

Classification of Gaussian Data with Sieve-Regularized Estimates

Natalia A. Schmid
Coordinated Science Laboratory
University of Illinois at Urbana-Champaign
nschmid@ifp.uiuc.edu

Joseph A. O'Sullivan
Electronic Systems and Signals Research Laboratory
Department of Electrical Engineering
Washington University in St. Louis
jao@ee.wustl.edu

ABSTRACT

Many classification problems use image or other high-dimensional data, and must be designed from training data. The design and analysis of such systems parameterized by unknown functions, based on a method of sieves to regularize the function estimates, is described. The test statistic is assumed to be the ideal test statistic with estimated functions substituted for the truth. The test statistic is decomposed into approximation error and estimation error components, providing analytical tools for determining the optimal sieve size.

Keywords: Recognition, function estimation, maximum likelihood, sieves, Kullback-Leibler information, splines.

1. INTRODUCTION

Practical classification systems are often designed using a limited amount of stochastic data, where the underlying stochastic processes are not completely known. These data are typically used to train or estimate finite or infinite dimensional parameters in a stochastic model (see Figure 1). This paper considers the case of infinite dimensional parameters (continuous functions) where estimates for the parameters are substituted into ideal test statistics in place of the true unknown parameters.

When continuous functions are estimated using a finite amount of data, some regularization method must be applied. In function estimation problems, the quality of estimates is usually measured with respect to a specified loss function (the average quadratic loss function or the relative entropy, for example) that admits a natural decomposition into two parts: the error of approximation (bias) and the error of estimation (variance). These parts exhibit different behavior with decreasing regularization parameter, resulting in a trade-off between them.

Intuitively, the best estimates of functions applied to a classification problem are expected to perform well. However, examples presented in [5] and in a sequence of works following it [4,9,18], have shown that the best function estimates are not always guaranteed to perform well in classification. For classification problems, a series of bias-variance decompositions for the 0/1-loss function was proposed. The components of these decompositions have a very different behavior from the behavior of bias and variance in estimation problem.

The existence of bias-variance decomposition and bias-variance trade-off for classification problems motivates our work. The model for the stochastic data in this paper differs from the one used in [5]. We assume that stochastic data are samples of two stationary complex Gaussian processes with zero mean and unknown real positive and distinct power spectral densities. Independent training vectors of a finite length are available to estimate the unknown parameters. The maximum likelihood (ML) estimation method is applied to obtain the estimates. To regularize the estimates of the unknown parameters, we use Grenander's method of sieves [6], following the work of Moulin, et al. [11,12].

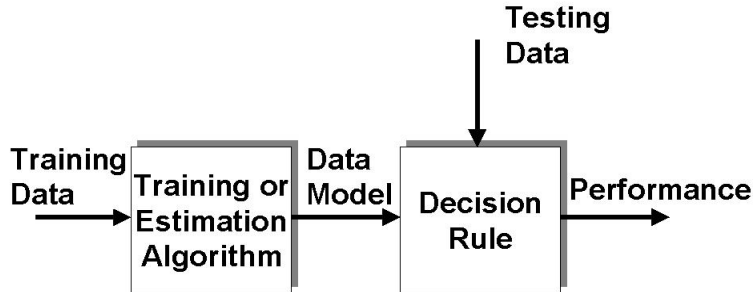


Figure 1. Block-diagram of a typical recognition system.

To analyze classification performance, we propose a bias-variance decomposition of the loglikelihood ratio and then analyze each term in the decomposition separately. The analysis is based upon a first order approximation of this decomposition. We find the conditions for the average probability of classification error to be consistent under the setting above and obtain an expression for the optimal rate of sieve growth.

In Section 2, we state the classification problem and define the test statistic. In Section 3, we present a decomposition for the test statistic and find asymptotic approximations for the components. In Section 4, we derive an expression for the optimal rate of the sieve growth for the classification problem. An iterative solution to evaluating restricted ML estimates is provided in Section 5. A summary is presented in Section 6.

2. PROBLEM STATEMENT

Consider a binary hypothesis testing problem. Assume that observed data are realizations of one of two stationary, zero-mean, discrete-time, Gaussian random processes. The power spectral densities under the two hypotheses are $f_S(\lambda)$ and $f_T(\lambda)$, $\lambda \in [-\frac{1}{2}, \frac{1}{2}]$. $f_S(\lambda)$ and $f_T(\lambda)$ are distinct, unknown, continuous positive valued functions defined on the interval $[-\frac{1}{2}, \frac{1}{2}]$. Suppose that a vector \mathbf{R} , a realization of n consecutive samples from one of the two processes, is observed and must be classified.

Suppose that two independent of \mathbf{R} and also mutually independent n -dimensional vectors \mathbf{S} and \mathbf{T} , samples of the two processes, are available to estimate the unknown functions. The question is how to optimally make use of the observed data to minimize the average classification error? This paper proposes a solution that relies on the following assumptions: (i), the test is designed to implement the loglikelihood ratio with the ML estimates substituted in place of the true parameters; and (ii), the ML estimates of the parameters are obtained using only the vectors \mathbf{S} and \mathbf{T} .

The loglikelihood function for the data \mathbf{S} is

$$l(\mathbf{S}) = -\mathbf{S}^\dagger \mathbf{K}^{-1}(f_S) \mathbf{S} - \log \det \mathbf{K}(f_S),$$

where $\mathbf{K}(f_S)$ is the Toeplitz Hermitian covariance matrix parameterized by the power spectral density $f_S(\lambda)$. A similar expression can be written for the loglikelihood of the data \mathbf{T} . Any parameter that maximizes the likelihood function of the data is a ML solution.

