Nonlinear Multiantenna Detection Methods

Sheng Chen

School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK
Email: sqc@ecs.soton.ac.uk

Lajos Hanzo

School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK
Email: lh@ecs.soton.ac.uk

Andreas Wolfgang

School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK
Email: aw03r@ecs.soton.ac.uk

A nonlinear detection technique designed for multiple-antenna assisted receivers employed in space-division multiple-access systems is investigated. We derive the optimal solution of the nonlinear spatial-processing assisted receiver for binary phase shift keying signalling, which we refer to as the Bayesian detector. It is shown that this optimal Bayesian receiver significantly outperforms the standard linear beamforming assisted receiver in terms of a reduced bit error rate, at the expense of an increased complexity, while the achievable system capacity is substantially enhanced with the advent of employing nonlinear detection. Specifically, when the spatial separation expressed in terms of the angle of arrival between the desired and interfering signals is below a certain threshold, a linear beamformer would fail to separate them, while a nonlinear detection assisted receiver is still capable of performing adequately. The adaptive implementation of the optimal Bayesian detector can be realized using a radial basis function network. Two techniques are presented for constructing block-data-based adaptive nonlinear multiple-antenna assisted receivers. One of them is based on the relevance vector machine invoked for classification, while the other on the orthogonal forward selection procedure combined with the Fisher ratio class-separability measure. A recursive sample-by-sample adaptation procedure is also proposed for training nonlinear detectors based on an amalgam of enhanced κ-means clustering techniques and the recursive least squares algorithm.

Keywords and phrases: smart antenna, adaptive beamforming, mean square error, bit error rate, Bayesian classification, radial basis function network.

1. INTRODUCTION

Spatial processing invoking adaptive antenna arrays has shown real promise in terms of attaining substantial capacity enhancements in mobile communication [1, 2, 3, 4, 5, 6, 7, 8]. Multiple-antenna aided receivers are capable of separating signals transmitted on the same carrier frequency, provided that signals are sufficiently separated in the spatial domain. Classically, beamforming algorithms create a linear combination of the signals received from the different elements of an antenna array. We refer to this classic beamforming principle as *linear* beamforming. A traditional approach to linear beamforming is based on the minimum mean square error (MMSE) principle that minimizes the mean square error (MSE) between the desired output generated from a known reference signal and the actual array output. Adaptive implementations of the linear MMSE (LMMSE) beamforming solution can readily be realized using the well-known family of temporal reference techniques [2, 3, 9, 10, 11, 12, 13]. Specifically, block-data-based beamformer weight adaptation can be achieved using the sample matrix inversion (SMI) algorithm [9, 10], while sample-by-sample based array-weight adaptation can be carried out using the least mean square (LMS) algorithm [11, 12, 13]. Recent work [14, 15] has investigated a linear beamforming technique based directly on minimizing the system's bit error rate (BER) rather than the MSE and developed both block-data-based and sample-by-sample adaptive algorithms for implementing linear minimum BER (LMBER) beamforming. The results of [14, 15] have demonstrated that LMBER beamforming is capable of providing considerable performance gains in terms of a reduced BER over the usual LMMSE beamforming.

In the context of space-division multiple access (SDMA), the spatial separation in angles of arrival between the desired
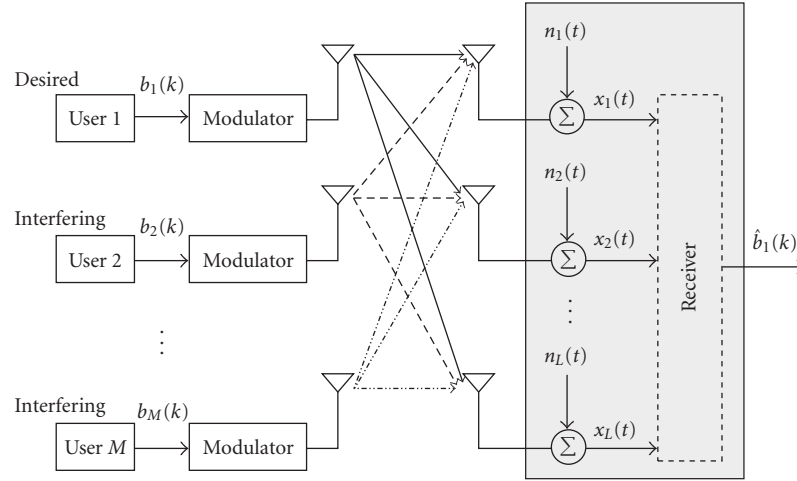
FIGURE 1: Multiantenna receiver configuration for the multiuser space-division multiple-access system.

signal and the closest interfering signal dominates the achievable system performance and hence the system's user capacity. When this angular separation is below a certain threshold, linear beamforming ultimately fails since the signals transmitted by the individual users become linearly inseparable, a situation that has also been observed in the context of single-user channel equalization and multiuser detection designed for code-division multiple access (CDMA) [16, 17, 18, 19, 20]. In fact, it has been observed even in linearly separable scenarios that a nonlinear processing technique is capable of providing a better performance than a linear one, although this is typically achieved at the cost of an increased complexity. In conjunction with nonlinear spatial processing, the achievable system capacity can be significantly increased since an adequate performance can be maintained even in case of a low angular separation compared to linear beamforming. These considerations motivate this study of nonlinear detection techniques contrived for multiantenna aided systems.

The outline of the paper is as follows. Section 2 introduces the system model, while Section 3 outlines our linear beamforming-based benchmarker. In Section 4, we derive the optimal solution of the nonlinear spatial processing assisted receiver for binary phase shift keying (BPSK) signalling, which is referred to as the Bayesian detection solution. It is shown that this Bayesian solution has an identical form to a radial basis function (RBF) network [17, 21]. In Section 5, two schemes are proposed for realizing block-data-based adaptive RBF detectors. One of them is based on the relevance vector machine (RVM) invoked for classification [22, 23] and the other one is the orthogonal forward selection (OFS) procedure using the Fisher ratio class-separability measure [24]. Finally, in Section 6, an adaptive sample-by-sample implementation of the RBF detector is also considered using an amalgam of the enhanced κ-means clustering and the recursive least squares (CRLS) algorithm [19, 25] before offering our conclusions in Section 7.

2. SYSTEM MODEL

We consider the multiple-antenna aided receiver configuration of Figure 1 invoked for assisting the operation of a multiuser SDMA system. It is assumed that the system supports M users (signal sources), and each user transmits a BPSK modulated signal on the same carrier frequency of $\omega = 2\pi f$. Let k denote the bit instance. Then the baseband signal of user i, sampled at symbol rate, is given by

$$m_i(k) = A_i b_i(k), \quad 1 \le i \le M, \tag{1}$$

where the complex-valued coefficient A_i models the multiplication of the channel coefficient of user i with the transmitted signal power of user i and therefore $|A_i|^2$ denotes the received signal power for user i, and $b_i(k) \in \{\pm 1\}$ is the kth bit of user i. Without any loss of generality, source 1 is assumed to be the desired user and the rest of the sources are the interfering users. A linear antenna array is considered which consists of L uniformly spaced elements, and the signals received by the L-element antenna array are given by

$$x_l(k) = \sum_{i=1}^{M} m_i(k) \exp\left(j\omega t_l(\theta_i)\right) + n_l(k) = \bar{x}_l(k) + n_l(k) \tag{2}$$

for $1 \le l \le L$, where $t_l(\theta_i)$ is the relative time delay at element l for source i, θ_i is the direction of arrival for source i, and $n_l(k)$ is a complex-valued white Gaussian noise with zero mean and $E[|n_l(k)|^2] = 2\sigma_n^2$. The desired user's signal-to-noise ratio is defined as SNR $= |A_1|^2/2\sigma_n^2$, and the desired signal-to-interference ratio with respect to interfering user i is defined by SIR$_i = |A_1|^2/|A_i|^2$ for $i = 2, \ldots, M$. In vectorial form, the antenna array output $\mathbf{x}(k) = [x_1(k) \ x_2(k) \ \cdots \ x_L(k)]^T$ can be expressed as

$$\mathbf{x}(k) = \bar{\mathbf{x}}(k) + \mathbf{n}(k) = \mathbf{P}\mathbf{b}(k) + \mathbf{n}(k), \tag{3}$$

where $\mathbf{n}(k) = [n_1(k) \ n_2(k) \ \cdots \ n_L(k)]^T$ has a covariance matrix of $E[\mathbf{n}(k)\mathbf{n}^H(k)] = 2\sigma_n^2 \mathbf{I}_L$ with $\mathbf{I}_L$ denoting the $L \times L$

identity matrix, the system matrix $\mathbf{P}$ is given by

$$\mathbf{P} = \begin{bmatrix} A_1\mathbf{s}_1 & A_2\mathbf{s}_2 & \cdots & A_M\mathbf{s}_M \end{bmatrix}, \qquad (4)$$

the steering vector for source i is formulated as

$$\mathbf{s}_i = \begin{bmatrix} \exp\left(j\omega t_1(\theta_i)\right) & \exp\left(j\omega t_2(\theta_i)\right) & \cdots & \exp\left(j\omega t_L(\theta_i)\right)^T \end{bmatrix}, \qquad (5)$$

and the transmitted bit vector is

$$\mathbf{b}(k) = \begin{bmatrix} b_1(k) & b_2(k) & \cdots & b_M(k) \end{bmatrix}^T. \qquad (6)$$

The task of the spatial-processing assisted receiver is to provide an estimate $\hat{b}_1(k)$ of the desired user's transmitted bit $b_1(k)$, given the input $\mathbf{x}(k)$. To keep our notations and the associated concepts relatively simple, we have used a BPSK modulation scheme, a narrowband channel model, and narrowband beamforming (space-only processing). The approach can be extended to other modulation schemes and wideband channels that induce intersymbol interference. The same idea can also be applied to broadband beamforming (space-time processing).

