Generalization Error of Linear Discriminant Analysis in Spatially-Correlated Sensor Networks

Kush R. Varshney

***Abstract*—Generalization error, the probability of error of a detection rule learned from training samples on new unseen samples, is a fundamental quantity to be characterized. However, characterizations of generalization error in the statistical learning theory literature are often loose and practically unusable for optimizing detection systems. In this work, focusing on learning linear discriminant analysis detection rules from spatially-correlated sensor measurements, a tight generalization error approximation is developed that can be used to optimize the parameters of a sensor network detection system. As such, the approximation is used to optimize network settings. The approximation is also used to derive a detection error exponent and select an optimal subset of deployed sensor nodes. A Gauss–Markov random field is used to model correlation and weak laws of large numbers in geometric probability are employed in the analysis.**

***Index Terms*—Distributed sensors, generalization error, geometric probability, linear discriminant analysis, signal detection.**

I. Introduction

Collections of spatially-distributed sensor nodes are now being used widely in environmental monitoring and surveillance. With the increasing penetration of smartphones loaded with a multitude of sensors, many new sensor network application domains are poised to emerge. In these domains, one of the key signal processing tasks is detection or classification. Sensor measurements in natural settings tend to exhibit spatial correlation [1]–[6], but more often than not, the design and analysis of detection rules in sensor networks is studied under the assumption of statistically independent measurements (conditioned on the hypothesis) [7]–[10]. Also, the focus of sensor network detection investigations is mostly on the case when the likelihood functions of the measurements are known. However, this *a priori* knowledge is often not available in real-world situations. In contrast, in this paper, the statistical learning framework of supervised classification is considered, in which training samples are provided rather than the likelihood functions, and sensor measurements are assumed spatially-correlated [11].

Two different errors may be examined in supervised classification: the training or empirical error and the generalization error. The training error, the fraction of misclassifications by the learned detection rule on the provided training samples, is of much less relevance than the generalization error, the probability of incorrect classification by the detection rule on a new sample drawn from the same distribution as the training samples. The training error is straightforward to report, but characterizing the generalization error of a learned detection rule is not; doing so is a main thrust in statistical learning theory.

Bounds on generalization error within the paradigm of empirical processes, e.g., based on Vapnik–Chervonenkis dimension [12] and Rademacher complexity [13], [14], are loose on real-world data and are not intended to directly serve as criteria for optimization [15, Sec. 1]. However, a less known paradigm for characterizing generalization error—specifically for linear discriminant analysis [16], [17] and closely-related plug-in supervised classification methods—described in [18] yields expressions that are not loose and can potentially be used as optimization criteria. In this paper, the generalization error approximations of [18] are expanded to the sensor network context and shown to be useful in optimizing sensor network attributes such as the number of sensors, number of training samples, and sensor spatial distribution. Linear discriminant analysis is limited as a classification algorithm because it relies on strong data distribution assumptions, but has been generalized into a powerful, competitive approach using nonlinear kernels [19], [20]. Although the analysis pursued in this correspondence does not take nonlinear kernels into account, such analysis may be undertaken using results on kernel Mahalanobis distances [21].

Like the model in [22], the spatially-correlated sensor system is modeled as a Gauss–Markov random field with nearest neighbor dependency among randomly placed sensors. Correlation between sensor nodes decays as a function of distance in a manner prescribed by spatial and geostatistics, e.g., according to the Matérn correlation function [1]–[3]. Gaussian random fields have historically been and currently are the predominant probability models for spatial phenomena [1]; Gauss–Markov random fields are becoming more and more popular within spatial and geostatistics [6], and are being used in sensor network analysis as well [22]–[24]. The asymptotic analysis of randomly placed points in growing bounded regions of the plane developed in [25] is employed for simplification purposes.

Common wisdom dictates that when monitoring a field, detection performance improves with more deployed sensors. Furthermore, it is only because of material costs that one limits the number of sensors [26]. However in this work, it is shown that increasing the number of sensors beyond a certain amount results in the degradation of detection performance when the detection rule is learned from a finite training set within the supervised classification context, independent of any power, communication, or network considerations.[1] This degradation behavior is a manifestation of the fundamental phenomenon of overfitting [28].

The analysis in [26] of detection using spatially-correlated sensor measurements corrupted by noise during communication to a fusion center shows that for stochastic signals subject to transmission power constraints, a finite rather than infinite sensor density in space is optimal. The conclusion herein is similar in the sense that with degraded and constrained information (here due to a finite training set), it is better to use a small-dimensional measurement vector rather than an infinite-dimensional one.

While the majority of prior work on detection with multiple sensors deals with known likelihood functions, there has also been work on the supervised learning of detection rules, such as [29]–[32]. These works address questions of what can be done when communication-limited sensors have a labeled training set. Here an even more fundamental question is addressed: even without limitation on communication, what is the generalization behavior of detection rules learned from spatially-correlated sensor measurements. A method for supervised classification with dimensionality reduction and information fusion in tree-structured sensor networks is developed in [33]. In that work, it is also apparent that for a fixed number of training samples, adding more sensors or dimensions of measurements beyond a certain point results in a degradation of detection performance.

Manuscript received March 09, 2011; revised August 25, 2011 and February 06, 2012; accepted February 15, 2012. Date of publication March 06, 2012; date of current version May 11, 2012. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Raviv Raich. Part of the material in this paper was presented at the IEEE International Conference on Acoustics, Speech, and Signal Processing, Prague, Czech Republic, May 2011.

The author is with the Business Analytics and Mathematical Sciences Department, IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598 USA (e-mail: krvarshn@us.ibm.com).

Digital Object Identifier 10.1109/TSP.2012.2190063

[1]In the same vein as [27], the term *sensor network* is used here because it is common, but the term *sensor ensemble* is more appropriate.

Works such as [27] model collections of sensor measurements as deterministic compressible signals rather than stochastically through Gauss–Markov random fields. Although such models do include spatially-independent stochastic measurement noise akin to the nugget effect discussed in Section II-A and signal compressibility does introduce a form of correlation between sensors, they are not the same as the model considered herein. Also, the task considered in such work is estimation rather than supervised classification. Despite these and other differences making conclusions difficult to compare, it can be noted that the optimal (misclassification) error scaling behavior we find is exponential in the number of sensors (cf. Section IV-D), whereas optimal (mean squared error) error scaling is only polynomial in that work [27, eq. 20]. Note however, that in that error scaling, a sublinear number of measurements is needed per sensor [27, eq. 19], whereas the relationship is linear in what is found herein.