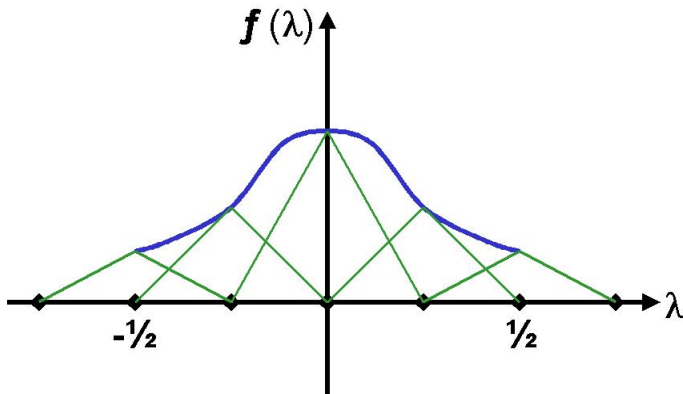


Figure 2. A function constructed from B-splines.

2.1. Spline sieve

The estimation of the continuous functions $f_S(\lambda)$ and $f_T(\lambda)$ from a finite amount of data is an ill-posed problem. Various approaches to constraining the set of solutions are possible (see [13,17] for an overview of regularization methods). In this work, we apply the method of sieves [6] to obtain stable ML estimates of the power spectral densities. A sieve in a parameter space \mathcal{F} is defined as a family of restricted subsets \mathcal{F}_μ , parameterized by a regularization parameter μ , called the mesh size. Each sieve has to satisfy two conditions: (i), a maximizer of the loglikelihood function, called restricted ML estimate, exists and is unique for each subset \mathcal{F}_μ of the sieve; (ii), any element in the parameter space \mathcal{F} can be arbitrary well approximated by an element of the subset \mathcal{F}_μ when $\mu \rightarrow 0$.

The sieve must be designed such that, as $n \rightarrow \infty$, the sequence of the restricted estimates converges to the true parameter in some sense. This can be usually achieved by requiring that the mesh size μ tends to zero at an appropriate rate with growth of the observation vector.

Following Moulin et al. [11,12], we use a spline sieve. The power spectral densities of Gaussian processes belong to a linear space of real valued functions defined on the interval $[-\frac{1}{2}, \frac{1}{2}]$. Thus, the solution can be sought in the space of all positive absolutely integrable functions $\mathbf{L}_+^1([-\frac{1}{2}, \frac{1}{2}])$. The parameter set is defined to be

$$\mathcal{F} = \{f(\lambda) \in \mathbf{L}_+^1([-\frac{1}{2}, \frac{1}{2}]) | f(\lambda) = \sum_{k \in \Lambda} a(k)\psi_k(\lambda), a(k) \geq 0\},$$

where Λ is a countable index set; $\{\psi_k(\lambda), k \in \Lambda\}$ is a set of B-splines spanning the interval $[-\frac{1}{2}, \frac{1}{2}]$, and $\{a(k), k \in \Lambda\}$ are the coefficients of the representation.

The spline sieve in the space \mathcal{F} is a family of subsets \mathcal{F}_μ . Each function in the subset \mathcal{F}_μ admits the following truncated representation

$$f_\mu(\lambda) = \sum_{k \in \Lambda_\mu} a(k)\psi_k(\lambda), \quad \lambda \in [-\frac{1}{2}, \frac{1}{2}],$$

where Λ_μ is a finite index set made to be related to the mesh size as follows. We assume that the interval $[-\frac{1}{2}, \frac{1}{2}]$ is partitioned into Q equal subintervals with the knots $-\frac{1}{2} = \lambda_0, \dots, \lambda_Q = \frac{1}{2}$ ($\lambda_k = -\frac{1}{2} + \frac{k}{Q}$). See Figure 2 for illustration. The indices of the knots form the set Λ_μ , with μ and Q related by $Q = \mu^{-1}$. For each given μ , the restricted ML estimates are sought in the subset \mathcal{F}_μ . The problem of function estimation reduces to the estimation of μ^{-1} unknown nonnegative coefficients $\{a(k)\}$. It was shown by Moulin et al. in [11,12] that the sets \mathcal{F}_μ described

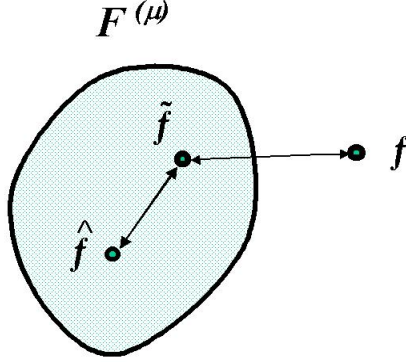


Figure 3. Elements of bias-variance decomposition within a subset of the sieve.

above satisfy the two conditions for the family of subsets to be a sieve. It was also proven that the spline sieve is consistent in the directed divergence sense, to be defined below, under the condition that the rate of the sieve growth is of the order of $o(n^{-1})$.