3. LINEAR BEAMFORMING ASSISTED RECEIVER

The output of the linear beamformer is given by

$$y(k) = \mathbf{w}^H\mathbf{x}(k) = \mathbf{w}^H\bar{\mathbf{x}}(k) + \mathbf{w}^H\mathbf{n}(k) = \bar{y}(k) + e(k), \quad (7)$$

where $\mathbf{w} = [w_1 \; w_2 \; \cdots \; w_L]^T$ is the complex-valued beamformer weight vector, and $e(k)$ is Gaussian distributed with a zero mean and a variance $E[|e(k)|^2] = 2\sigma_n^2\mathbf{w}^H\mathbf{w}$. The estimate of the transmitted bit $b_1(k)$ is given by

$$\hat{b}_1(k) = \operatorname{sgn}\left(y_R(k)\right) = \begin{cases} +1, & y_R(k) > 0, \\ -1, & y_R(k) \le 0, \end{cases} \qquad (8)$$

where $y_R(k) = \Re[y(k)]$ denotes the real part of $y(k)$. Classically, the linear beamformer's weight vector is determined by minimizing the MSE term of $E[|b_1(k) - y(k)|^2]$ between the desired user's transmitted bit and the beamformer's output, which leads to the following LMMSE solution:

$$\mathbf{w}_{\mathrm{MMSE}} = \left(\mathbf{P}\mathbf{P}^H + 2\sigma_n^2\mathbf{I}_L\right)^{-1}\mathbf{p}_1 \qquad (9)$$

with $\mathbf{p}_1$ being the first column of $\mathbf{P}$. Using a temporal reference technique aided approach [7], the LMMSE beamforming solution can be readily realized using the block-data-based SMI algorithm [7], and recursive sample-by-sample adaptation can be performed using the LMS or RLS algorithm [21].

In order to derive the BER formula of the linear beamformer with the weight vector $\mathbf{w}$, firstly note that there are $N_b = 2^M$ possible sequences of $\mathbf{b}(k)$, which are denoted as $\mathbf{b}_q$, $1 \le q \le N_b$. Furthermore, denote the first element of $\mathbf{b}_q$, corresponding to the desired user, as $b_{q,1}$. As expected, the noiseless part of the beamformer input signal, $\bar{\mathbf{x}}(k)$, assumes encountering values only from the signal set defined as

$$\mathcal{X} \triangleq \{\bar{\mathbf{x}}_q = \mathbf{P}\mathbf{b}_q, \; 1 \le q \le N_b\}. \qquad (10)$$

This set can be partitioned into two subsets depending on the specific value of $b_1(k)$ as follows:

$$\mathcal{X}^{(\pm)} \triangleq \{\bar{\mathbf{x}}_q^{(\pm)} \in \mathcal{X} : b_1(k) = \pm 1\}. \qquad (11)$$

Similarly, $\bar{y}(k)$ takes values from the scalar set

$$\mathcal{Y} \triangleq \{\bar{y}_q = \mathbf{w}^H\bar{\mathbf{x}}_q, \; 1 \le q \le N_b\} \qquad (12)$$

which can be divided into the two subsets defined as

$$\mathcal{Y}^{(\pm)} \triangleq \{\bar{y}_q^{(\pm)} \in \mathcal{Y} : b_1(k) = \pm 1\}. \qquad (13)$$

Thus, $\bar{y}_R(k)$ can only take values from the set

$$\mathcal{Y}_R \triangleq \{\bar{y}_{R,q} = \Re[\bar{y}_q], \; 1 \le q \le N_b\} \qquad (14)$$

which can be partitioned into the two subsets conditioned on the value of $b_1(k)$:

$$\mathcal{Y}_R^{(\pm)} \triangleq \{\bar{y}_{R,q}^{(\pm)} \in \mathcal{Y}_R : b_1(k) = \pm 1\}. \qquad (15)$$

It can be readily seen that the conditional probability density function (pdf) of $y(k)$ given $b_1(k) = +1$ is a Gaussian mixture given by

$$p(y|+1) = \frac{1}{N_{sb}} \sum_{q=1}^{N_{sb}} \frac{1}{2\pi\sigma_n^2\mathbf{w}^H\mathbf{w}} \exp\left(-\frac{|y - \bar{y}_q^{(+)}|^2}{2\sigma_n^2\mathbf{w}^H\mathbf{w}}\right), \qquad (16)$$

where $\bar{y}_q^{(+)} \in \mathcal{Y}^{(+)}$ and $N_{sb} = N_b/2$ is the number of the points in $\mathcal{Y}^{(+)}$. Therefore, the conditional marginal pdf of $y_R(k)$ given $b_1(k) = +1$ is formulated as follows:

$$p(y_R|+1) = \frac{1}{N_{sb}} \sum_{q=1}^{N_{sb}} \frac{1}{\sqrt{2\pi\sigma_n^2\mathbf{w}^H\mathbf{w}}} \exp\left(-\frac{(y_R - \bar{y}_{R,q}^{(+)})^2}{2\sigma_n^2\mathbf{w}^H\mathbf{w}}\right), \qquad (17)$$

where $\bar{y}_{R,q}^{(+)} \in \mathcal{Y}_R^{(+)}$. Thus, it can be shown that the BER of the linear beamformer associated with the weight vector $\mathbf{w}$ is given by [14, 15]

$$P_E(\mathbf{w}) = \frac{1}{N_{sb}} \sum_{q=1}^{N_{sb}} Q(g_{q,+}(\mathbf{w})), \qquad (18)$$

where

$$Q(u) = \frac{1}{\sqrt{2\pi}} \int_u^\infty \exp\left(-\frac{v^2}{2}\right) dv,$$

$$g_{q,+}(\mathbf{w}) = \frac{\operatorname{sgn}(b_{q,1})\bar{y}_{R,q}^{(+)}}{\sigma_n\sqrt{\mathbf{w}^H\mathbf{w}}} = \frac{\operatorname{sgn}(b_{q,1})\Re[\mathbf{w}^H\bar{\mathbf{x}}_q^{(+)}]}{\sigma_n\sqrt{\mathbf{w}^H\mathbf{w}}}. \qquad (19)$$
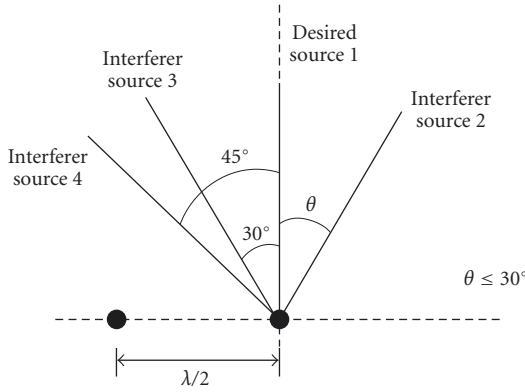
FIGURE 2: Locations of the desired source and the interfering sources with respect to the two-element linear antenna array having $\lambda/2$ element spacing, where λ is the wavelength.

The LMBER beamforming solution is then defined as follows:

$$\mathbf{w}_{\text{MBER}} = \arg \min_{\mathbf{w}} P_E(\mathbf{w}). \qquad (20)$$

Unlike the LMMSE solution (9), there exists no closed-form LMBER solution. In [14, 15], a simplified conjugate gradient method [26, 27] is used to obtain numerical solutions. Both the block-data-based gradient and LMS-style stochastic gradient adaptive algorithms have been derived in [14, 15] to realize the LMBER beamforming solution.

For the linear beamformer to work adequately, the underlying system must be linearly separable. The linear separability means that there exists a weight vector $\mathbf{w}$ such that $\mathcal{Y}_R^{(-)}$ and $\mathcal{Y}_R^{(+)}$ are completely separated by the decision threshold $y_R = 0$. When the minimum spatial separation expressed in angles of arrival between the desired user and interfering users is below a certain threshold, the system inevitably becomes linearly inseparable. In such a situation, the linear beamformer will have a high irreducible BER floor, and nonlinear processing has to be adopted for the sake of achieving an adequate BER performance. In general, nonlinear spatial processing is capable of achieving a better performance than a linear receiver, regardless whether the output of the system is linearly separable or not. The limitation of a linear beamforming assisted receiver is illustrated in the following example, which is also used throughout this paper for investigating the proposed nonlinear multiantenna detection techniques.

Simulation example

The example consisted of four signal sources and a two-element antenna array. Figure 2 shows the locations of the desired source and the three interfering sources in a graphical form. The simulated channel conditions were $A_i = 1 + j0$, $1 \leq i \leq 4$. The desired user and all the three interfering users had equal signal power, and therefore we had $\text{SIR}_i = 0$ dB for $i = 2, 3, 4$. The minimum spatial separation in this example was the difference in angles of arrival between the desired user 1 and the interferer 2, which was $\theta \leq 30°$. Figure 3 com-

pares the BERs of the LMMSE and LMBER beamformers for the two cases of $\theta = 30°$ and $\theta = 10°$, respectively. It can be seen from Figure 3a that for $\theta = 30°$, the underlying system scenario was linearly separable as was confirmed by the performance of the LMBER beamformer, while the LMMSE beamformer was unable to achieve the linear separability of the signal constellation and hence exhibited a high BER floor. Figure 4 plots the conditional pdfs $p(y| + 1)$, the conditional marginal pdfs $p(y_R| + 1)$, and the conditional subsets $\mathcal{Y}^{(+)}$ and $\mathcal{Y}_R^{(+)}$ for the LMMSE and LMBER beamformers, given $\theta = 30°$ and SNR $= 10$ dB, which represented a typical condition in Figure 3a. It is clearly seen from Figure 4 that the LMBER beamformer was "smarter" than the LMMSE scheme and hence achieved the desired linear separability. However, when the minimum spatial separation was reduced to $\theta = 10°$, the system became inherently linearly inseparable, and any linear beamformer failed to perform adequately as can be seen in Figure 3b. Figure 5 depicts the conditional pdfs $p(y| + 1)$, the conditional marginal pdfs $p(y_R| + 1)$, and the conditional subsets $\mathcal{Y}^{(+)}$ and $\mathcal{Y}_R^{(+)}$ for the LMMSE and LMBER beamformers, given $\theta = 10°$ and SNR $= 10$ dB, which provided a typical condition in Figure 3b. The results of Figure 5 confirm that the underlying system was linearly inseparable, and it also explains why the LMBER solution did better than the LMMSE scheme, resulting in a lower BER floor. This example clearly demonstrates the need for invoking a nonlinear spatial-processing assisted receiver structure.