The correspondence is organized as follows. Section II presents the Gauss–Markov random field sensor system model and the linear discriminant analysis detection rule. Section III finds approximate expressions for the detection generalization error. Section IV presents several sensor network optimization scenarios in which the derived generalization error expressions may be applied. Section V concludes.

II. Sensor and Detection Rule Models

This section describes the specific setup for which generalization error is characterized. The Gauss–Markov random field model of measurements from randomly placed sensors that is used in the remainder of the paper is given. The linear discriminant analysis detection rule that is learned from training samples is also given.

A. Sensors

Consider the system with p sensor nodes randomly deployed on the plane. Specifically, they are deployed within a square of area p. Thus as the number of sensors grows, the deployment area also grows but the spatial density of sensors remains the same. The location of sensor i, denoted $\mathbf{v}_i \in \mathbb{R}^2$, is drawn according to the distribution $f_{\mathbf{V}}(\mathbf{v})$. This distribution is supported on the same square with area p.

Each sensor measures a scalar random variable $X_i,\ i = 1, \ldots, p$. The overall measurement $\mathbf{X} \in \mathbb{R}^p$ is governed by the two hypotheses $Y \in \{0,1\}$ through the Gaussian likelihood functions $f_{\mathbf{X}|Y}(\mathbf{x}|Y = 0) \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ and $f_{\mathbf{X}|Y}(\mathbf{x}|Y = 1) \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$. The variables $X_i,\ i = 1, \ldots, p$ conditioned on the hypothesis clearly form a Gaussian random field, but moreover form a Gauss–Markov random field through the structure of the covariance matrices.

Modeling similar physical processes under the two hypotheses, we take $\boldsymbol{\Sigma}_0$ to equal $\boldsymbol{\Sigma}_1$ with the particulars of the covariance matrix coming from the spatial locations of the sensors.[2] We assume that correlation decays as a function of distance between sensors. The covariance structure contains Markov relationships according to the Euclidean nearest neighbor graph of the sensors. The nearest neighbor of sensor i is the sensor $j \neq i$ for which the Euclidean distance $d(\mathbf{v}_i, \mathbf{v}_j)$ is minimized. The (undirected) nearest neighbor graph contains an edge between sensor i and sensor j:

- if sensor i is the nearest neighbor of sensor j *or*
- if sensor j is the nearest neighbor of sensor i.

The set of edges in the nearest neighbor graph is denoted $\mathcal{E}$.

It is most convenient to specify the p^2 entries of the common covariance matrix, denoted $\boldsymbol{\Sigma}$, in three parts. First, the diagonal elements of $\boldsymbol{\Sigma}$ are all equal to the constant σ^2. Second, the elements of $\boldsymbol{\Sigma}$ corresponding to edges in the nearest neighbor graph are as follows:

$$\{\boldsymbol{\Sigma}\}_{ij} = \sigma^2 g(d(\mathbf{v}_i, \mathbf{v}_j)), \quad (i,j) \in \mathcal{E} \tag{1}$$

where $g(\cdot) : \mathbb{R}^+ \rightarrow (0,1)$ is a monotonically decreasing function that encodes correlation decay with distance. This decay function is known as the semivariogram [1]–[3]. Often in geostatistics and elsewhere, $g(0^+) < c < 1$ and $g(\cdot) : \mathbb{R}^+ \rightarrow (0, c)$ due to the nugget effect [1]–[3].

Third, the inverse covariance matrix or information matrix, denoted $\mathbf{J} = \boldsymbol{\Sigma}^{-1}$, is used to specify the remaining elements. The off-diagonal elements of $\mathbf{J}$ corresponding to sensor pairs (i,j) that do not have an edge in the nearest neighbor graph are zero, i.e.,

$$\{\mathbf{J}\}_{ij} = 0, \quad i \neq j,\ (i,j) \notin \mathcal{E}. \tag{2}$$

These conditions fully specify all p^2 elements of $\boldsymbol{\Sigma}$ [22].

B. Detection Rule

The detection or binary classification task is to determine the hypothesis y based on the measurement vector $\mathbf{x}$ using a detection rule or classifier function $\hat{y}(\cdot) : \mathbb{R}^p \rightarrow \{0,1\}$. If the distribution $f_{\mathbf{X},Y}(\mathbf{x}, y)$ were given, then the likelihood ratio test detection rule based on it would minimize error [34]. However, this density is not given *a priori*. Only a set of n i.i.d. training samples $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ drawn from the distribution are given. A detection rule must be learned from the training set.

The detection rule studied here is the linear discriminant analysis rule—a simple, classical, often-used technique—given by [17]

$$\hat{y}(\mathbf{x}) = \mathrm{step}(\mathbf{w}^T\mathbf{x} + \theta) \tag{3}$$

where

$$\mathbf{w} = \left(\hat{\boldsymbol{\Sigma}}_0 + \hat{\boldsymbol{\Sigma}}_1\right)^{-1}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)$$
$$\theta = -\frac{1}{2}\mathbf{w}^T(\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1)$$

and $\hat{\boldsymbol{\mu}}_0$, $\hat{\boldsymbol{\mu}}_1$, $\hat{\boldsymbol{\Sigma}}_0$, and $\hat{\boldsymbol{\Sigma}}_1$ are the conditional sample means and covariances of the n training samples. The linear discriminant analysis rule follows from the likelihood ratio test for optimal signal detection between Gaussian signals with the same covariance and different means. Once the detection rule $\hat{y}(\cdot)$ is learned, it is applied to new unseen and unlabeled measurements $\mathbf{x}$.

III. Generalization Error Approximation

As the linear discriminant analysis detection rule is learned from training samples but applied to new test samples, the performance metric of interest is the generalization error $\Pr[\hat{y}(\mathbf{X}) \neq Y]$, which is always greater than or equal to the Bayes optimal detection error achieved by the optimal likelihood ratio test detection rule [28]. Generalization error for the sensor system model of Section II is first studied for a given realization of sensor locations and then as an average across realizations.

A. Linear Discriminant Analysis Generalization Error

Despite extensive study by many researchers, an exact closed form expression for the generalization error of linear discriminant analysis has not yet been found, but several highly accurate approximations exist [18]. One of the best ones when the true likelihood functions are

[2]As another case, we could consider $\boldsymbol{\Sigma}_0 = \sigma^2\mathbf{I}$, where σ^2 is a positive constant, to model independent noise in the absence of a target.