2.2. Measure of the estimation performance

In estimation problem, for a sieve to be consistent, it has to be designed such that the mesh size decreases at an appropriate rate as the size of the observed vectors increases. The consistency of the estimates in [12] is studied with respect to the directed divergence, a distance measure derived from the relative entropy between two parametric distributions. Directed divergence between the elements of the parameter space \mathcal{F} is defined as follows. Consider two probability density functions $f_{\mathbf{a}}(\mathbf{s})$ and $f_{\mathbf{a}'}(\mathbf{s})$, parameterized by \mathbf{a} and $\mathbf{a}' \in \mathcal{F}$, where $\mathbf{s} \in \mathcal{R}^n$. Then the directed divergence between parameters \mathbf{a} and \mathbf{a}' is given by

$$d^{(n)}(\mathbf{a}, \mathbf{a}') = \frac{1}{n} D(f_{\mathbf{a}} \| f_{\mathbf{a}'}) = \frac{1}{n} \int f_{\mathbf{a}}(\mathbf{s}) \log \frac{f_{\mathbf{a}}(\mathbf{s})}{f_{\mathbf{a}'}(\mathbf{s})} d\mathbf{s}. \quad (1)$$

If the parameter \mathbf{a}' is a restricted ML estimate, it is natural to define the distance between the parameters \mathbf{a} and $\hat{\mathbf{a}}_{\mu}$ such that it is independent of the data

$$\bar{d}^{(n)}(\mathbf{a}, \hat{\mathbf{a}}_{\mu}) \equiv E \left[d^{(n)}(\mathbf{a}, \hat{\mathbf{a}}_{\mu}) \right] = \frac{1}{n} \int f_{\mathbf{a}}(\mathbf{s}) D(f_{\mathbf{a}} \| f_{\hat{\mathbf{a}}_{\mu}(\mathbf{s})}) d\mathbf{s}. \quad (2)$$

It was shown in [12] that the directed divergence between f and \hat{f}_{μ} , when \hat{f}_{μ} is sought within a subset of the sieve \mathcal{F}_{μ} , can be naturally decomposed into two parts

$$d^{(n)}(f, \hat{f}_{\mu}) \sim d^{(n)}(f, \tilde{f}_{\mu}) + d^{(n)}(\tilde{f}_{\mu}, \hat{f}_{\mu}), \quad (3)$$

where \tilde{f}_{μ} is the asymptotic in n restricted ML estimate of the function f obtained within the subset of the sieve \mathcal{F}_{μ} . Elements of this bias-variance decomposition within a subset of the sieve are shown in Figure 3. For i.i.d. samples, the sequence of the ML estimates \hat{f}_{μ} converges to \tilde{f}_{μ} under some regularity conditions provided in [19,7,8]. The parameter \tilde{f}_{μ} is the closest point of the subset \mathcal{F}_{μ} to the true parameter f in the directed divergence sense

$$d^{(n)}(f, \tilde{f}_{\mu}) \leq d^{(n)}(f, \hat{f}_{\mu}),$$

where f_μ is any element of the subset \mathcal{F}_μ . This inequality is true for every finite n and also in the limit (provided it exists).

As $\mu \rightarrow 0$, every element of the parameter set \mathcal{F} can be arbitrary well approximated by some element of the subset \mathcal{F}_μ in the information sense, i.e.

$$\bar{d}(f, \tilde{f}_\mu) \rightarrow 0, \quad \text{as } \mu \rightarrow 0.$$

For more information about sieves, their design, and applications see [6,2,11,12,16].

2.3. Test statistic

Denote by $\{\hat{f}_{\mu,S}(\lambda)\}$ and $\{\hat{f}_{\mu,T}(\lambda)\}$ the sequences of restricted ML estimates for the power spectral densities parameterized by μ . Assume that for each given μ , the system is designed to implement the loglikelihood ratio test with the test statistic given by

$$\begin{aligned} \hat{l}_\mu(\mathbf{R}) &= \hat{l}_{\mu,S}(\mathbf{R}) - \hat{l}_{\mu,T}(\mathbf{R}) \\ &= -\mathbf{R}^\dagger(\mathbf{K}^{-1}(\hat{f}_{\mu,S}) - \mathbf{K}^{-1}(\hat{f}_{\mu,T}))\mathbf{R} - \log \det(\mathbf{K}(\hat{f}_{\mu,S})\mathbf{K}^{-1}(\hat{f}_{\mu,T})), \end{aligned} \quad (4)$$

where $\mathbf{K}(\hat{f}_{\mu,S})$ and $\mathbf{K}(\hat{f}_{\mu,T})$ are the Toeplitz Hermitian matrices parameterized by the functions $\hat{f}_{\mu,S}$ and $\hat{f}_{\mu,T}$.

3. LOGLIKELIHOOD DECOMPOSITION

Let $l_S(\mathbf{R})$ and $l_T(\mathbf{R})$ denote the loglikelihood functions of the data \mathbf{R} parameterized by the functions $f_S(\cdot)$ and $f_T(\cdot)$ (the true likelihood functions). In this and the following sections, for the purpose of analysis, we assume that the true parameters are known. Let $\tilde{l}_{\mu,S}$ and $\tilde{l}_{\mu,T}$ denote the loglikelihood function of the data \mathbf{R} when the asymptotic ML estimates of the power spectral densities $\tilde{f}_{\mu,S}$ and $\tilde{f}_{\mu,T}$, restricted to be in the sieve subset \mathcal{F}_μ , are used in the loglikelihood functions in place of the true parameters. With this notation, the following decomposition of the test statistic holds

$$\begin{aligned} \hat{l}_{\mu,S}(\mathbf{R}) - \hat{l}_{\mu,T}(\mathbf{R}) &= l_S(\mathbf{R}) - l_T(\mathbf{R}) + [\tilde{l}_{\mu,S}(\mathbf{R}) - l_S(\mathbf{R})] \\ &\quad - [\tilde{l}_{\mu,T}(\mathbf{R}) - l_T(\mathbf{R})] + [\hat{l}_{\mu,S}(\mathbf{R}) - \tilde{l}_{\mu,S}(\mathbf{R})] - [\hat{l}_{\mu,T}(\mathbf{R}) - \tilde{l}_{\mu,T}(\mathbf{R})]. \end{aligned} \quad (5)$$

This decomposition has five components: the true loglikelihood ratio, two components of the estimation error, and two components of the approximation error. Asymptotic approximations for the estimation and approximation error components are obtained below.