4. BAYESIAN DETECTION SCHEME

Given the observation vector $\mathbf{x}(k)$, the optimal solution to the multiantenna aided spatial processing problem in terms of the achievable BER is the maximum a posteriori probability solution, which is similar to the case of single-user channel equalization [17, 18], and therefore can readily be formulated. The posterior probabilities or decision variables for $b_1(k) = \pm 1$ given $\mathbf{x}(k)$ are given by

$$\eta^{(\pm)}(k) = \sum_{q=1}^{N_{sb}} \frac{\xi_q^{(\pm)}}{(2\pi\sigma_n^2)^L} \exp\left(-\frac{||\mathbf{x}(k) - \bar{\mathbf{x}}_q^{(\pm)}||^2}{2\sigma_n^2}\right), \qquad (21)$$

where $\xi_q^{(\pm)}$ are a priori probabilities of $\bar{\mathbf{x}}_q^{(\pm)}$ and $||\mathbf{x}||^2 = \mathbf{x}^H\mathbf{x}$. Typically, all the states $\bar{\mathbf{x}}_q^{(\pm)}$ are equiprobable, and thus we have $\xi_q^{(\pm)} = 1/N_b$. The optimal decision regarding the transmitted bit $b_1(k)$ is given by

$$\hat{b}_1(k) = \begin{cases} +1, & \eta^{(+)}(k) \geq \eta^{(-)}(k), \\ -1, & \text{otherwise.} \end{cases} \qquad (22)$$

We redefine a single decision variable as

$$y_B(k) = \sum_{q=1}^{N_b} v_q \exp\left(-\frac{||\mathbf{x}(k) - \bar{\mathbf{x}}_q||^2}{2\sigma_n^2}\right), \qquad (23)$$
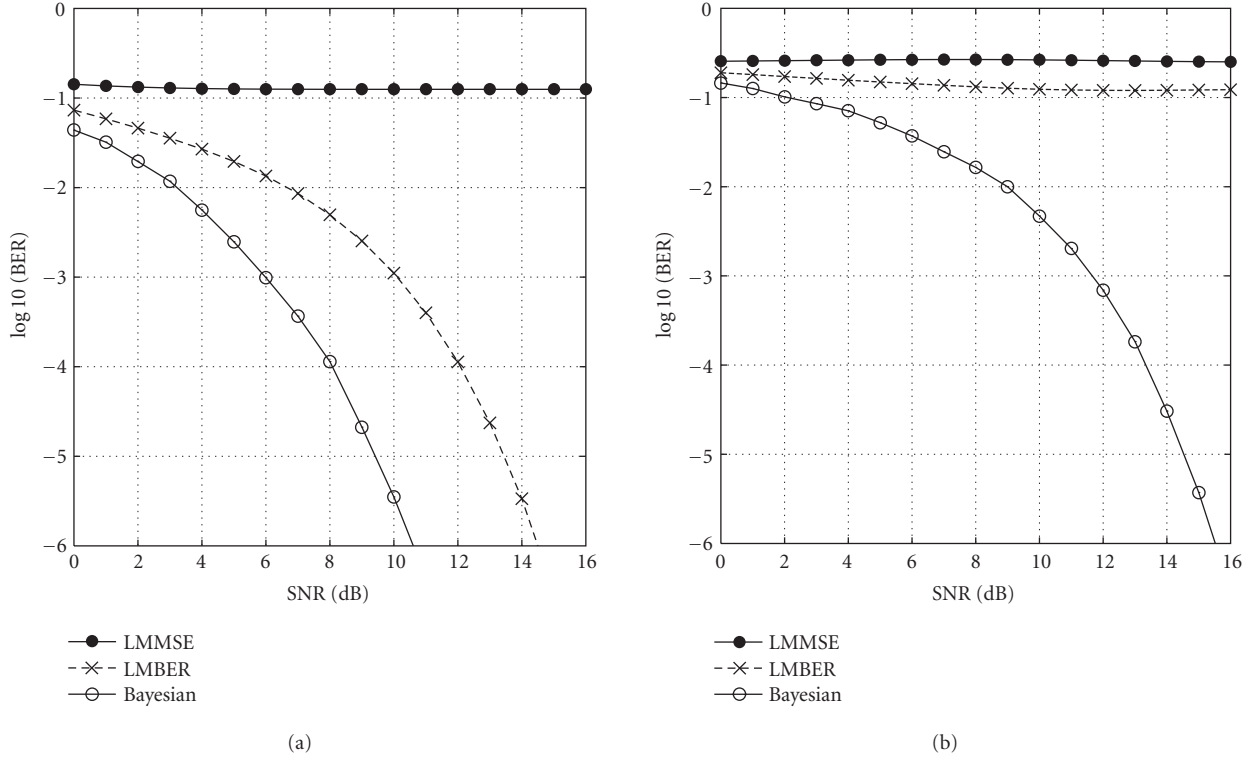
FIGURE 3: Comparison of the bit error rates of three theoretical detection schemes: the LMMSE and LMBER beamformers, and the optimal Bayesian detector. (a) $\theta = 30°$. (b) $\theta = 10°$.
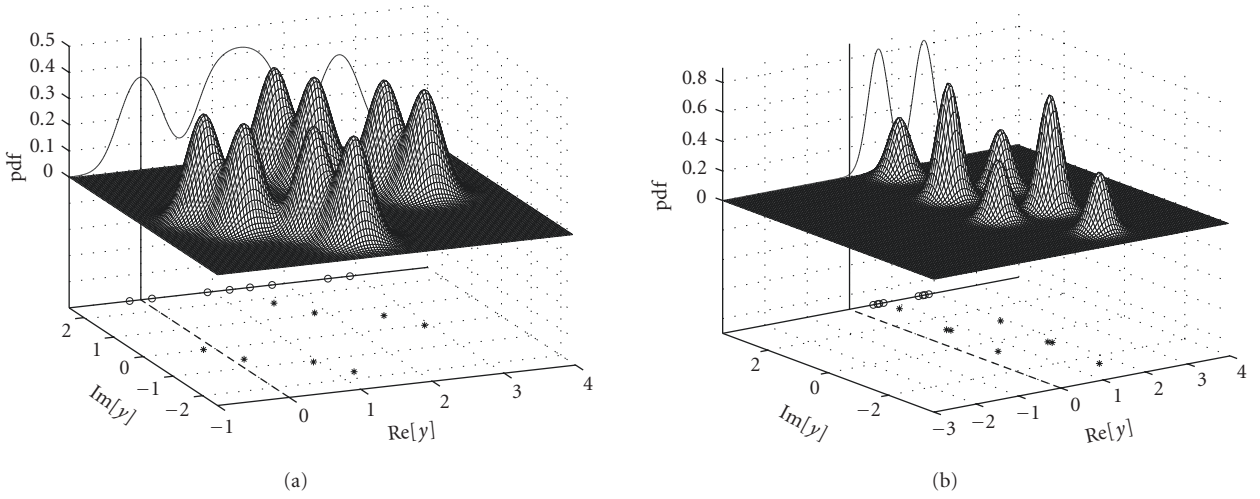

FIGURE 4: Conditional pdfs $p(y| + 1)$ (surface), conditional marginal pdfs $p(y_R| + 1)$ (curve), and conditional subsets $\mathcal{Y}^{(+)}$ (symbol $*$) and $\mathcal{Y}_R^{(+)}$ (symbol $\circ$), given $\theta = 30°$ and SNR = 10 dB. Beamformer weight vector has been normalized to a unit length. (a) LMMSE beamformer. (b) LMBER beamformer.