Gaussian and the true prior probabilities are equal, applicable for a wide range of p and n values, is the following [18]:

$$\Pr[\hat{y}(\mathbf{X}) \neq Y] \approx \Phi\left(-\frac{\delta}{2}\left[\left(1+\frac{4p}{n\delta^2}\right)\frac{n}{n-p}\right]^{-\frac{1}{2}}\right) \quad (4)$$

where $\Phi(\cdot)$ is the standard Gaussian cumulative distribution function and

$$\delta^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T \mathbf{J} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \quad (5)$$

is a squared Mahalanobis distance [35]. The $1 + \frac{4p}{n\delta^2}$ term is due to mean vector estimation and the $\frac{n}{n-p}$ term is due to covariance matrix estimation. The difference between the left and right sides of (4) converges to zero as p and n simultaneously go to infinity at a rate such that their ratio converges to a positive constant [18].

The expression when the true prior probabilities are unequal,

$$\begin{aligned}\Pr[\hat{y}(\mathbf{X}) \neq Y] &\approx \pi_0 \Phi\left(-\frac{\delta^2 - \frac{p}{n_0} + \frac{p}{n_1}}{2\sqrt{\left(\delta^2 + \frac{p}{n_0} + \frac{p}{n_1}\right)\frac{n_0+n_1}{n_0+n_1-p}}}\right) \\ &\quad + \pi_1 \Phi\left(-\frac{\delta^2 + \frac{p}{n_0} - \frac{p}{n_1}}{2\sqrt{\left(\delta^2 + \frac{p}{n_0} + \frac{p}{n_1}\right)\frac{n_0+n_1}{n_0+n_1-p}}}\right) \end{aligned} \quad (6)$$

where $\pi_0 = \Pr[Y=0]$, $\pi_1 = \Pr[Y=1]$, n_0 is the number of class 0 training samples and n_1 is the number of class 1 training samples, is similar but ungainly and can be used in a similar way to (4) for detection scenarios that call for $\pi_0 \neq \pi_1$. For simplicity in the remainder of the correspondence, let $\pi_0 = \pi_1 = \frac{1}{2}$. Also for simplicity, let $\boldsymbol{\mu}_0 = \mathbf{0}$ (the length p vector of all zeroes) and let $\boldsymbol{\mu}_1 = \mathbf{1}$ (the length p vector of all ones).

The high quality of (4) is empirically verified in [36]. The analytical mean squared error between the approximation and the true generalization error has not yet been found. In fact, an analytical expression for the case of known covariance matrix has only very recently been found [37], with the authors writing, "As has generally been historically the case, the results for known covariance matrix have been obtained prior to those for unknown covariance matrix, the latter typically being significantly more difficult."

Theorem 1: The squared Mahalanobis distance for the Gauss–Markov random field described in Section II-A is

$$\delta^2 = \frac{p}{\sigma^2} - \frac{2}{\sigma^2}\sum_{(i,j)\in\mathcal{E}} \frac{g(d(\mathbf{v}_i,\mathbf{v}_j))}{1+g(d(\mathbf{v}_i,\mathbf{v}_j))}. \quad (7)$$

Proof: Considering $\boldsymbol{\mu}_0 = \mathbf{0}$, $\boldsymbol{\mu}_1 = \mathbf{1}$, the squared Mahalanobis distance simplifies to

$$\delta^2 = \sum_{i=1}^{p}\sum_{j=1}^{p} \{\mathbf{J}\}_{ij}. \quad (8)$$

Off-diagonal entries of the information matrix with $(i,j) \in \mathcal{E}$ take the value [22]

$$\{\mathbf{J}\}_{ij} = \frac{1}{\sigma^2} \cdot \frac{-g(d(\mathbf{v}_i,\mathbf{v}_j))}{1-g(d(\mathbf{v}_i,\mathbf{v}_j))^2}. \quad (9)$$

Therefore the sum of the off-diagonal elements is

$$\sum_{i=1}^{p}\sum_{j\neq i} \{\mathbf{J}\}_{ij} = 2\sum_{(i,j)\in\mathcal{E}} \{\mathbf{J}\}_{ij} = \frac{2}{\sigma^2}\sum_{(i,j)\in\mathcal{E}} \frac{-g(d(\mathbf{v}_i,\mathbf{v}_j))}{1-g(d(\mathbf{v}_i,\mathbf{v}_j))^2}. \quad (10)$$

Diagonal entries of the information matrix take the value [22]

$$\{\mathbf{J}\}_{ii} = \frac{1}{\sigma^2}\left(1 + \sum_{\{j|(i,j)\in\mathcal{E}\}} \frac{g(d(\mathbf{v}_i,\mathbf{v}_j))^2}{1-g(d(\mathbf{v}_i,\mathbf{v}_j))^2}\right). \quad (11)$$

Consequently, the sum of the diagonal elements is

$$\begin{aligned}\sum_{i=1}^{p}\{\mathbf{J}\}_{ii} &= \frac{1}{\sigma^2}\sum_{i=1}^{p}\left(1 + \sum_{\{j|(i,j)\in\mathcal{E}\}} \frac{g(d(\mathbf{v}_i,\mathbf{v}_j))^2}{1-g(d(\mathbf{v}_i,\mathbf{v}_j))^2}\right) \\ &= \frac{p}{\sigma^2} + \frac{2}{\sigma^2}\sum_{(i,j)\in\mathcal{E}} \frac{g(d(\mathbf{v}_i,\mathbf{v}_j))^2}{1-g(d(\mathbf{v}_i,\mathbf{v}_j))^2}. \end{aligned} \quad (12)$$

The expression (7) is obtained by combining (10) and (12). ∎

This squared Mahalanobis distance is an exact expression. It is only when it is substituted into (4) that there is approximation.

B. Mahalanobis Distance

The Mahalanobis distances depend on the Euclidean distances $d(\mathbf{v}_i,\mathbf{v}_j)$ which in turn depend on the particular realization of the random deployment of sensor locations. For analysis purposes, it is useful to characterize the average behavior of δ across realizations of $\{\mathbf{v}_1,\ldots,\mathbf{v}_p\}$. Such characterization falls within the scope of stochastic geometry and random geometric graphs; space limitations prevent us from providing more background details on these topics, but the survey paper [38] may be consulted for a nice introduction. For the purposes of this correspondence, we will describe the average behavior of functionals of the nearest neighbor graph using average behavior of homogenous Poisson point processes [25].