3.1. Estimation error

Consider the joint probability density function for the random n -dimensional vector \mathbf{R} parameterized by an element of the subset \mathcal{F}_μ . Assume that the function $\log p(\mathbf{r} : \mathbf{a}_\mu)$ is twice continuously differentiable with respect to the components $a(k)$, $k \in \Lambda_\mu$; and that the limit of the expected values $\frac{1}{n}E\{\nabla \log p(\mathbf{r} : \mathbf{a}_\mu)\}$, $\frac{1}{n}E\{\nabla^2 \log p(\mathbf{r} : \mathbf{a}_\mu)\}$, and $\frac{1}{n}E\{\nabla \log p(\mathbf{r} : \mathbf{a}_\mu)\nabla^T \log p(\mathbf{r} : \mathbf{a}_\mu)\}$, exists for every $f_\mu \in \mathcal{F}_\mu$. Here and below, \mathbf{a}_μ is a Q -dimensional vector of coefficients in the representation $f_\mu(\lambda) = \sum_{k=1}^Q a_\mu(k)\psi_k(\lambda)$. Suppose that the law of large numbers holds for the normalized sequences $\frac{1}{n}\{\nabla \log p(\mathbf{r} : \mathbf{a}_\mu)\}$, $\frac{1}{n}\{\nabla \log p(\mathbf{r} : \mathbf{a}_\mu)\nabla^T \log p(\mathbf{r} : \mathbf{a}_\mu)\}$, and $\frac{1}{n}\{\nabla^2 \log p(\mathbf{r} : \mathbf{a}_\mu)\}$. Then the sufficient conditions for consistency and asymptotic normality of the restricted ML estimates stated in [19,7] can be extended to the case of dependent random variables. The conditions stated by Huber in [7], under which ML estimates converge to a well-defined limit, are very general and rely on the works of Wald; however, Huber does not explicitly discuss the information theoretic interpretation of this limit. This interpretation has been emphasized by Akaike in [1], who has observed that when the true distribution is unknown, the ML estimate is a natural estimate for the parameters which minimize the Kullback-Leibler Information Criterion. White's derivations in [19] are motivated by Akaike's observation. The notations and the theory used and developed in the rest of this paper will rely on those of White's [19].

Denote by $L(f_k, \mathbf{a}_{\mu,l})$ the limiting value of the sequence $\frac{1}{n}E\{\nabla \log p(\mathbf{R} : \mathbf{a}_{\mu,l})|H_k\}$, $k, l = \{S, T\}$; by $J(f_k, \mathbf{a}_{\mu,l})$ the limiting value of the sequence $\frac{1}{n}E\{\nabla^2 \log p(\mathbf{R} : \mathbf{a}_{\mu,l})|H_k\}$; and by $I(f_k, \mathbf{a}_{\mu,l})$ the limiting value of the sequence $\frac{1}{n}E\{\nabla \log p(\mathbf{R} : \mathbf{a}_{\mu,l})\nabla^T \log p(\mathbf{R} : \mathbf{a}_{\mu,l})|H_k\}$. Note that all expectations are taken with respect to the true probability distribution.

Theorem 1: Let \mathbf{R} be drawn from the distribution under H_1 . Then as $n \rightarrow \infty$

(i) $2(\hat{l}_{\mu,1} - \tilde{l}_{\mu,1})$ is asymptotically approximated by a linear combination of $Q = 1/\mu$ independent central chi-square distributed random variables with one degree of freedom $\chi_k^2(0, 1)$

$$2(\hat{l}_{\mu,1} - \tilde{l}_{\mu,1}) \sim \sum_{k=1}^Q \zeta_1(k) \chi_k^2(0, 1) \quad (6)$$

(ii) Under H_1 , $E(\hat{l}_{\mu,1} - \tilde{l}_{\mu,1} | H_1) = \frac{1}{2} \sum_{k=1}^Q \zeta_1(k)$;

(iii) $2(\hat{l}_{\mu,2} - \tilde{l}_{\mu,2})$ is asymptotically distributed as a sum of a Gaussian random variable and a linear combination of Q independent central chi-square distributed random variables

$$2(\hat{l}_{\mu,2} - \tilde{l}_{\mu,2}) \sim \sum_{k=1}^Q \rho_1(k) \chi_k^2(0, 1) + 2\sqrt{n} \mathcal{N}(\mathbf{0}, L^T(f_S, \tilde{\mathbf{a}}_{\mu,2}) J^{-1}(f_T, \tilde{\mathbf{a}}_{\mu,2}) I(f_T, \tilde{\mathbf{a}}_{\mu,2}) J^{-1}(f_T, \tilde{\mathbf{a}}_{\mu,2}) L(f_S, \tilde{\mathbf{a}}_{\mu,2})); \quad (7)$$

(iv) Under H_1 , $\frac{1}{n} E(\hat{l}_{\mu,2} - \tilde{l}_{\mu,2} | H_1) \sim \frac{1}{n} \sum_{k=1}^Q \rho_1(k)$.

The proof of (i) and (ii) is similar to the proof of statements in Lemma 1 in [12, p.63] and is based on a generalization of the results from [20, Ch.13.4] and [19]. (iii) and (iv) can be proven using similar techniques. For details see [14]. Here $\{\zeta_1(k)\}$ and $\{\rho_1(k)\}$ are the eigenvalues of the matrix products

$$[J^{-1/2}(f_S, \tilde{\mathbf{a}}_{\mu,1}) I(f_S, \tilde{\mathbf{a}}_{\mu,1}) J^{-1/2}(f_S, \tilde{\mathbf{a}}_{\mu,1})]$$

and

$$[J^{1/2}(f_S, \tilde{\mathbf{a}}_{\mu,2}) J^{-1}(f_T, \tilde{\mathbf{a}}_{\mu,2}) I(f_T, \tilde{\mathbf{a}}_{\mu,2}) J^{-1}(f_T, \tilde{\mathbf{a}}_{\mu,2}) J^{1/2}(f_S, \tilde{\mathbf{a}}_{\mu,2})],$$

respectively.