where

$$v_q = \frac{\text{sgn}(b_{q,1})}{N_b (2\pi\sigma_n^2)^L}. \tag{24}$$

Then the optimal decision (22) is equivalent to

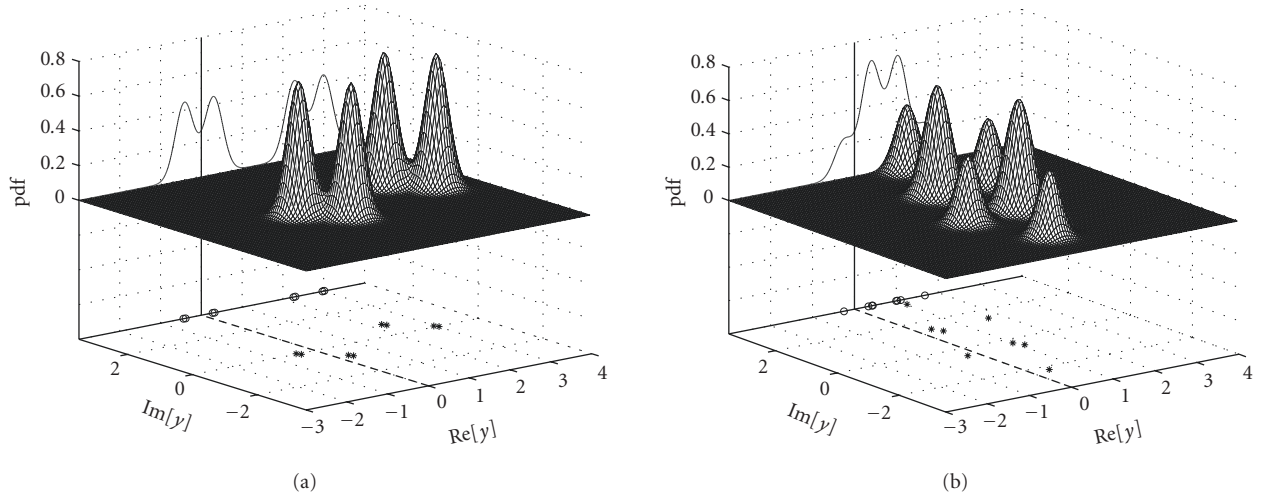$$\hat{b}_1(k) = \text{sgn}(y_B(k)) = \begin{cases} +1, & y_B(k) \geq 0, \\ -1, & y_B(k) < 0. \end{cases} \tag{25}$$

FIGURE 5: Conditional pdfs $p(y \mid +1)$ (surface), conditional marginal pdfs $p(y_R \mid +1)$ (curve), and conditional subsets $\mathcal{Y}^{(+)}$ (symbol $*$) and $\mathcal{Y}_R^{(+)}$ (symbol $\circ$), given $\theta = 10°$ and SNR = 10 dB. Beamformer weight vector has been normalized to a unit length. (a) LMMSE beamformer. (b) LMBER beamformer.

Note that (23) has the exact form of the RBF network in conjunction with a Gaussian kernel function.

The BER performance of the optimal Bayesian detection scheme was evaluated using the simulation example of Section 3 under the two conditions of having minimum spatial separations of $\theta = 30°$ and $\theta = 10°$, and the results are plotted in Figures 3a and 3b, respectively, in comparison to the BERs of linear beamformers. It can be seen from Figure 3a that the Bayesian detector achieved an SNR improvement of 4 dB at the BER of 10^{-4} over the LMBER beamformer. In the linearly inseparable case, the achievable performance improvement over the linear beamformer was even greater. In particular, Figure 3b shows that the Bayesian spatial processing assisted receiver removed the irreducible BER that was experienced by the linear beamforming aided receiver. The Bayesian detection scheme (23) may be viewed as a nonlinear "beamforming" process, and this nonlinear beamformer is clearly more complex than the simple linear beamformer (7). Therefore, the performance improvement achieved by the Bayesian detection scheme is attained at the expense of considerably increased computational complexity.

5. BLOCK-DATA KERNEL-BASED NONLINEAR DETECTOR CONSTRUCTION

In reality, the signal subsets $\mathcal{X}^{(\pm)}$ are unknown and have to be estimated in order to realize the Bayesian solution. We will adopt a temporal reference technique to construct a nonlinear detector. Given a block of N training data $\{\mathbf{x}(k), b_1(k)\}_{k=1}^N$, consider the nonlinear detector of the form

$$y(\mathbf{x}) = \sum_{l=1}^{N} \beta_l \phi_l(\mathbf{x}), \qquad (26)$$

where β_l represents the real-valued weights and $\phi_l(\mathbf{x}) = \phi(\mathbf{x}, \mathbf{x}(l))$ are the appropriately chosen kernel basis functions with $\mathbf{x}(l)$ denoting the lth training input. In our spatial processing aided application, $\phi(\cdot, \cdot)$ can be chosen as the Gaussian kernel function of the form

$$\phi(\mathbf{x}, \mathbf{x}(l)) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}(l)\|^2}{2\rho^2}\right), \qquad (27)$$

where the kernel variance ρ^2 is an estimate of the noise variance σ_n^2. Define the modelling residual as

$$\epsilon(k) = t(k) - y(k) = b_1(k) - y(\mathbf{x}(k)). \qquad (28)$$

Then the kernel model (26) generated for the training data set can be formulated as

$$\mathbf{t} = \mathbf{\Phi}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad (29)$$

where the target vector $\mathbf{t}$ is defined as

$$\begin{aligned}\mathbf{t} &= \begin{bmatrix} t(1) & t(2) & \cdots & t(N) \end{bmatrix}^T \\ &= \begin{bmatrix} b_1(1) & b_1(2) & \cdots & b_1(N) \end{bmatrix}^T, \end{aligned} \qquad (30)$$

the kernel weight vector is given by $\boldsymbol{\beta} = [\beta_1 \ \beta_2 \ \cdots \ \beta_N]^T$, the residual vector is formulated as $\boldsymbol{\epsilon} = [\epsilon(1) \ \epsilon(2) \ \cdots \ \epsilon(N)]^T$, and the regression matrix $\mathbf{\Phi}$ is given by

$$\mathbf{\Phi} = \begin{bmatrix} \boldsymbol{\phi}_1 & \boldsymbol{\phi}_2 & \cdots & \boldsymbol{\phi}_N \end{bmatrix} \qquad (31)$$

with

$$\begin{aligned}\boldsymbol{\phi}_i &= \begin{bmatrix} \phi_i(1) & \phi_i(2) & \cdots & \phi_i(N) \end{bmatrix}^T \\ &= \begin{bmatrix} \phi(\mathbf{x}(1), \mathbf{x}(i)) & \phi(\mathbf{x}(2), \mathbf{x}(i)) & \cdots & \phi(\mathbf{x}(N), \mathbf{x}(i)) \end{bmatrix}^T, \end{aligned} \qquad (32)$$

for $1 \leq i \leq N$. We adopt two different techniques for constructing a sparse detector model having $N_{\text{spa}}(\ll N)$ number of terms from the full model (26).

5.1. Relevance vector machine for sparse kernel detector construction

The RVM method [22, 23] can readily be applied for constructing a sparse kernel model having N_{spa} number of terms from the full model (26). The introduction of an individual hyperparameter α_i for every weight β_i of the model (26) is the key feature of the RVM, and is ultimately responsible for the sparsity properties of the RVM method [22]. During the optimization process, many of the α_i coefficients are driven to large values so that the corresponding model weights β_i are effectively pruned out. Thus the corresponding model terms $\phi_i(\cdot)$ can be removed from the trained model. The construction procedure produces a beamformer having a sparse final kernel structure consisting of N_{spa} number of significant terms. The detailed RVM method used is summarized in Appendix A.

The RVM method is known to be able to produce very sparse models while exhibiting excellent generalization capabilities [22]. A drawback of the RVM method is its high computational complexity. The algorithm contains two loops, with the inner loop used for updating the kernel weights and the outer loop for the associated hyperparameters (see Appendix A). Both loops involve "expensive" nonlinear optimization, and therefore converge relatively slowly, while incurring high computational costs. Furthermore, the RVM method starts with the full model set $\mathbf{\Phi}$ and removes those kernel terms that have large values in their associated hyperparameters. In other words, it is based on the backward elimination principle. Since the Hessian matrix $\mathbf{H}$ associated with the full model set ((A.8) in Appendix A) is typically ill-conditioned and may even be non invertible, the RVM method is inherently ill-conditioned and its iterative procedure may converge at a slow rate, requiring numerous iterations. The threshold Lg employed by the pruning process (see Appendix A) is problem-dependent and has to be determined empirically. Provided that the value of Lg is tuned appropriately, the RVM algorithm is in general capable of identifying a sparse detector from the full model (26), which closely approximates the Bayesian performance.

5.2. Orthogonal forward selection with Fisher ratio class-separability measure for sparse kernel detector construction

An alternative way of constructing a sparse kernel model from the full model (26) is offered by the OFS procedure based on Fisher ratio class-separability measure [24], which is computationally attractive and numerically very robust. Let an orthogonal decomposition of the regression matrix $\mathbf{\Phi}$ be

$$\mathbf{\Phi} = \mathbf{UD}, \tag{33}$$

where

$$\mathbf{D} = \begin{bmatrix} 1 & d_{1,2} & \cdots & d_{1,N} \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & d_{N-1,N} \\ 0 & \cdots & 0 & 1 \end{bmatrix},$$

$$\mathbf{U} = \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_N \end{bmatrix} = \begin{bmatrix} u_{1,1} & u_{1,2} & \cdots & u_{1,N} \\ u_{2,1} & u_{2,2} & \cdots & u_{2,N} \\ \vdots & \vdots & \vdots & \vdots \\ u_{N,1} & u_{N,2} & \cdots & u_{N,N} \end{bmatrix}, \tag{34}$$

with orthogonal columns that satisfy $\mathbf{u}_i^T \mathbf{u}_q = 0$ if $i \neq q$. The kernel model (29) can alternatively be expressed as

$$\mathbf{t} = \mathbf{Ug} + \boldsymbol{\epsilon}, \tag{35}$$

where the orthogonal weight vector $\mathbf{g} = \begin{bmatrix} g_1 & g_2 & \cdots & g_N \end{bmatrix}^T$ satisfies the triangular system $\mathbf{D\beta} = \mathbf{g}$.