Theorem 2: The squared Mahalanobis distance (7) can be approximated as

$$\delta^2 \approx \frac{p}{\sigma^2}(1-\zeta) \quad (13)$$

where

$$\zeta = \int \mathrm{E}\left[\sum_{(0,a)\in\mathcal{F}} \phi\left(\frac{d(\mathbf{0},\mathbf{w}_a)}{\sqrt{f_{\mathbf{V}}(\mathbf{v})}}\right)\right] f_{\mathbf{V}}(\mathbf{v})\,d\mathbf{v} \quad (14)$$

and

$$\phi(\cdot) = \frac{g(\cdot)}{1+g(\cdot)}. \quad (15)$$

Proof: For a given function $\phi(\cdot)$, as p grows,

$$\frac{1}{p}\sum_{(i,j)\in\mathcal{E}} \phi(d(\mathbf{v}_i,\mathbf{v}_j)) \to \frac{1}{2}\int \mathrm{E}\left[\sum_{(0,a)\in\mathcal{F}} \phi\left(\frac{d(\mathbf{0},\mathbf{w}_a)}{\sqrt{f_{\mathbf{V}}(\mathbf{v})}}\right)\right] f_{\mathbf{V}}(\mathbf{v})\,d\mathbf{v} \quad (16)$$

where $\mathbf{w}_k$ are spatial locations drawn according to the Poisson point process with unit rate over a unit square centered at the origin, and $\mathcal{F}$ is the set of edges of the nearest neighbor graph constructed from the origin point $\mathbf{0}$ and those points $\mathbf{w}_k$ [25, Theorem 2.2].

Substituting the right side of (16) for the left side of (16) in (7) yields the result. ∎

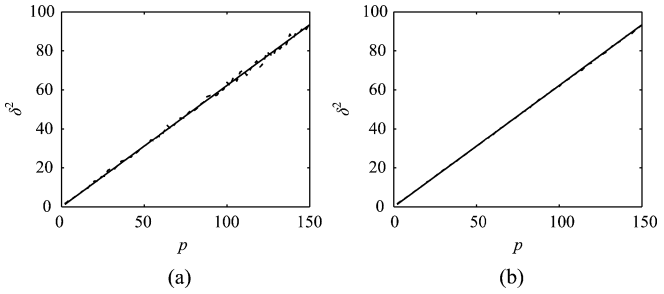

Fig. 1. Squared Mahalanobis distance as a function of the number of sensors with $g(d) = 1/2\exp(-d/2)$ and $\sigma^2 = 1$. The dashed line is the exact value for (a) one uniform sensor location realization and (b) average over 20 realizations. The solid line is the approximation based on a Poisson point process.

It may be noted that the squared Mahalanobis distance is approximately a linear function of the number of sensors p. The approximation is in fact quite tight. As an empirical verification, the squared Mahalanobis distance approximation (13) is compared to the true value (7). The particular semivariogram that is considered is: $g(d) = 1/2\exp(-d/2)$; such exponential models of correlation decay are used to model agricultural, epidemiological, and many other spatial signals, and are a special case of Matérn correlation decay [2], [3]. With this $g(\cdot)$, $\phi(\cdot)$ is a variant of the logistic function.

The exact squared Mahalanobis distance is calculated with the uniform distribution for different numbers of sensors. In Fig. 1(a), the dashed line is the exact value of squared Mahalanobis distance as a function of p for one realization of $\mathbf{V}$. The figure also includes the approximation to squared Mahalanobis distance using Poisson point processes as a solid line. In Fig. 1(b), the dashed line represents the average squared Mahalanobis distance over twenty realizations of the sensor placement distribution. The approximation passes through the single realization values. Averaged over twenty realizations, even for small p, the approximation is so good that the dashed and solid line are nearly indistinguishable.

Whereas finding the exact squared Mahalanobis distance involves constructing the $p \times p$ information matrix for each realization of $\mathbf{V}$ and each p, the approximation only involves finding ζ once,[3] and it is applicable for all p, even as p approaches infinity.

IV. Optimization for Generalization Error

Combining the approximations of Sections III-A and III-B, the overall approximation for the generalization error is

$$\Pr[\hat{y}(\mathbf{X}) \neq Y] \approx \Phi\left(-\sqrt{\frac{p(1-\zeta)}{4\sigma^2}}\left[\left(1+\frac{4\sigma^2}{n(1-\zeta)}\right)\frac{n}{n-p}\right]^{-\frac{1}{2}}\right). \tag{17}$$

This approximation to the generalization error of linear discriminant analysis from spatially-distributed sensors with Gauss–Markov nearest neighbor dependency is used in this section to optimize the parameters and settings of sensor networks.

A. Optimal Number of Sensors for Fixed Number of Training Samples

In certain sensor deployment scenarios, it is known beforehand how much time and resources are available for training. (The assumption of i.i.d. training samples requires some time to elapse between training sample acquisitions to allow temporal correlation that occurs in real-world signals to sufficiently decay.) In these scenarios, the number of training samples n may be viewed as a fixed parameter. The question then is to determine the optimal value of the number of sensors p for fixed n. By differentiating (17) with respect to p, setting equal to zero, and solving for p, we find that the number of sensors that minimizes the generalization error approximation is $p = \frac{n}{2}$, irrespective of σ^2 and ζ. Examining Fig. 2(a), it can be confirmed that test error is minimized when $p = \frac{n}{2}$.

[3]It is not possible to analytically determine ζ; it must be estimated in a Monte Carlo fashion. However, this need only be done once for a distribution $f_{\mathbf{V}}(\mathbf{v})$ and semivariogram $g(\cdot)$. The estimation may be done quite rapidly because it does not depend on p and the number of edges incident on $\mathbf{0}$ in the nearest neighbor graph from the Poisson point process is usually only one or two.