Note that for every finite μ , if the true parameter is not in \mathcal{F}_μ , the asymptotic covariance matrices of the restricted ML estimates do not equal the inverse of Fisher's information matrix.

Under certain regularity conditions, as Q gets large, the sums of the $\{\rho_1(k)\}$ and $\{\zeta_1(k)\}$ converge to integrals $\frac{Q}{2} \int_{-1/2}^{1/2} \tilde{\rho}_1(\lambda) d\lambda$ and $\frac{Q}{2} \int_{-1/2}^{1/2} \tilde{\zeta}_1(\lambda) d\lambda$, respectively.

3.2. Approximation error

Suppose that the functions f_S and f_T belong to the class \mathbf{L}_q^2 , the class of functions with absolutely continuous derivatives to the order of $(q-1)$ and with the derivative of order q belonging to \mathbf{L}^2 . As proven in [15], the smoothness of functions and the order of splines used for their approximation are related. We assume that the order of the B-splines is L , $L \leq q$.

To analytically treat the approximation error terms in (5) is a difficult task, therefore we apply Toeplitz theory to overcome this problem. This yields

$$\lim_{n \rightarrow \infty} \frac{1}{n} E(\tilde{l}_{\mu,1} - l_1 | H_1) = \int \left[1 - \frac{f_S(\lambda)}{\tilde{f}_S(\lambda)} + \log \frac{f_S(\lambda)}{\tilde{f}_S(\lambda)} \right] d\lambda,$$

$$\lim_{n \rightarrow \infty} \frac{1}{n} E(\tilde{l}_{\mu,1} - l_1 | H_2) = - \int \frac{f_T(\lambda)}{f_S(\lambda)} \left[\frac{f_S(\lambda)}{\tilde{f}_S(\lambda)} - 1 \right] d\lambda - \int \log \left[\frac{\tilde{f}_S(\lambda)}{f_S(\lambda)} \right] d\lambda.$$

Similar expressions can be obtained for $\lim_{n \rightarrow \infty} \frac{1}{n} E(\tilde{l}_{\mu,2} - l_2 | H_i)$, $i = 1, 2$.

Proposition 1: Under H_1 , the two parts of the average approximation error in (5) are

$$\lim_{n \rightarrow \infty} \frac{1}{n} E(\tilde{l}_{\mu,1} - l_1 | H_1) = - \frac{C_L \mu^{2L}}{4} \int \left| \frac{\partial^L f_S(\lambda) / \partial \lambda^L}{f_S(\lambda)} \right|^2 d\lambda,$$

where C_L is a constant that depends only on L ,

$$\lim_{n \rightarrow \infty} \frac{1}{n} E(\tilde{l}_{\mu,2} - l_2 | H_1) \sim \frac{C_L \mu^{2L}}{4} \int \left| \frac{\partial^L f_T(\lambda) / \partial \lambda^L}{f_T(\lambda)} \right|^2 d\lambda - \frac{C_L \mu^{2L}}{2} \int \frac{f_S(\lambda)}{f_T(\lambda)} \left| \frac{\partial^L f_T(\lambda) / \partial \lambda^L}{f_T(\lambda)} \right|^2 d\lambda.$$

The proof is based on the Taylor series expansion and the theory of splines (see [12,14] for details).

A similar expression can be derived for the average approximation error under the hypothesis H_2 .

Substituting these results in the decomposition (5) results in first order asymptotic approximations under each hypothesis. Under H_1 ,

$$\begin{aligned} \frac{1}{n} [\hat{l}_{\mu,1} - \hat{l}_{\mu,2}] &\sim \frac{1}{n} [l_1 - l_2] + \frac{1}{2n\mu} \int \tilde{\zeta}_1(\lambda) d\lambda - \frac{1}{2n\mu} \int \tilde{\rho}_1(\lambda) d\lambda \\ &- \frac{C_L \mu^{2L}}{4} \int \left| \frac{\partial^L f_S(\lambda) / \partial \lambda^L}{f_S(\lambda)} \right|^2 d\lambda - \frac{C_L \mu^{2L}}{4} \int \left| \frac{\partial^L f_T(\lambda) / \partial \lambda^L}{f_T(\lambda)} \right|^2 d\lambda + \frac{C_L \mu^{2L}}{2} \int \frac{f_S(\lambda)}{f_T(\lambda)} \left| \frac{\partial^L f_T(\lambda) / \partial \lambda^L}{f_T(\lambda)} \right|^2 d\lambda. \end{aligned} \quad (8)$$

4. CLASSIFICATION PERFORMANCE

In this section, we analyze classification performance of the system described in Section 2, based upon the first order asymptotic approximations. A quick analysis of (8) (and the analogous expression under H_2) shows that for classification performance to be close to the ideal, the mesh size $\mu(n)$ must be of the order of $o(n^{-1})$. To find the optimal rate of sieve growth in the classification sense, we apply the large deviation theory to obtain the asymptotic exponential rates for the probability of false alarm and the probability of missed detection defined as $P_{FA}(\gamma) = Pr \left\{ \frac{\hat{l}}{n} > \gamma | H_2 \right\}$ and $P_{MD}(\gamma) = Pr \left\{ \frac{\hat{l}}{n} < \gamma | H_1 \right\}$, respectively, where γ is a recognition threshold.