A sparse N_{spa}-term model can be selected by incrementally maximizing a class separability measure in an OFS procedure, as is presented in [24]. Define the two class sets $\mathbf{X}_\pm = \{\mathbf{x}(k) : d(k) = \pm 1\}$, and let the numbers of points in $\mathbf{X}_\pm$ be $N_\pm$, respectively, with $N_+ + N_- = N$. The means and variances of training samples belonging to class $\mathbf{X}_+$ and class $\mathbf{X}_-$ in the direction of basis $\mathbf{u}_l$ are given by

$$m_{+,l} = \frac{1}{N_+} \sum_{i=1}^{N} \delta(t(i) - 1) u_{i,l},$$

$$\sigma_{+,l}^2 = \frac{1}{N_+} \sum_{i=1}^{N} \delta(t(i) - 1)(u_{i,l} - m_{+,l})^2,$$

$$m_{-,l} = \frac{1}{N_-} \sum_{i=1}^{N} \delta(t(i) + 1) u_{i,l}, \tag{36}$$

$$\sigma_{-,l}^2 = \frac{1}{N_-} \sum_{i=1}^{N} \delta(t(i) + 1)(u_{i,l} - m_{-,l})^2,$$

respectively, where $\delta(x) = 1$ for $x = 0$ and $\delta(x) = 0$ for $x \neq 0$. Fisher ratio is defined as the ratio of the interclass difference and the intraclass spread encountered in the direction of $\mathbf{u}_l$, which is given by [28]

$$F_l = \frac{(m_{+,l} - m_{-,l})^2}{\sigma_{+,l}^2 + \sigma_{-,l}^2}. \tag{37}$$

Based on this Fisher ratio for class separability measure, significant kernel terms can be selected with the aid of an OFS procedure. At the lth stage, a term is chosen as the lth term in the selected model if it produces the largest F_l among the candidate terms $\mathbf{u}_i$, $l \leq i \leq N$. The procedure is terminated with a sparse N_{spa}-term model when we have

$$\frac{F_{N_{\text{spa}}}}{\sum_{l=1}^{N_{\text{spa}}} F_l} < \xi, \tag{38}$$
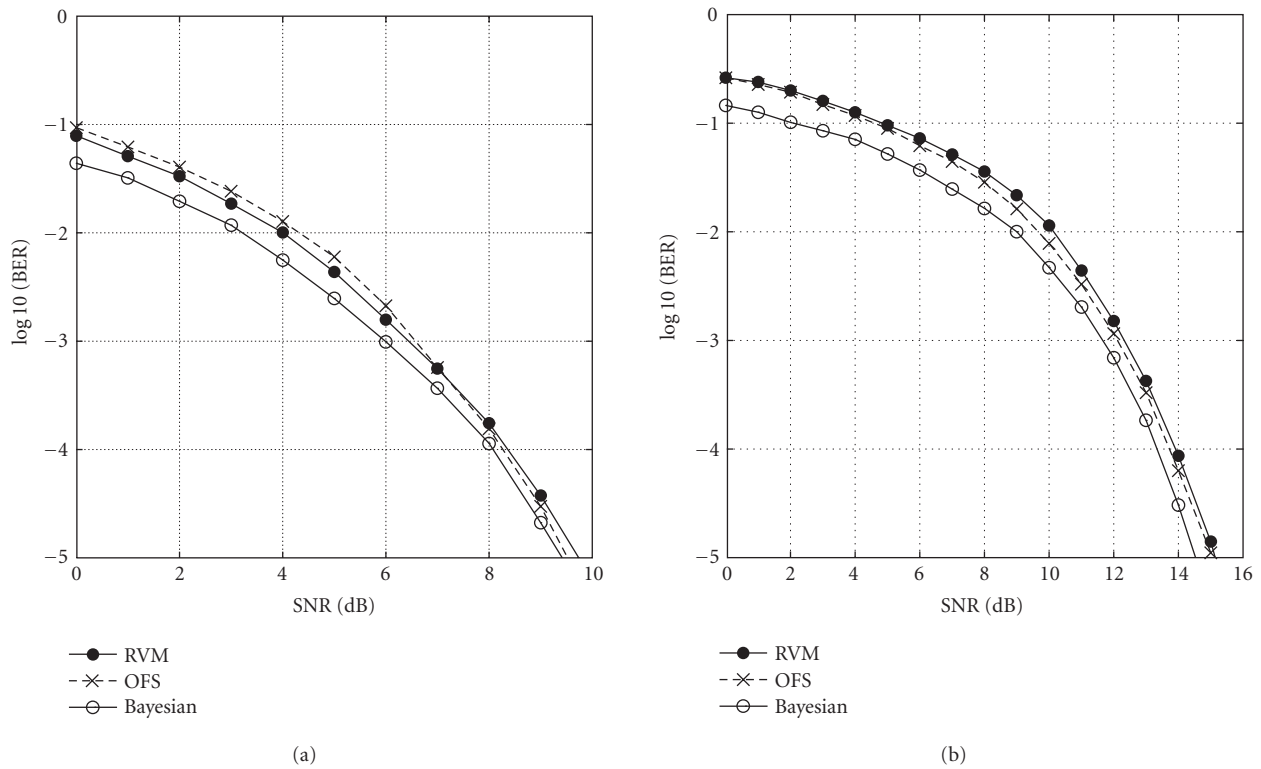
(a)

(b)

FIGURE 6: Performance comparison of the Bayesian detector with the RBF detectors constructed by the RVM algorithm and the OFS with Fisher ratio, respectively. (a) $\theta = 30°$. (b) $\theta = 10°$.

where the threshold ξ determines the sparsity of the selected model. The appropriate value for ξ depends on the application concerned, and in our spatial processing oriented application, we have found out empirically that the appropriate values for ξ is in the range of 0.005 to 0.01. The least square solution for the corresponding sparse model weight vector $\boldsymbol{\beta}_{N_{\mathrm{spa}}}$ is readily available given the least square solution of $\mathbf{g}_{N_{\mathrm{spa}}}$.

The detailed construction algorithm is summarized in Appendix B. This algorithm involves only linear optimization and is computationally significantly more attractive compared with the RVM method. In the selection procedure, if $\mathbf{u}_i^T \mathbf{u}_i$ is too small, this term will not be selected. Thus, any ill-conditioning problem or singular situations are automatically avoided. The construction process is guaranteed to converge and, to arrive at the sparsest possible kernel detector that is also capable of closely approximating the optimum Bayesian performance, the only algorithmic parameter that requires tuning is the threshold ξ.

5.3. Simulation study

The example given in Section 3 was used for testing the two block-data kernel-based construction algorithms. Two conditions of $\theta = 30°$ and $\theta = 10°$ were simulated, representing the linearly separable and inseparable cases, respectively. In each case, the OFS algorithm employing the Fisher ratio and the RVM algorithm were used for constructing a RBF detec-

tor. The number of training data used for each SNR value was $N = 160$. The Gaussian kernel variance ρ^2 was determined empirically and the appropriate values of ρ^2 were found to be in the range spanning from $2\sigma_n^2$ to $10\sigma_n^2$, depending on the SNR. The number of RBF centers or kernel terms identified by the two algorithms for the given SNR values was similar, ranging from $N_{\mathrm{spa}} = 14$ to 20, having typical values of $N_{\mathrm{spa}} = 18$. The BERs of the RVM and OFS detectors are compared in Figure 6. It can be seen that both kernel-based detectors had a similar performance at a similar model sparsity, and the two RBF detectors constructed from noisy training data closely approximated the optimal Bayesian performance. However, the OFS algorithm based on the Fisher ratio is known to have considerable computational and numerical advantages over the RVM algorithm.

6. RECURSIVE ADAPTIVE RBF DETECTOR USING THE COMBINED CLUSTERING AND RLS ALGORITHM

In practice, it is often desirable to update a detector on a recursive sample-by-sample basis. Consider again the RBF detector of the form

$$y(\mathbf{x}(k)) = \sum_{i=1}^{N_c} \beta_i \phi(\mathbf{x}(k), \mathbf{c}_i), \qquad (39)$$

where $\mathbf{c}_i$ are the complex-valued kernel centers and the number of kernel centers N_c is assumed to be given. We propose

to apply a combined enhanced κ-means clustering and RLS algorithm [19, 25] for a recursive sample-by-sample based adaptation of this RBF detector.

The enhanced κ-means clustering algorithm [29], which recursively updates the RBF centers, is described by

$$\mathbf{c}_i(k) = \mathbf{c}_i(k-1) + \mathcal{M}_i(\mathbf{x}(k))\left(\bar{g}_c\left(\mathbf{x}(k) - \mathbf{c}_i(k-1)\right)\right) \quad (40)$$

for $1 \leq i \leq N_c$, where $0 < \bar{g}_c < 1.0$ defines the learning rate, the membership function $\mathcal{M}_i(\mathbf{x}(k))$ is defined as follows:

$$\mathcal{M}_i(\mathbf{x}) = \begin{cases} 1, & \text{if } \bar{v}_i\|\mathbf{x} - \mathbf{c}_i\|^2 \leq \bar{v}_l\|\mathbf{x} - \mathbf{c}_l\|^2 \; \forall l \neq i, \\ 0, & \text{otherwise}, \end{cases} \quad (41)$$

and $\bar{v}_i$ is the variation of the ith cluster. In order to estimate the associated variation $\bar{v}_i$, the following updating rule is used:

$$\begin{aligned} \bar{v}_i(k) = &\; \bar{g}_v \bar{v}_i(k-1) \\ &+ (1 - \bar{g}_v)\left(\mathcal{M}_i(\mathbf{x}(k))\|\mathbf{x}(k) - \mathbf{c}_i(k-1)\|^2\right), \end{aligned} \quad (42)$$

where $\bar{g}_v$ is a constant slightly less than 1.0. The initial variations $\bar{v}_i(0)$, $1 \leq i \leq N_c$, are set to the same small number. The learning rate $\bar{g}_c$ can either be set to a fixed small positive number or be self-adjusting, based on an entropy formula [29].