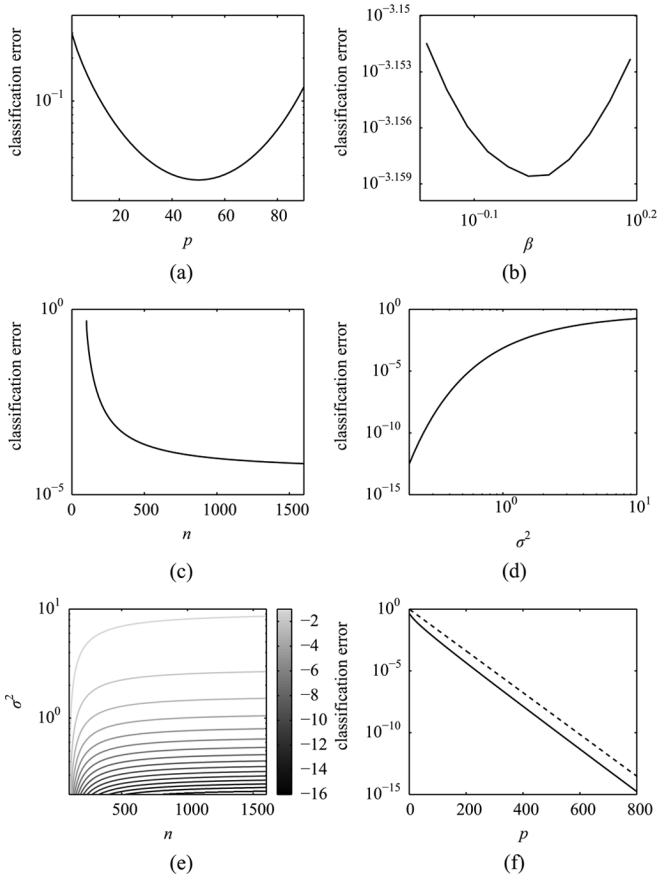

Fig. 2. Generalization error approximation with $g(d) = 1/2\exp(-d/2)$. (a) As a function of the number of sensors for $n = 100$, $\sigma^2 = 1$, and uniform distribution. (b) As a function of the sensor spatial location beta distribution parameter β for $n = 100, p = 50$, and $\sigma^2 = 1$. (c) As a function of the number of samples for $p = 100$, $\sigma^2 = 1$, and uniform distribution. (d) As a function of σ^2 for $n = 300, p = 100$, and uniform distribution. (e) Contours of equal generalization error as a function of n and σ^2 for $p = 100$ and uniform distribution. The shading is labeled by the base ten logarithm of generalization error. (f) As a function of the number of sensors for $p = \frac{n}{2}$, $\sigma^2 = 1$, and uniform distribution. The Chernoff bound is marked as a dashed line.

B. Optimal Sensor Placement Distribution

Another sensor deployment scenario is when the number of training samples is fixed beforehand and the number of sensors is set optimally; the question is how to place the sensors in space. Examining the error expression (17), it can be noted that the approximate generalization error is a monotonically increasing function of the Poisson point process expectation $\zeta \in \left[0, \frac{1}{2}\right]$, implying that the sensor placement distribution $f_{\mathbf{V}}(\mathbf{v})$ should be chosen to minimize ζ in order to achieve the best system performance.

A specific example of such a question is whether the sensors should be clustered in the center of the deployment region, uniformly distributed, or clustered at the edges? Taking the sensor location distribution $f_{\mathbf{V}}(\mathbf{v})$ to be an appropriately scaled and shifted version of the

beta distribution independent and identically distributed in both components of $\mathbf{V}$ with both parameters of the beta distribution equal to β, the sensors are placed uniformly over the square when $\beta = 1$, they are concentrated in the middle of the square when $\beta > 1$, and concentrated at the edges of the square when $\beta < 1$. Thus we can use this distribution to gain insight into the placement distribution question. Based on Monte Carlo estimation, we see that ζ is minimized when $\beta = 1$ and is larger for $\beta > 1$ and for $\beta < 1$, consequently reflected in Fig. 2(b), a plot of the generalization error approximation as a function of β. Therefore, among the family of i.i.d. beta distributions, the uniform distribution optimizes detection performance. This guideline may apply to choices among other similar sensor placement distributions.

C. Constrained Optimization and Tradeoffs for Fixed Number of Sensors

Another scenario involves a fixed number of sensors, but the possibility of choosing the number of training samples. The expression (17) unsurprisingly reveals that the generalization error is a monotonically decreasing function of n, the number of training samples. Thus ideally, without resource constraints, an infinite number of training samples ought to be gathered to approach the Bayes optimal error; however, sensor networks are resource-constrained. Also, slight suboptimality in accuracy is often tolerable.

The question that can be asked in this scenario is: to achieve a particular generalization error probability, how much training is required? Fig. 2(c) plots the generalization error approximation as a function of the number of samples for a fixed number of sensors. Such a plot can be used to determine the number of samples to be acquired in a training phase after deployment of the sensor network.

Similarly, expression (17) shows that generalization error is a monotonically increasing function of σ^2. Better (and costlier) sensor nodes usually induce a smaller magnitude of measurement noise. Thus the sensor network designer may have the ability to set σ^2 to a certain degree. The question in this scenario is similar to that for selecting the number of training samples with a fixed number of sensors. Plots such as that of Fig. 2(d) may be used.

With a fixed number of sensors, it is also possible to examine generalization error on the n–σ^2 plane. A given generalization error requirement defines an isoerror contour in the n–σ^2 plane; the various operating points on this contour represent the tradeoff between the quality of the measurements and the amount of time available for training. Such isoerror contours are shown in Fig. 2(e).

D. Error Exponent for a Fixed Ratio of Sensors and Training Samples

An interesting characterization of the sensor network is to look at generalization error for a fixed ratio of $p = \frac{n}{2}$, as those two variables grow. A plot of this generalization error is shown in Fig. 2(f). With the fixed ratio and large p and n, the generalization error simplifies to

$$\Pr[\hat{y}(\mathbf{X}) \neq Y] \approx \Phi\left(-\sqrt{\frac{1-\zeta}{8\sigma^2}}\sqrt{p}\right). \tag{18}$$

The Chernoff bound/approximation for the Gaussian cumulative distribution function, $\Phi(-\alpha) \lesssim 1/2\exp(-\alpha^2/2)$, may be applied to (18) to obtain

$$\Pr[\hat{y}(\mathbf{X}) \neq Y] \approx \frac{1}{2}\exp\left(-\frac{1-\zeta}{16\sigma^2}p\right). \tag{19}$$

This function is also plotted in Fig. 2(f) and has the same slope as (17). The detection error exponent $\frac{1-\zeta}{16\sigma^2}$ can be compared to error exponents of detection with known likelihood functions (rather than supervised classification from training data) (cf. [39] and [40]). Note that error exponents are often derived when it is not possible to state error expressions for small p, but that is not the case here.