The large deviation rate function is the Fenchel-Legendre transform given by $I(t) = \sup_s [st - \bar{\phi}(s)]$, where $\bar{\phi}(s)$ is the asymptotic log-moment generating function and s is a real-valued parameter. The asymptotic log-moment generating functions of the ideal test statistic under hypotheses H_1 and H_2 , $\bar{\phi}_1(s) = \lim_{n \rightarrow \infty} \frac{1}{n} \log E[\exp(sl) | H_1]$, and $\bar{\phi}_2(s) = \lim_{n \rightarrow \infty} \frac{1}{n} \log E[\exp(sl) | H_2]$, can be obtained in closed form by invoking Toeplitz theory.

Let γ_1 denote all the terms to the right from $\frac{1}{n} [l_1 - l_2]$ in (8) and γ_2 be the analogous notation under H_2 . Then the asymptotic log-moment generating functions for \hat{l} are simply $\bar{\phi}_1$ and $\bar{\phi}_2$ shifted by γ_1 and γ_2 , i.e. $\hat{\phi}_1(s) = \bar{\phi}_1(s) + s\gamma_1$ and $\hat{\phi}_2(s) = \bar{\phi}_2(s) + s\gamma_2$. This yields the following relationship between the rate functions

$$\hat{I}_1(t) = \sup_s [st - \bar{\phi}_1(s) - s\gamma_1] = I_1(t - \gamma_1), \quad (9)$$

$$\hat{I}_2(t) = \sup_\theta [\theta t - \bar{\phi}_2(\theta) - \theta\gamma_2] = I_2(t - \gamma_2), \quad (10)$$

with $I_2(t) = I_1(t) + t$.

For the average classification error to be as small as possible, the rate functions $\hat{I}_1(t)$ and $\hat{I}_2(t)$ have to be well separated. This takes place when the distance between γ_1 and γ_2 is maximized, i.e.

$$\mu^*(n, L) = \arg \max_{\mu} (|\gamma_1 - \gamma_2|). \quad (11)$$

This results in the following value for the optimal μ^*

$$\begin{aligned} [\mu^*(n, L)]^{(2L+1)} &= (nL)^{-1} \left[\frac{1}{2} \int (\tilde{\rho}(\lambda) + \tilde{\eta}(\lambda) - \tilde{\zeta}_1(\lambda) - \tilde{\zeta}_2(\lambda)) d\lambda \right] \\ &\times \left[C_L \int \left\{ \left| \frac{\partial^L f_S(\lambda) / \partial \lambda^L}{f_S(\lambda)} \right|^2 \left[\frac{f_S(\lambda) - f_T(\lambda)}{f_S(\lambda)} \right] + \left| \frac{\partial^L f_T(\lambda) / \partial \lambda^L}{f_T(\lambda)} \right|^2 \left[\frac{f_T(\lambda) - f_S(\lambda)}{f_T(\lambda)} \right] \right\} d\lambda \right]^{-1}. \end{aligned} \quad (12)$$

Here $\{\zeta_2(k)\}$ and $\{\rho_2(k)\}$ are the eigenvalues of the matrix products

$$[J^{-1/2}(f_T, \tilde{\mathbf{a}}_{\mu,2}) I(f_T, \tilde{\mathbf{a}}_{\mu,2}) J^{-1/2}(f_T, \tilde{\mathbf{a}}_{\mu,2})],$$

and

$$[J^{1/2}(f_T, \tilde{\mathbf{a}}_{\mu,1})J^{-1}(f_S, \tilde{\mathbf{a}}_{\mu,1})I(f_S, \tilde{\mathbf{a}}_{\mu,1})J^{-1}(f_S, \tilde{\mathbf{a}}_{\mu,1})J^{1/2}(f_T, \tilde{\mathbf{a}}_{\mu,1})],$$

respectively.

The expression (12) shows that the optimal rate of sieve growth for the recognition problem differs from the optimal rate of sieve growth obtained in [12] for the function estimation problem. The optimal mesh size for the recognition problem is obtained in a simplified setting, under a first order approximation. In a general setting, the expression for the optimal mesh size is expected to be more complex.

5. NUMERICAL SOLUTION

In more realistic situations, the true parameters $f_S(\cdot)$ and $f_T(\cdot)$ are not known, and the training vectors have a finite length. A typical approach to finding the optimal parameter μ in this case is, for every μ from a selected range of those, to estimate the average classification performance and select the parameter μ, μ^* , that minimizes the estimated probability of error. This requires finding restricted ML estimates of the functions $f_S(\cdot)$ and $f_T(\cdot)$ from the vectors \mathbf{S} and \mathbf{T} , respectively, for each given μ . Below we provide an iterative solution to evaluating restricted ML estimates.