The traditional κ-means clustering algorithm [28] can only achieve a local optimal solution in partitioning the input data set into N_c clusters, and the solution obtained depends on the initial locations of cluster centers. A consequence of this local optimality is that some initial centers may become trapped in regions of the input domain, which have only a few or no input patterns, and never move to regions where they are needed. This wastes resources and results in an unnecessarily large network. The enhanced κ-means clustering algorithm [29] overcomes the above-mentioned drawback. When using a cluster variation-weighted measure, we always achieve an optimal center configuration in the sense that after convergence, all clusters have an equal cluster variance. The above-mentioned enhanced κ-means clustering algorithm is an unsupervised one. In order to take full advantage of training, the algorithm can be modified in order to create a semisupervised one. Let the RBF center set be divided into the two subsets

$$\begin{aligned} \mathcal{C}^{(+)} &= \left\{\mathbf{c}_i, \; 1 \leq i \leq \frac{N_c}{2}\right\}, \\ \mathcal{C}^{(-)} &= \left\{\mathbf{c}_i, \; 1 + \frac{N_c}{2} \leq i \leq N_c\right\}, \end{aligned} \quad (43)$$

corresponding to the two classes $b_1(k) = \pm 1$. During the training instance k, the enhanced κ-means clustering algorithm is applied only to the center subset $\mathcal{C}^{(+)}$ if we have $b_1(k) = +1$. Otherwise, it is applied to $\mathcal{C}^{(-)}$ provided that we have $b_1(k) = -1$. This "semisupervised" clustering technique was found to be more effective in dealing with linearly inseparable cases.

The RBF weights β_i are updated using the classic RLS algorithm. Thus the combined CRLS algorithm used for training the RBF detector (39) can readily be summa-

rized as follows. At the instance k, given the center set $\{\mathbf{c}_i(k-1), \; 1 \leq i \leq N_c\}$ and weight vector $\boldsymbol{\beta}(k-1) = [\beta_1(k-1) \; \beta_2(k-1) \; \cdots \; \beta_{N_c}(k-1)]^T$, we invoke the following procedure:

RBF center updating: use the enhanced κ-means clustering algorithm for obtaining an updated RBF center set $\{\mathbf{c}_i(k), \; 1 \leq i \leq N_c\}$;

RBF weight updating: employ the RLS algorithm for obtaining an updated RBF weight vector $\boldsymbol{\beta}(k)$.

The enhanced κ-means clustering process is guaranteed to converge to the optimal center configuration if either the learning rate $\bar{g}_c$ is self-adjusting based on an entropy formula or it is fixed to a positive constant that is not too large [29]. The convergence properties of the standard RLS algorithm are well known. It is therefore reasonable to believe that the above-mentioned combined κ-means clustering and RLS algorithm is capable of guaranteeing convergence, provided that the algorithmic parameters are set appropriately.

The example given in Section 3 was employed again for investigating the CRLS algorithm used for training the RBF detector of (39). Two conditions associated with $\theta = 30°$ and $\theta = 10°$ were simulated. For this example, the number of states that defined the Bayesian detector was $N_b = 16$, and $N_c = 16$ was assumed for the RBF detector. The training data length was $N = 1000$. The first N_c number of samples $\mathbf{x}(k)$ were used as the initial RBF centers and the two adaptive parameters of the clustering algorithm were set to $\bar{g}_c = 0.2$ and $\bar{g}_v = 0.995$. Half of the RBF weights were set initially to $+0.001$ and the other half to -0.001. The initial condition of the RLS algorithm was chosen as $\boldsymbol{\Psi}(0) = \text{diag}\{1000.0, 1000.0, \dots, 1000.0\}$ with the forgetting factor given by $\mu = 0.995$. Figure 7 depicts the achievable BER of the CRLS RBF detector in comparison to the optimal Bayesian performance. For the CRLS RBF detector, the results obtained using the unsupervised and semisupervised clustering algorithms were similar in the linearly separable case ($\theta = 30°$). By contrast, for the linearly inseparable scenario of $\theta = 10°$, it was observed that the semisupervised clustering performed better than the unsupervised one. The results given in Figure 7 are those obtained with the aid of semi-supervised clustering. From Figure 7, it can be seen that the performance of the CRLS RBF detector closely matched the optimal Bayesian performance.

7. CONCLUSIONS AND DISCUSSIONS

A nonlinear detection technique has been investigated in the context of a multiantenna assisted receiver. The optimal solution of the nonlinear spatial processing aided receiver has been derived for binary phase shift keying signalling. It has been shown that this optimal Bayesian detector significantly outperforms the linear beamformer in terms of a reduced bit error rate, at the expense of an increased complexity. The results presented in this paper have demonstrated the potential system capacity enhancements that may be achieved by employing nonlinear spatial processing. Both block-data-based and recursive sample-by-sample adaptive implementations
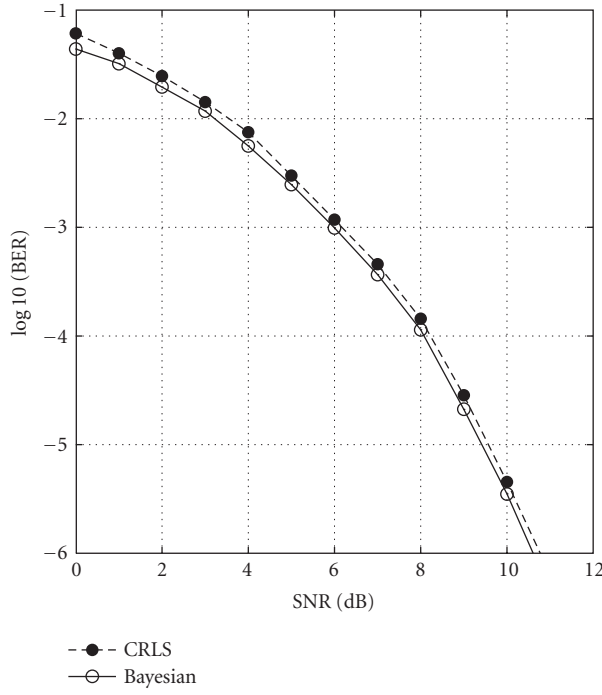
(a)
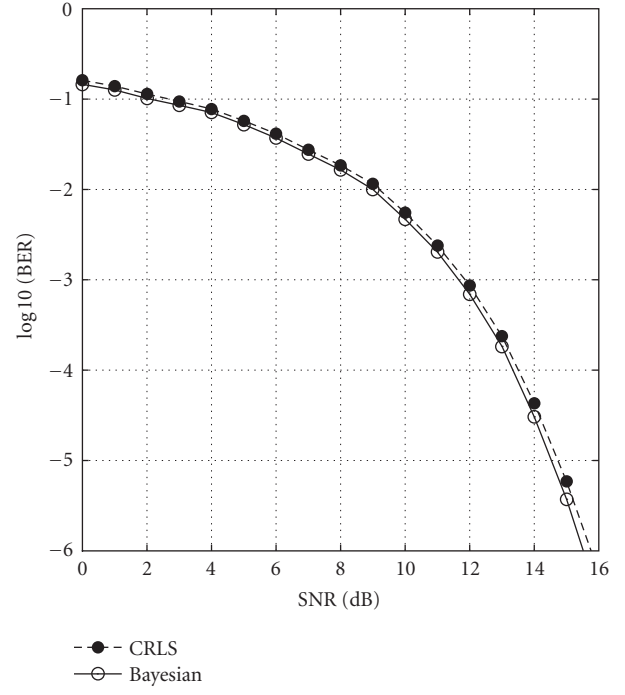
(b)

FIGURE 7: Performance comparison of the Bayesian detector with the RBF detector trained by the CRLS algorithm. (a) $\theta = 30°$. (b) $\theta = 10°$.

of the optimal Bayesian detector have been considered using a radial basis function network. For block-data-based adaptation, both the RVM algorithm and the orthogonal forward selection procedure employing the Fisher-ratio-based class separability measure have been considered. Both algorithms have been shown to produce similarly good performance, but the latter is known to have considerable computational advantages. For recursive sample-by-sample based adaptation, the combination of the enhanced κ-means clustering and the recursive least squares algorithm has been invoked.

The nonlinear detection scheme proposed in this paper is based on what we refer to as a "direct" approach, namely, on estimating the RBF centers directly from received training data contaminated by the channel. Alternatively, an "indirect" approach can be adopted, where the system matrix **P** defined in (4) is first identified and then used for constructing the nonlinear detector. This indirect approach has the advantage of requiring a significantly shorter training time, since estimating the channel matrix needs a shorter training sequence than estimating the noiseless channel states that define RBF centers. This indirect approach is not applicable in the SDMA assisted multiuser downlink, since the receiver in this case only has access to the one desired user's training sequence. However, this indirect scheme becomes attractive in the uplink, as the receiver has to detect all the users' data and has access to the training sequences of all the users. Moreover, numerous complexity-reduction schemes can be adopted for the RBF detector [21]. Indeed, it was demonstrated in [21] that the complexity of the RBF detector may be rendered comparable to that of classic linear detectors. For example,

decision feedback can be employed not only to improve the performance significantly but also to reduce the complexity dramatically of the RBF detector, similar to the case of single-user channel equalization [18, 30]. This nonlinear detection scheme designed for the SDMA assisted multiuser uplink is currently under investigation.

