pp. 694-699). New York: ACM.

Figures Captions

Figure 1. Approximating a unit step function using a sigmoid function.

Figure 2. Experiment 1: $N_0 = N_1 = 0.5N$, $w_1 = w_2 = 0.5$, $\beta = 0.5$, $l_{10}^l = l_{20}^l = 0$, $l_{10}^u = l_{20}^u = 0.8$, $l_{11}^l = l_{21}^l = 0.2$, $l_{11}^u = l_{21}^u = 1$.

Figure 3. Experiment 2: $N_0 = 0.2N$ $N_1 = 0.8N$, $w_1 = w_2 = 0.5$, $\beta = 0.5$, $l_{10}^l = l_{20}^l = 0$, $l_{10}^u = l_{20}^u = 0.8$, $l_{11}^l = l_{21}^l = 0.2$, $l_{11}^u = l_{21}^u = 1$.

Figure 4. Experiment 3: $N_0 = N_1 = 0.5N$, $w_1 = w_2 = 0.5$, $\beta = 0.9$, $l_{10}^l = l_{20}^l = 0$, $l_{10}^u = l_{20}^u = 0.8$, $l_{11}^l = l_{21}^l = 0.2$, $l_{11}^u = l_{21}^u = 1$.

Figure 5. Experiment 4: $N_0 = N_1 = 0.5N$, $w_1 = w_2 = 0.5$, $\beta = 0.1$, $l_{10}^l = l_{20}^l = 0$, $l_{10}^u = l_{20}^u = 0.8$, $l_{11}^l = l_{21}^l = 0.2$, $l_{11}^u = l_{21}^u = 1$.

Figure 6. Experiment 5: $N_0 = N_1 = 0.5N$, $\alpha = -1$, $\beta = 0.5$, $l_{10}^l = l_{20}^l = 0$, $l_{10}^u = l_{20}^u = 0.8$, $l_{11}^l = l_{21}^l = 0.2$, $l_{11}^u = l_{21}^u = 1$.

Figure 7. Experiment 6: $N_0 = N_1 = 0.5N$, $\alpha = -1$, $\beta = 0.5$, $l_{10}^l = l_{20}^l = 0$, $l_{10}^u = l_{20}^u = 0.8$, $l_{11}^l = 0.4$ $l_{21}^l = 0.2$, $l_{11}^u = l_{21}^u = 1$.

Figure 8. Experiment 7: $N_0 = N_1 = 0.5N$, $w_1 = w_2 = 0.5$, $\beta = 0.5$, $l_{10}^l = l_{20}^l = 0$, $l_{10}^u = l_{20}^u = 0.8$, $l_{11}^l = l_{21}^l = 0.2$, $l_{11}^u = l_{21}^u = 1$. Parameters are optimized by AUCmax ($nAUC_0^1$).

Figure 9. Experiment 8: $N_0 = N_1 = 0.5N$, $w_1 = w_2 = 0.5$, $\beta = 0.5$, $l_{10}^l = l_{20}^l = 0$, $l_{10}^u = l_{20}^u = 0.8$, $l_{11}^l = l_{21}^l = 0.2$, $l_{11}^u = l_{21}^u = 1$. Parameters are optimized by AUCmax ($nAUC_0^{0.1}$).

Figure 10. Experiment 9: $N_0 = N_1 = 0.5N$, $w_1 = w_2 = 0.5$, $\beta = 0.5$, $l_{10}^l = l_{20}^l = 0$, $l_{10}^u = l_{20}^u = 0.8$, $l_{11}^l = l_{21}^l = 0.2$, $l_{11}^u = l_{21}^u = 1$. Parameters are optimized by AUCmax ($nAUC_{0.6}^1$).

Table captions

Table I. Description of the BSSR1partition content.

Table II. Experiment 1. GAR (%) corresponding to different methods applied to partition 1 of BSSR1. Scores were normalized by using (25).

Table III. Experiment 2. GAR (%) corresponding to different methods applied to partition 1 of BSSR1. Scores are not normalized.

Table IV. Experiment 3. GAR (%) corresponding to different methods applied to partition 2 of BSSR1. Scores are not normalized.

Table V. Experiment 4. GAR (%) corresponding to different methods applied to partition 3 of BSSR1. Scores are not normalized.

Table VI. Left side: Percentage of decisions coincident with the expert decisions corresponding to EEG scores, ECG scores and α-integrated scores. Right side: estimated α-integration parameters with the MPE criterion.