Suppose that the spline sieve is designed as it is described in Subsection 2.1. Given a value of the mesh-size, the problem of function estimation (infinite dimensional parameter) is reduced to a problem of estimating of a finite number of parameters. The loglikelihood function for the n -dimensional vector \mathbf{S} , when the estimate is sought in the subset $\mathcal{F}_\mu, \mu = Q^{-1}$, is given by

$$l_\mu(\mathbf{S}) = -\mathbf{S}^\dagger \mathbf{K}^{-1}(f_{\mu,S})\mathbf{S} - \log \det \mathbf{K}(f_{\mu,S}), \quad (13)$$

where $f_{\mu,S}$ is a power spectral density function restricted to be in the subset \mathcal{F}_μ , i.e. having the representation $f_{\mu,S} = \sum_{m \in \Lambda_\mu} a(m)\psi_m(\lambda)$, with unknown parameters $\{a(m)\}$. From here, the matrix $\mathbf{K}(f_{\mu,S})$, the Toeplitz covariance matrix parameterized by $f_{\mu,S}$, can be represented as a linear combination of the Toeplitz matrices Ψ_m with entries given by the Fourier transform of m -th B-spline, that is

$$\mathbf{K}(f_{\mu,S}) = \sum_{m \in \Lambda_\mu} a(m)\Psi_m, \quad (14)$$

$$\Psi_m = \begin{bmatrix} t_{m,0} & t_{m,1} & t_{m,2} & \dots & t_{m,n-1} \\ t_{m,-1} & t_{m,0} & t_{m,1} & \dots & t_{m,n-2} \\ \cdot & \cdot & \cdot & \dots & \cdot \\ t_{m,-n+1} & t_{m,-n+2} & t_{m,-n+3} & \dots & t_{m,0} \end{bmatrix}$$

and

$$t_{m,k} = \int_{-1/2}^{1/2} \psi_m(\lambda) \exp(2\pi j \lambda k) d\lambda.$$

To find the maximum likelihood solution, a derivative of $l_\mu(\mathbf{S})$ has to be taken with respect to the parameters $\{a(m)\}$. The necessary conditions for a local maximizer can be derived (for details see [12]). However, in this case the maximizer is a solution to a system of nonlinear equations and cannot be written in closed form. As an alternative, an iterative solution converging to the maximizer can be obtained by applying the expectation-maximization (EM) method [3,10].

An EM algorithm requires existence of two data sets: complete data set and incomplete data set. The latter is simply observed data. The complete data set has to be selected such that the loglikelihood function of the set is possibly convenient to manipulate; and there exists a function that maps the complete data set into incomplete data set. Since the complete data are not known, a sequence of conditional expected loglikelihood functions for complete data, given incomplete data, is formed. The functions are updated iteratively with updating the parameters. Under mild regularity conditions imposed on the incomplete loglikelihood function, the sequence of parameters converges to a stationary point. The algorithm is called alternating algorithm because of the two steps involved in its implementation that are used alternatively: taking the conditional expectation of the complete loglikelihood and maximizing the resulting function with respect to the parameters.

For the spline sieve, we define the complete data set to be consisting of Q independent n -dimensional vectors \mathbf{S}_m , $m \in \Lambda_\mu$, each drawn from the complex Gaussian distribution with zero mean and covariance matrix $[a(m)\Psi_m]$. Then the mapping between the complete and incomplete data is given by

$$\mathbf{S}_{id} = \sum_{m=1}^Q \mathbf{S}_m,$$

where subscript *id* denotes ‘‘incomplete data.’’ The notation *cd* will be used for ‘‘complete data.’’

The loglikelihood function for the complete data is the sum of the loglikelihoods of the vectors \mathbf{S}_m

$$l_{cd}(\mathbf{S}_1, \dots, \mathbf{S}_Q) = - \sum_{m=1}^Q \mathbf{S}_m^\dagger [a(m)\Psi_m]^{-1} \mathbf{S}_m - \sum_{m=1}^Q \log \det [a(m)\Psi_m]. \quad (15)$$

By taking the expectation of the right side in the equation (15), given the incomplete data and the estimates obtained over the previous (assume k -th) step, we obtain

$$E[l_{cd}|\mathbf{S}_{id}, \hat{\mathbf{a}}^{(k)}] = - \sum_{m=1}^Q n \log a(m) - \sum_{m=1}^Q \frac{1}{a(m)} E[\mathbf{S}_m^\dagger \Psi_m^{-1} \mathbf{S}_m | \mathbf{S}_{id}, \hat{\mathbf{a}}^{(k)}],$$

where $\hat{\mathbf{a}}^{(k)}$ is the abbreviated notation for $\{\hat{a}^{(k)}(1), \dots, \hat{a}^{(k)}(Q)\}$.

Differentiating the conditional complete loglikelihood with respect to $a(m)$, we obtain the following update-equation for the estimate of $a(m)$ after $(k+1)$ iterations

$$\begin{aligned} a^{(k+1)}(m) &= \frac{1}{n} \text{Tr} \left\{ \Psi_m^{-1} E[\mathbf{S}_m \mathbf{S}_m^\dagger | \mathbf{S}_{id}, \hat{\mathbf{a}}^{(k)}] \right\} \\ &= \frac{1}{n} \text{Tr} \left\{ \Psi_m^{-1} \text{cov}[\mathbf{S}_m | \mathbf{S}_{id}, \hat{\mathbf{a}}^{(k)}] + \Psi_m^{-1} E[\mathbf{S}_m | \mathbf{S}_{id}, \hat{\mathbf{a}}^{(k)}] E^\dagger[\mathbf{S}_m | \mathbf{S}_{id}, \hat{\mathbf{a}}^{(k)}] \right\}. \end{aligned}$$

To find the conditional expectation and covariance, we appeal to the conditional probability density function for the vector \mathbf{S}_m , given the incomplete data and the estimates obtained over the previous iteration

$$p(\mathbf{S}_m | \mathbf{S}_{id}, \hat{\mathbf{a}}^{(k)}) \sim \exp\left\{ -(\mathbf{S}_{id} - \mathbf{S}_m)^\dagger \left[\sum_{l \neq m} a^{(k)}(l) \Psi_l \right]^{-1} (\mathbf{S}_{id} - \mathbf{S}_m) - \mathbf{S}_m^\dagger [a(m)\Psi_m]^{-1} \mathbf{S}_m \right\},$$

where we dropped all terms irrelevant to the computation of the conditional expectation and covariance. Completing the squares results in the following expressions for the conditional expectation and covariance

$$\text{cov}(\mathbf{S}_m | \mathbf{S}_{id}, \hat{\mathbf{a}}^{(k)}) = \left(\left[\sum_{l \neq m} a^{(k)}(l) \Psi_l \right]^{-1} [a^{(k)}(m)\Psi_m]^{-1} \right)^{-1}, \quad (16)$$

$$E(\mathbf{S}_m | \mathbf{S}_{id}, \hat{\mathbf{a}}^{(k)}) = (\mathbf{I} + \left[\sum_{l \neq m} a^{(k)}(l) \Psi_l \right] [a^{(k)}(m)\Psi_m]^{-1})^{-1} \mathbf{S}. \quad (17)$$