APPENDICES

A. RELEVANCE VECTOR MACHINE METHOD

The posterior probability of the kernel detector weight vector $\boldsymbol{\beta}$ is defined by

$$p(\boldsymbol{\beta}|\mathbf{t}, \boldsymbol{\alpha}) = \frac{p(\mathbf{t}|\boldsymbol{\beta}, \boldsymbol{\alpha})p(\boldsymbol{\beta}|\boldsymbol{\alpha})}{p(\mathbf{t}|\boldsymbol{\alpha})}, \tag{A.1}$$

where $p(\boldsymbol{\beta}|\boldsymbol{\alpha})$ is the prior with $\boldsymbol{\alpha} = [\alpha_1 \ \alpha_2 \ \cdots \ \alpha_N]^T$ denoting the vector of hyperparameters, $p(\mathbf{t}|\boldsymbol{\beta}, \boldsymbol{\alpha})$ is the likelihood, and $p(\mathbf{t}|\boldsymbol{\alpha})$ the evidence. Following the Bayesian classification framework [22, 23], the likelihood is expressed as

$$p(\mathbf{t}|\boldsymbol{\beta}, \boldsymbol{\alpha})$$
$$= \prod_{l=1}^{N} \big(f\big(y(\mathbf{x}(l))\big)\big)^{(t(l)+1)/2} \big(1 - f\big(y(\mathbf{x}(l))\big)\big)^{(1-t(l))/2}, \tag{A.2}$$

where

$$f(y) = \frac{1}{1 + \exp(-y)} \tag{A.3}$$

is the logistic sigmoid function. The Gaussian prior is chosen:

$$p(\boldsymbol{\beta}|\boldsymbol{\alpha}) = \prod_{l=1}^{N} \frac{\sqrt{\alpha_l}}{\sqrt{2\pi}} \exp\left(-\frac{\alpha_l\beta_l^2}{2}\right). \tag{A.4}$$

As the marginal likelihood $p(\mathbf{t}|\boldsymbol{\alpha})$ cannot be obtained analytically by integrating out the weights from (A.1), an iterative procedure is necessitated [22].

With a fixed given $\boldsymbol{\alpha}$, the maximum a posteriori probability (MAP) solution $\hat{\boldsymbol{\beta}}$ can be obtained by maximizing $\log(p(\boldsymbol{\beta}|\mathbf{t},\boldsymbol{\alpha}))$ or, equivalently, by minimizing the following cost function:

$$J(\boldsymbol{\beta}|\mathbf{t},\boldsymbol{\alpha}) = \sum_{l=1}^{N} \left(\frac{\alpha_l\beta_l^2}{2} - \frac{t(l)+1}{2}\log\left(f(y(\mathbf{x}(l)))\right)\right.$$
$$\left. - \frac{1-t(l)}{2}\log\left(1 - f(y(\mathbf{x}(l)))\right)\right). \tag{A.5}$$

The gradient of J with respect to $\boldsymbol{\beta}$ is

$$\nabla J = \mathbf{A}\boldsymbol{\beta} + \boldsymbol{\Phi}^T\left(\mathbf{f} - \frac{1}{2}(\mathbf{t}+\mathbf{1}_N)\right), \tag{A.6}$$

where

$$\mathbf{A} = \mathrm{diag}\{\alpha_1, \alpha_2, \dots, \alpha_N\},$$
$$\mathbf{f} = \begin{bmatrix} f(y(\mathbf{x}(1))) & f(y(\mathbf{x}(2))) & \cdots & f(y(\mathbf{x}(N))) \end{bmatrix}^T,$$
$$\mathbf{1}_N = \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}^T, \tag{A.7}$$

and $\boldsymbol{\Phi}$ is the regression matrix defined in (31). The Hessian of J is

$$\mathbf{H} = \nabla^2 J = \boldsymbol{\Phi}^T\mathbf{B}\boldsymbol{\Phi} + \mathbf{A}, \tag{A.8}$$

where

$$\mathbf{B} = \mathrm{diag}\{f(y(\mathbf{x}(1)))(1 - f(y(\mathbf{x}(1)))), \dots,$$
$$f(y(\mathbf{x}(N)))(1 - f(y(\mathbf{x}(N))))\}. \tag{A.9}$$

The hyperparameters $\boldsymbol{\alpha}$ are updated using

$$\alpha_i^{\mathrm{new}} = \frac{1 - \alpha_i^{\mathrm{old}}\gamma_{i,i}}{\hat{\beta}_i^2} \tag{A.10}$$

with $\gamma_{i,i}$ being the diagonal elements of $\boldsymbol{\Gamma}$ which is defined by

$$\boldsymbol{\Gamma} = \left(\mathbf{H}|_{\hat{\boldsymbol{\beta}}}\right)^{-1}. \tag{A.11}$$

The following simple iterative procedure can be adopted to construct a sparse RVM detector.

Initialization

The $N \times N_{\mathrm{spa}}$ kernel matrix $\boldsymbol{\Phi}$ is initialized with $N_{\mathrm{spa}} = N$, that is, every training data point is considered as a candidate kernel. Each weight β_i is initially associated with the same value of the hyperparameter α_i.

Step 1. Given current value $\boldsymbol{\alpha}$, find $\hat{\boldsymbol{\beta}}$ by minimizing the cost function (A.5). A simplified conjugate gradient algorithm [26, 27] is used in our application.

Step 2. The hyperparameters are updated using (A.10). If a $\alpha_i > Lg$, where Lg is a preset large positive value, $N_{\mathrm{spa}} := N_{\mathrm{spa}} - 1$, the corresponding column in $\boldsymbol{\Phi}$ is removed, and thus the corresponding weight β_i and model term $\phi_i(\cdot)$ are pruned out the model.

Test

If the hyperparameters $\boldsymbol{\alpha}$ remain sufficiently unchanged in two successive iterations (no removal of hyperparameters) or a preset maximum iteration number is reached, stop; otherwise, go to Step 1.

B. ORTHOGONAL FORWARD SELECTION ALGORITHM

The modified Gram-Schmidt orthogonalization procedure [31] calculates the $\mathbf{D}$ matrix row by row and orthogonalizes $\boldsymbol{\Phi}$ as follows: at the lth stage, make the columns $\boldsymbol{\phi}_i$, $l+1 \leq i \leq N$, orthogonal to the lth column and repeat the operation for $1 \leq l \leq N-1$. Specifically, denoting $\boldsymbol{\phi}_i^{(0)} = \boldsymbol{\phi}_i$, $1 \leq i \leq N$, then

$$\mathbf{u}_l = \boldsymbol{\phi}_l^{(l-1)},$$
$$d_{l,i} = \frac{\mathbf{u}_l^T\boldsymbol{\phi}_i^{(l-1)}}{(\mathbf{u}_l^T\mathbf{u}_l)}, \quad l+1 \leq i \leq N, \tag{B.1}$$
$$\boldsymbol{\phi}_i^{(l)} = \boldsymbol{\phi}_i^{(l-1)} - d_{l,i}\mathbf{u}_l, \quad l+1 \leq i \leq N,$$

where $l = 1, 2, \dots, N-1$.

The last stage of the procedure is simply $\mathbf{u}_N = \boldsymbol{\phi}_N^{(N-1)}$. The elements of $\mathbf{g}$ are computed by transforming $\mathbf{t}^{(0)} = \mathbf{t}$ in a similar way:

$$g_l = \frac{\mathbf{u}_l^T\mathbf{t}^{(l-1)}}{\left(\mathbf{u}_l^T\mathbf{u}_l\right)},$$
$$\mathbf{t}^{(l)} = \mathbf{t}^{(l-1)} - g_l\mathbf{u}_l, \tag{B.2}$$

where $1 \leq l \leq N$.

This orthogonalization scheme can be used to derive a simple and efficient algorithm for selecting subset models in a forward-regression manner [31]. First define

$$\boldsymbol{\Phi}^{(l-1)} = \begin{bmatrix} \mathbf{u}_1 \cdots \mathbf{u}_{l-1} & \boldsymbol{\phi}_l^{(l-1)} \cdots \boldsymbol{\phi}_N^{(l-1)} \end{bmatrix}. \tag{B.3}$$

If some of the columns $\boldsymbol{\phi}_l^{(l-1)}, \dots, \boldsymbol{\phi}_N^{(l-1)}$ in $\boldsymbol{\Phi}^{(l-1)}$ have been interchanged, this will still be referred to as $\boldsymbol{\Phi}^{(l-1)}$ for notational convenience. With the notation $\boldsymbol{\phi}_q^{(l-1)} = [\phi_{1,q}^{(l-1)} \ \phi_{2,q}^{(l-1)} \ \cdots \ \phi_{N,q}^{(l-1)}]^T$, the lth stage of the selection procedure is given as follows.

Step 1. For $l \leq q \leq N$, compute

$$m_{+,l}^{(q)} = \frac{1}{N_+} \sum_{i=1}^{N} \delta(t(i) - 1) \phi_{i,q}^{(l-1)},$$

$$\left(\sigma_{+,l}^{(q)}\right)^2 = \frac{1}{N_+} \sum_{i=1}^{N} \delta(t(i) - 1) \left(\phi_{i,q}^{(l-1)} - m_{+,l}^{(q)}\right)^2,$$

$$m_{-,l}^{(q)} = \frac{1}{N_-} \sum_{i=1}^{N} \delta(t(i) + 1) \phi_{i,q}^{(l-1)}, \quad \text{(B.4)}$$

$$\left(\sigma_{-,l}^{(q)}\right)^2 = \frac{1}{N_-} \sum_{i=1}^{N} \delta(t(i) + 1) \left(\phi_{i,q}^{(l-1)} - m_{-,l}^{(q)}\right)^2,$$

$$F_l^{(q)} = \frac{\left(m_{+,l}^{(q)} - m_{-,l}^{(q)}\right)^2}{\left(\sigma_{+,l}^{(q)}\right)^2 + \left(\sigma_{-,l}^{(q)}\right)^2}.$$

Step 2. Find

$$F_l = F_l^{(q_l)} = \max\{F_l^{(q)}, l \leq q \leq N\}. \quad \text{(B.5)}$$

Then the q_lth column of $\mathbf{\Phi}^{(l-1)}$ is interchanged with the lth column of $\mathbf{\Phi}^{(l-1)}$, and the q_lth column of $\mathbf{D}$ is interchanged with the lth column of $\mathbf{D}$ up to the $(l-1)$th row. This effectively selects the q_lth candidate as the lth kernel term in the subset model.