E. Sensor Subset Selection

A scenario that would arise in the operation, rather than the design or deployment of a sensor network, is to choose a subset of k sensors to be active while the remaining $(p-k)$ sensors are inactive. Measurements from active sensors are used as input to the linear discriminant analysis; the active subset should minimize generalization error.

The nearest neighbor graph used to define Markov dependency is acyclic and has a tree or forest structure. A Gauss–Markov random field defined on nodes and edges that are a subforest of a larger forest-structured Gauss–Markov random field maintain the statistical relationships of the larger forest. Selecting a subset of variables (sensor nodes) ought not change covariance relationships in the probabilistic model; therefore for tractability, the specific problem that is considered is to find a cardinality k subtree or subforest (with k predetermined) of the full cardinality p nearest neighbor tree or forest, denoting the edge set of a subtree with k nodes as $\mathcal{E}_k$ and the subtree itself as $\mathcal{T}_k$.

The squared Mahalanobis distance of subtree or subforest $\mathcal{T}_k$ is:

$$\delta^2_{\mathcal{T}_k} = \frac{k}{\sigma^2} - \frac{2}{\sigma^2}\sum_{(i,j)\in\mathcal{E}_k}\phi(d(\mathbf{v}_i,\mathbf{v}_j)). \tag{20}$$

Remembering that all parameters including k are fixed, the optimization problem is

$$\arg\min_{\mathcal{T}_k}\Phi\left(-\frac{\delta_{\mathcal{T}_k}}{2}\left[\left(1+\frac{4k}{n\delta^2_{\mathcal{T}_k}}\right)\frac{n}{n-k}\right]^{-\frac{1}{2}}\right). \tag{21}$$

Since Φ is a monotonically increasing function, its argument is monotonically decreasing in $\delta_{\mathcal{T}_k}$ (for $n \geq k$, which is required for the generalization error approximation to be valid), and Mahalanobis distance is non-negative, the optimization problem (21) is equivalent to

$$\arg\min_{\mathcal{T}_k} -\delta^2_{\mathcal{T}_k} \tag{22}$$

and

$$\arg\min_{\mathcal{T}_k}\sum_{(i,j)\in\mathcal{E}_k}\phi(d(\mathbf{v}_i,\mathbf{v}_j)). \tag{23}$$

The optimization problem (23) is the problem of minimizing the total weight of edges in a graph. A dynamic programming algorithm presented in [41] solves the optimization problem of finding a subtree within a tree to minimize the sum of the edge weights.

As an easy-to-understand illustration of this scenario, consider a uniform deployment of $p = 3$ sensor nodes with a budget for $n = 6$ training samples, with $g(d) = 1/2\exp(-d/2)$ and $\sigma^2 = 1$. The sensors are randomly located at $(-0.6461, 0.6880)$, $(0.8084, -0.2135)$, and $(-0.4149, -0.2837)$ with edges between the first and third sensor and between the second and third sensor. Fixing $k = 3 = p$, there is only one choice, the full tree, with edge weight sum 0.446, squared Mahalanobis distance 2.11, and generalization error 0.357. There are two $k = 2$ subtrees, the one with nodes 1 and 3, and the one with nodes 2 and 3. The edge weight of the first subtree is 0.233, and it has squared Mahalanobis distance 1.53 and generalization error 0.356. The edge weight of the second subtree is 0.213, and it has squared Mahalanobis distance 1.57 and generalization error 0.353. For $k = 2$, the second subtree minimizes the edge weight and consequently also minimizes generalization error.

The optimal subtree is the one that has the sensor nodes farther apart, which is due to more correlation decay between farther nodes. In this example it turns out that even though $p = 3$ sensors are deployed, only $k = 2$ sensors, the sensors 2 and 3, should be activated for the

best classification accuracy. (Using $k = 1$ sensor node would result in 0.362 generalization error.) In deployments of power-limited sensor nodes, taking advantage of this effect may yield significant extensions of monitoring lifetime.

V. Conclusion

It is a fundamental truth that if nothing else, time is a limited resource, and limits a system to finite sets of training samples. It is shown in this work that when learning a linear discriminant analysis detection rule for spatially-correlated sensor measurements with local Gauss–Markov dependency and constant-density random sensor placement, it is optimal to use precisely half the number of sensors as training sample instances. This result that a finite rather than infinite number of sensors is optimal follows from the phenomenon of overfitting. Less is more.

In developing this result, generalization error has been approximated using an expression by Raudys that involves Mahalanobis distance. Mahalanobis distance has been exactly stated for Gauss–Markov sensor measurements, and has also been approximated using weak laws of large numbers. It should be noted that much of the analysis goes through unchanged for other types of Gauss–Markov dependency besides nearest neighbor dependency.

Besides optimizing the number of sensors for a fixed number of training samples, it has been seen that within a family of sensor placement beta distributions, the uniform distribution minimizes a Poisson point process expectation parameter and thus consequently minimizes generalization error.[4]

The overall guideline is then that when using linear discriminant analysis detection given a budget n on the number of training samples, $\frac{n}{2}$ sensors placed uniformly should be used. On the other hand, if the number of sensors is fixed, it is seen that as many training samples as possible should be used. However, in some sense, power or communication budgets often considered in sensor network studies are replaced by a budget on the cardinality of the labeled training set in the learning setting. A detection error exponent for growing numbers of sensors and training samples in fixed ratio has also been derived. In considering the problem of selecting a subset of sensor nodes to activate, it has also been seen that less could be more.

One may argue that the analysis provided here should be taken *cum grano salis* because linear discriminant analysis may provide poor detection performance compared to other classification methods from the statistical learning literature. Linear discriminant analysis has been chosen because it leads to simple, analytic characterization of generalization error with small approximation error, which is not always the case with existing generalization error characterizations of other classification methods [15]. The main thing to take away from the analysis is not only the specifics, but the general theme that irrespective of communication or power constraints, supervised classification by multiple sensors is fundamentally affected by a finite training set: too many sensors degrade performance. This general theme or guiding principle is applicable to all detection rules learned from finite training data.