After substituting (16) and (17) in the update equation, we obtain the estimate of the parameter $a(m)$ at $(k+1)$ iteration of the algorithm

$$a^{(k+1)}(m) = a^{(k)}(m) + \frac{1}{n} a^{(k)}(m) \text{Tr} \left\{ \left(\sum_{l=1}^Q a^{(k)}(l) \Psi_l \right)^{-1} [\mathbf{S} \mathbf{S}^\dagger - \sum_{l=1}^Q a^{(k)}(l) \Psi_l] \left(\sum_{l=1}^Q a^{(k)}(l) \Psi_l \right)^{-1} \Psi_m \right\}. \quad (18)$$

6. SUMMARY

We consider the binary classification problem where the stationary Gaussian data are parameterized by unknown positive power spectral densities. We assume that the unknown spectral densities are estimated using training sets of a limited size, independent of the testing data. The recognition system implements the loglikelihood ratio test with estimated parameters placed instead of the true but unknown parameters. Grenander’s method of sieves is applied to regularize the estimates, using the spline sieve designed and studied in [11,12]. To analyze the recognition performance we first decomposed the loglikelihood ratio with estimated parameters into five components including estimation and approximation error terms. Then we analyzed each component separately. Using asymptotic first order expansions of these components, we find the optimal rate of sieve growth for the recognition problem.

7. ACKNOWLEDGMENTS

This work was supported in part by the US Army Research Office grant DAAH 04-95-1-04-94, by the Office of Naval Research grant N00014-98-1-06-06, and by the Boeing-McDonnell Foundation.

REFERENCES

1. H. Akaike, "Information Theory and an Extension of the Likelihood Principle," in *Proceedings of the Second International Symposium of Information Theory*, ed. B. N. Petrov and F. Csaki, Budapest: Akademiai Kiado, 1973.
2. Yu. Chow and U. Grenander, "A Sieve Method for the Spectral Density," *The Annals of Statistics*, v. 13, no. 3, 1985, pp. 998-1010.
3. A. D. Dempster, N. M. Laird, and D. B. Rubin, "Maximum-Likelihood from Incomplete Data via EM Algorithm," *J. Roy. Statist. Soc.*, v. B39, no. 1, pp. 1-37, 1977.
4. T. G. Dietterich and E. B. Kong "Machine Learning Bias, Statistical Bias, and Statistical Variance of Decision Tree Algorithms," Dept. of Computer Science, Oregon State University Technical Report, 1997.
5. J. H. Friedman, "On Bias, Variance, $0/1$ - loss, and the Curse-of-Dimensionality," Dept. of Statistics, Stanford University Technical Report, 1996.
6. U. Grenander, *Abstract Inference*, John Wiley and Sons, New York, 1981.
7. P. J. Huber, "The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions," *Proc. 5th Berkeley Symp. Math. Statist. and Probab.* Los Angeles, CA: Univ. of California Press, vol. 1, 1967, pp. 221-233.
8. P. J. Huber, *Robust Statistics*, John Wiley and Sons, New York, 1981.
9. R. Kohavi and D. H. Wolpert, "Bias Plus Variance Decomposition for Zero-One Loss Functions," *Proceedings of the Thirteenth International Conference on Machine Learning*, 1996.
10. G. J. McLachlan, T. Krishnan, *The EM algorithm and extensions*, John Wiley and sons, New York, 1997.
11. P. Moulin, J. A. O'Sullivan, and D. L. Snyder, "A Method of Sieves for Multiresolution Spectrum Estimation and Radar Imaging," *IEEE Trans. on Info. Theory*, v. 38, no. 2, March 1992, pp. 801-813.
12. P. Moulin, "A Method of Sieves for Radar Imaging and Spectrum Estimation," D.Sc. Dissert., Washington Univ., St. Louis, MO, 1990.
13. J. A. O'Sullivan, R. E. Blahut, and D. L. Snyder, "Information Theoretic Image Formation," *IEEE Trans. Info. Theory*, v. 44, no. 6, Oct. 1998, pp. 2094-2123.
14. N. A. Schmid, "Performance Analysis in Stochastic Recognition and Authentication," D.Sc. Dissert., Washington Univ., St. Louis, MO, 2000.
15. L. L. Schumaker, *Spline Functions: Basic Theory*, John Wiley and Sons, New York, 1981.
16. X. Shen, "On Method of Sieves and Penalization," *The Annals of Statistics*, vol. 25, no. 6, pp. 2555-2591.
17. J. R. Thompson and R. A. Tapia, *Nonparametric Function Estimation, Modeling, and Simulation*, SIAM, Philadelphia, 1990.
18. R. Tibshirani, "Bias, Variance, and Prediction Error for Classification Rules," Dept. of Statistics, University of Toronto Technical Report, 1996.
19. H. White, "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, v. 50, no. 1, 1982, pp. 1-25.
20. S. S. Wilks, *Mathematical Statistics*, John Wiley and Sons, New York, 1962.