Step 3. Perform the orthogonalization as indicated in (B.1) to derive the lth row of $\mathbf{D}$ and to transform $\mathbf{\Phi}^{(l-1)}$ into $\mathbf{\Phi}^{(l)}$. Calculate g_l and update $\mathbf{t}^{(l-1)}$ into $\mathbf{t}^{(l)}$ in the way shown in (B.2).

The selection is terminated at the N_{spa} stage when the criterion (38) is satisfied and this produces a sparse subset model containing N_{spa} significant kernel terms.

REFERENCES

[1] J. H. Winters, J. Salz, and R. D. Gitlin, "The impact of antenna diversity on the capacity of wireless communication systems," *IEEE Trans. Communications*, vol. 42, no. 2/3/4, pp. 1740–1751, 1994.

[2] J. Litva and T. K. Y. Lo, *Digital Beamforming in Wireless Communications*, Artech House, London, UK, 1996.

[3] L. C. Godara, "Applications of antenna arrays to mobile communications. I. Performance improvement, feasibility, and system considerations," *Proceedings of the IEEE*, vol. 85, no. 7, pp. 1031–1060, 1997.

[4] A. J. Paulraj and C. B. Papadias, "Space-time processing for wireless communications," *IEEE Signal Processing Magazine*, vol. 14, no. 6, pp. 49–83, 1997.

[5] J. H. Winters, "Smart antennas for wireless systems," *IEEE Personal Communications*, vol. 5, no. 1, pp. 23–27, 1998.

[6] R. Kohno, "Spatial and temporal communication theory using adaptive antenna array," *IEEE Personal Communications*, vol. 5, no. 1, pp. 28–35, 1998.

[7] J. S. Blogh and L. Hanzo, *Third Generation Systems and Intelligent Wireless Networking: Smart Antenna and Adaptive Modulation*, John Wiley & Sons, Chichester, UK, 2002.

[8] R. A. Soni, R. M. Buehrer, and R. D. Benning, "Intelligent antenna system for cdma2000," *IEEE Signal Processing Magazine*, vol. 19, no. 4, pp. 54–67, 2002.

[9] I. S. Reed, J. D. Mallett, and L. E. Brennan, "Rapid convergence rate in adaptive arrays," *IEEE Trans. on Aerospace and Electronics Systems*, vol. 10, no. 6, pp. 853–863, 1974.

[10] M. W. Ganz, R. L. Moses, and S. L. Wilson, "Convergence of the SMI and the diagonally loaded SMI algorithms with weak interference [adaptive array]," *IEEE Trans. Antennas and Propagation*, vol. 38, no. 3, pp. 394–399, 1990.

[11] B. Widrow, P. E. Mantey, L. J. Griffiths, and B. B. Goode, "Adaptive antenna systems," *Proceedings of the IEEE*, vol. 55, no. 12, pp. 2143–2159, 1967.

[12] L. J. Griffiths, "A simple adaptive algorithm for real-time processing in antenna arrays," *Proceedings of the IEEE*, vol. 57, no. 10, pp. 1696–1704, 1969.

[13] S. Haykin, *Adaptive Filter Theory*, Prentice-Hall, Upper Saddle River, NJ, USA, 3rd edition, 1996.

[14] S. Chen, L. Hanzo, and N. N. Ahmad, "Adaptive minimum bit error rate beamforming assisted receiver for wireless communications," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 4, pp. 640–643, Hong Kong, April 2003.

[15] S. Chen, N. N. Ahmad, and L. Hanzo, "Adaptive minimum bit error rate beamforming," to appear in *IEEE Trans. Wireless Communications*.

[16] S. Chen, G. J. Gibson, C. F. N. Cowan, and P. M. Grant, "Reconstruction of binary signals using an adaptive radial-basis-function equalizer," *Signal Processing*, vol. 22, no. 1, pp. 77–93, 1991.

[17] S. Chen, B. Mulgrew, and P. M. Grant, "A clustering technique for digital communications channel equalization using radial basis function networks," *IEEE Transactions on Neural Networks*, vol. 4, no. 4, pp. 570–590, 1993.

[18] S. Chen, B. Mulgrew, and S. McLaughlin, "Adaptive Bayesian equalizer with decision feedback," *IEEE Trans. Signal Processing*, vol. 41, no. 9, pp. 2918–2927, 1993.

[19] S. Chen, A. K. Samingan, and L. Hanzo, "Support vector machine multiuser receiver for DS-CDMA signals in multipath channels," *IEEE Transactions on Neural Networks*, vol. 12, no. 3, pp. 604–611, 2001.

[20] S. Chen and L. Hanzo, "Block-adaptive kernel-based CDMA multiuser detection," in *Proc. IEEE International Conference on Communications*, vol. 2, pp. 682–686, New York, NY, USA, April–May 2002.

[21] L. Hanzo, C. H. Wong, and M. S. Yee, *Adaptive Wireless Transceivers: Turbo-Coded, Turbo-Equalized and Space-Time Coded TDMA, CDMA, and OFDM Systems*, John Wiley & Sons, New York, NY, USA, 2002.

[22] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, vol. 1, pp. 211–244, 2001.

[23] S. Chen, S. R. Gunn, and C. J. Harris, "The relevance vector machine technique for channel equalization application," *IEEE Transactions on Neural Networks*, vol. 12, no. 6, pp. 1529–1532, 2001.

[24] K. Z. Mao, "RBF neural network center selection based on Fisher ratio class separability measure," *IEEE Transactions on Neural Networks*, vol. 13, no. 5, pp. 1211–1217, 2002.

[25] S. Chen, "Nonlinear time series modelling and prediction using Gaussian RBF networks with enhanced clustering and RLS learning," *Electronics Letters*, vol. 31, no. 2, pp. 117–118, 1995.

[26] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear Programming: Theory and Algorithms*, John Wiley & Sons, New York, NY, USA, 2nd edition, 1993.

[27] S. Chen, A. K. Samingan, B. Mulgrew, and L. Hanzo, "Adaptive minimum-BER linear multiuser detection for DS-CDMA

signals in multipath channels," *IEEE Trans. Signal Processing*, vol. 49, no. 6, pp. 1240–1247, 2001.

[28] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, John Wiley & Sons, New York, NY, USA, 1973.

[29] C. Chinrungrueng and C. H. Sequin, "Optimal adaptive κ-means algorithm with dynamic adjustment of learning rate," *IEEE Transactions on Neural Networks*, vol. 6, no. 1, pp. 157–169, 1995.

[30] S. Chen, S. McLaughlin, B. Mulgrew, and P. M. Grant, "Bayesian decision feedback equaliser for overcoming co-channel interference," *IEE Proceedings-Communication*, vol. 143, no. 4, pp. 219–225, 1996.

[31] S. Chen, S. A. Billings, and W. Luo, "Orthogonal least squares methods and their application to non-linear system identification," *International Journal of Control*, vol. 50, no. 5, pp. 1873–1896, 1989.

Sheng Chen obtained his B.Eng. degree in control engineering from the East China Petroleum Institute, China, in 1982, and his Ph.D. degree in control engineering from the City University at London in 1986. He joined the Department of Electronics and Computer Science at the University of Southampton, UK, in September 1999. He has previously held research and academic appointments at the Universities of She eld, Edinburgh, and Portsmouth, UK. Dr. Chen is a Senior Member of IEEE. His recent research works include adaptive nonlinear signal processing, modeling and identification of nonlinear systems, neural network research, finite-precision digital controller design, and evolutionary computation methods and optimization. He has published over 200 research papers.

Lajos Hanzo received his degree in electronics in 1976 and his doctorate in 1983. During his career in telecommunications, he has held various research and academic posts in Hungary, Germany, and the UK. Since 1986, he has been with the Department of Electronics and Computer Science, University of Southampton, UK, where he holds the Chair in telecommunications. He coauthored 10 books totalling 8000 pages on mobile radio communications, published about 450 research papers, organized and chaired conference sessions, presented overview lectures, and has been awarded a number of distinctions. Currently he heads an academic research team, working on a range of research projects in the field of wireless multimedia communications sponsored by industry, the Engineering and Physical Sciences Research Council (EPSRC), UK, the European IST Programme, and the Mobile Virtual Centre of Excellence (VCE), UK. He is an enthusiastic supporter of industrial and academic liaison and he o ers a range of industrial courses. L. Hanzo is also an IEEE Distinguished Lecturer at both the Communications as well as the Vehicular Technology Society, a Fellow of the IEE, and a Fellow of the IEEE.

Andreas Wolfgang received his Dipl.-Ing. degree in electrical engineering from Karlsruhe University of Technology, Germany, in 2003. Currently he is with the Communications Research Group at the University of Southampton, UK, where he is pursuing the Ph.D. degree. He was a member of the Antenna Group at Chalmers University, Gothenburg, Sweden, where he worked for the development of measurement methods for terminal antennas. His current research interests are in wireless communications with emphasis on nonlinear filter structures designed for multiple antenna systems.