The performance of linear discriminant analysis may be improved through regularization or through further constraints. Generalization error approximations to many such extensions of the basic linear discriminant analysis are provided in [18], and are also based on Mahalanobis distance. Therefore, the Mahalanobis distance development in this paper applies, and the generalization error may be similarly analyzed for the extensions. An extension of this work to generalized kernel discriminant analysis [19]–[21], which does have competitive classification performance, may also lead to simple, analytic characterization.

In future work, it would be interesting to combine the analysis of this paper with analysis that does consider sensor communication and power constraints. Such combined analysis may reveal several interesting design principles for sensor networks in which the likelihood functions are not known *a priori* but labeled training samples may be collected.

[4]Analytically determining whether the uniform distribution is optimal among all possible distributions with support on a square would be interesting future work.

Acknowledgment

The author would like to thank V. Y. F. Tan for valuable discussions that led to Section IV-E.

References

[1] G. Matheron, *The Theory of Regionalized Variables and Its Applications*. Paris, France: École Nationale Supérieure des Mines de Paris, 1971.
[2] O. Schabenberger and F. J. Pierce, *Contemporary Statistical Models for the Plant and Soil Sciences*. Boca Raton, FL: CRC Press, 2002.
[3] L. A. Waller and C. A. Gotway, *Applied Spatial Statistics for Public Health Data*. Hoboken, NJ: Wiley, 2004.
[4] M. C. Vuran, Ö. B. Akan, and I. F. Akyildiz, "Spatio-temporal correlation: Theory and applications for wireless sensor networks," *Comput. Netw.*, vol. 45, no. 3, pp. 245–259, Jun. 2004.
[5] A. Jindal and K. Psounis, "Modeling spatially correlated data in sensor networks," *ACM Trans. Sensor Netw.*, vol. 2, no. 4, pp. 466–499, Nov. 2006.
[6] F. Lindgren, H. Rue, and J. Lindström, "An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach," *J. Roy. Stat. Soc. B*, vol. 73, no. 4, pp. 423–498, Sep. 2011.
[7] R. R. Tenney and N. R. Sandell, Jr., "Detection with distributed sensors," *IEEE Trans. Aerosp. Electron. Syst.*, vol. AES-17, no. 4, pp. 501–510, Jul. 1981.
[8] J. N. Tsitsiklis, "Decentralized detection," in *Advances in Statistical Signal Processing*, H. V. Poor and J. B. Thomas, Eds. Greenwich, CT: JAI Press, 1993, vol. 2, pp. 297–344.
[9] R. Viswanathan and P. K. Varshney, "Distributed detection with multiple sensors: Part I—Fundamentals," *Proc. IEEE*, vol. 85, no. 1, pp. 54–63, Jan. 1997.
[10] J.-F. Chamberland and V. V. Veeravalli, "Decentralized detection in sensor networks," *IEEE Trans. Signal Process.*, vol. 51, no. 2, pp. 407–416, Feb. 2003.
[11] K. R. Varshney, "Spatially-correlated sensor discriminant analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Prague, Czech Republic, May 2011, pp. 3680–3683.
[12] V. N. Vapnik and A. Y. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory Probab. Appl.*, vol. 16, no. 2, pp. 264–280, 1971.
[13] V. Koltchinskii, "Rademacher penalties and structural risk minimization," *IEEE Trans. Inf. Theory*, vol. 47, no. 5, pp. 1902–1914, Jul. 2001.
[14] P. L. Bartlett and S. Mendelson, "Rademacher and Gaussian complexities: Risk bounds and structural results," *J. Mach. Learn. Res.*, vol. 3, pp. 463–482, Nov. 2002.
[15] O. Bousquet, "New approaches to statistical learning theory," *Ann. Inst. Statist. Math.*, vol. 55, no. 2, pp. 371–389, Jun. 2003.
[16] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, pp. 179–188, 1936.
[17] T. W. Anderson, *Introduction to Multivariate Statistical Analysis*. New York: Wiley, 1958.
[18] Š. Raudys and D. M. Young, "Results in statistical discriminant analysis: A review of the former Soviet Union literature," *J. Multivar. Anal.*, vol. 89, no. 1, pp. 1–35, Apr. 2004.
[19] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K.-R. Müller, "Fisher discriminant analysis with kernels," in *Proc. IEEE Workshop Neural Netw. Signal Process.*, Madison, WI, Aug. 1999, pp. 41–48.
[20] G. Baudat and F. Anouar, "Generalized discriminant analysis using a kernel approach," *Neural Comput.*, vol. 12, no. 10, pp. 2385–2404, Oct. 2000.
[21] B. Haasdonk and E. Pękalska, "Classification with kernel Mahalanobis distance classifiers," in *Advances in Data Analysis, Data Handling and Business Intelligence*, A. Fink, B. Lausen, W. Seidel, and A. Ultsch, Eds. Heidelberg, Germany: Springer, 2010, pp. 351–361.

[22] A. Anandkumar, L. Tong, and A. Swami, "Detection of Gauss-Markov random fields with nearest-neighbor dependency," *IEEE Trans. Inf. Theory*, vol. 55, no. 2, pp. 816–827, Feb. 2009.
[23] Y. Sung, H. V. Poor, and H. Yu, "How much information can one get from a wireless ad hoc sensor network over a correlated random field?," *IEEE Trans. Inf. Theory*, vol. 55, no. 6, pp. 2827–2847, Jun. 2009.
[24] J. Fang and H. Li, "Distributed estimation of Gauss-Markov random fields with one-bit quantized data," *IEEE Signal Process. Lett.*, vol. 17, no. 5, pp. 449–452, May 2010.
[25] M. D. Penrose and J. E. Yukich, "Weak laws of large numbers in geometric probability," *Ann. Appl. Prob.*, vol. 13, no. 1, pp. 277–303, Jan. 2003.
[26] J.-F. Chamberland and V. V. Veeravalli, "How dense should a sensor network be for detection with correlated observations?," *Ann. Appl. Prob.*, vol. 52, no. 11, pp. 5099–5106, Nov. 2006.
[27] W. U. Bajwa, J. D. Haupt, A. M. Sayeed, and R. D. Nowak, "Joint source-channel communication for distributed estimation in sensor networks," *IEEE Trans. Inf. Theory*, vol. 53, no. 10, pp. 3629–3653, Oct. 2007.
[28] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer, 1995.
[29] X. Nguyen, M. J. Wainwright, and M. I. Jordan, "Nonparametric decentralized detection using kernel methods," *IEEE Trans. Inf. Theory*, vol. 53, no. 11, pp. 4053–4066, Nov. 2005.
[30] J. B. Predd, S. R. Kulkarni, and H. V. Poor, "Consistency in models for distributed learning under communication constraints," *IEEE Trans. Signal Process.*, vol. 52, no. 1, pp. 52–63, Jan. 2006.
[31] J. B. Predd, S. R. Kulkarni, and H. V. Poor, "Distributed learning in wireless sensor networks," *IEEE Trans. Inf. Theory*, vol. 23, no. 4, pp. 56–69, Jul. 2006.
[32] H. Zheng, S. R. Kulkarni, and H. V. Poor, "Attribute-distributed learning: Models, limits, and algorithms," *IEEE Signal Process. Mag.*, vol. 59, no. 1, pp. 386–398, Jan. 2011.
[33] K. R. Varshney and A. S. Willsky, "Linear dimensionality reduction for margin-based classification: High-dimensional data and sensor networks," *IEEE Trans. Signal Process.*, vol. 59, no. 6, pp. 2496–2512, Jun. 2011.
[34] H. L. Van Trees, *Detection, Estimation, and Modulation Theory*. New York: Wiley, 1968.
[35] P. C. Mahalanobis, "On the generalized distance in statistics," *P. Nat. Inst. Sci. India*, vol. 2, no. 1, pp. 49–55, Apr. 1936.
[36] F. J. Wyman, D. M. Young, and D. W. Turner, "A comparison of asymptotic error rate expansions for the sample linear discriminant function," *Pattern Recogn.*, vol. 23, no. 7, pp. 775–783, Jul. 1990.
[37] A. Zollanvari, U. M. Braga-Neto, and E. R. Dougherty, "Analytic study of performance of error estimators for linear discriminant analysis," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4238–4255, Sep. 2011.
[38] M. Haenggi, J. G. Andrews, F. Baccelli, O. Dousse, and M. Franceschetti, "Stochastic geometry and random graphs for the analysis and design of wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 27, no. 7, pp. 1029–1046, Sep. 2009.
[39] S. Misra and L. Tong, "Error exponents for Bayesian detection with randomly spaced sensors," in *Proc. IEEE Workshop Signal Process. Adv. Wireless Commun.*, Helsinki, Finland, Jun. 2007.
[40] A. Anandkumar, A. S. Willsky, and L. Tong, "Detection error exponent for spatially dependent samples in random networks," in *Proc. IEEE Int. Symp. Inf. Theory*, Seoul, Korea, Jun./Jul. 2009, pp. 2882–2886.
[41] C. Blum, "Revisiting dynamic programming for finding optimal subtrees in trees," *Eur. J. Oper. Res.*, vol. 177, no. 1, pp. 102–115, Feb. 2007.

An Enhanced IAF-PNLMS Adaptive Algorithm for Sparse Impulse Response Identification

Francisco das Chagas de Souza, Rui Seara, and Dennis R. Morgan

***Abstract*—This correspondence presents an individual-activation-factor proportionate normalized least-mean-square (IAF-PNLMS) algorithm that (during the adaptive process) uses a new gain distribution strategy for updating the filter coefficients. This strategy consists of increasing the gain assigned to the inactive coefficients as the active ones approach convergence. For such, whenever a predefined threshold is crossed during the learning process, a new gain distribution is carried out, rather than to assign gains proportional to coefficient magnitudes as the IAF-PNLMS algorithm does. This new version of the IAF-PNLMS algorithm leads to a better distribution of the adaptation energy over the whole learning process. As a consequence, for impulse responses exhibiting high sparseness, the proposed algorithm achieves faster convergence, outperforming the IAF-PNLMS and other well-known PNLMS-type algorithms.**

***Index Terms*—Adaptive filtering, gain redistribution, proportionate normalized least-mean-square (PNLMS) algorithm, sparse impulse response, system identification, thresholding technique.**

I. Introduction

Sparse impulse responses are encountered in many real-world applications, such as communications, acoustics, and seismic processes [1]–[5]. Such responses are qualitatively classified as sparse if most of the coefficients take values near zero and only a few have significant values [1], [6]. For this class of plant impulse responses, classical adaptive algorithms using the same step-size value for all filter coefficients, such as the normalized least-mean-square (NLMS) algorithm, are outperformed by algorithms that exploit the sparse nature of the impulse response [2], [5], such as proportionate NLMS (PNLMS) [7]–[9], in which each filter coefficient is updated proportionally to its magnitude, resulting in higher convergence speed. However, the standard PNLMS algorithm suffers some performance degradation as the sparseness decreases [10]; furthermore, its fast initial convergence is not maintained over the whole adaptation process [11]–[13]. Improved versions of the PNLMS algorithm, aiming to deal with impulse responses exhibiting medium sparseness, are the PNLMS++ [7], [9] and improved PNLMS (IPNLMS) [10]. Nevertheless, these algorithms do not provide the same fast initial convergence obtained with the PNLMS for impulse responses having high sparseness [14], [15]. A version of the PNLMS algorithm that takes into account the sparseness variation of the plant is the sparseness-controlled PNLMS (SC-PNLMS) [16]. This algorithm performs well for both very high sparseness and medium dispersions; however, such performance is obtained at the expense of higher computational complexity, as compared with the standard PNLMS algorithm. Aiming to preserve the fast initial convergence over the whole adaptation process, the μ-law PNLMS (MPNLMS) and adaptive MPNLMS (AMPNLMS) algorithms are, respectively, proposed in [11] and [12] at the expense

Manuscript received June 22, 2011; revised October 14, 2011 and January 11, 2012; accepted February 22, 2012. Date of publication March 08, 2012; date of current version May 11, 2012. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Konstantinos Slavakis.

F. das Chagas de Souza and R. Seara are with the LINSE—Circuits and Signal Processing Laboratory, Department of Electrical Engineering, Federal University of Santa Catarina, Trindade, 88040-900, Florianópolis, SC, Brazil (e-mail: fsouza@linse.ufsc.br; seara@linse.ufsc.br).

D. R. Morgan is with the Bell Laboratories, Alcatel-Lucent, Murray Hill, NJ 07974-0636 USA (e-mail: drrm@bell-labs.com).

Digital Object Identifier 10.1109/TSP.2012.2190407