A Signal Theory Approach to Support Vector Classification: The Sinc Kernel*

James D. B. Nelson, Robert I. Damper, Steve R. Gunn
and Baofeng Guo

Information: Signals, Images, Systems Research Group,

School of Electronics and Computer Science,

University of Southampton,

Southampton SO17 1BJ, UK

Tel +44 (0)23 8059 5000

{jn|rid|srg|bg}@ecs.soton.ac.uk

*This research was supported by the Data and Information Fusion (DIF) Defence Technology Centre, United Kingdom, under DTC Project 8.2, funded by the UK Ministry of Defence and managed by General Dynamics Limited and QinetiQ.

Abstract

Fourier-based regularisation is considered for the support vector machine classification problem over absolutely integrable loss functions. By invoking the modest assumption that the decision function belongs to a Paley-Wiener space, it is shown that the classification problem can be developed in the context of signal theory. Furthermore, by employing the Paley-Wiener reproducing kernel, namely the sinc function, it is shown that a principled and finite kernel hyper-parameter search space can be discerned, *a priori*. Subsequent simulations performed on a commonly-available hyperspectral image data set reveal that the approach yields results that surpass state-of-the-art benchmarks.

Keywords: hyperspectral imaging, parameter estimation, regularisation, reproducing kernel Hilbert spaces, sequency analysis, signal theory, sinc kernel, support vector machines

List of Symbols

$\mathbb{R}$	set of real numbers
$\mathbb{N}$	set of natural numbers
$\mathbb{Z}$	set of integers
X	input space
$\mathcal{H}, \mathcal{F}$	Hilbert spaces
k	reproducing kernel function
Γ	regularisation operator
sgn	signum functional
φ	informative part of data
ε	non-informative part of data
$\bigoplus$	direct sum
Ω^*	frequency support of φ
PW	Paley-Wiener function space
$\wedge$	Fourier transform operator
$\langle\cdot,\cdot\rangle$	inner product
$\overline{\cdot}$	complex conjugate
cal	$\operatorname{sgn}\cos$
sal	$\operatorname{sgn}\sin$
μ	Möbius function
$n \mid m$	n divides m
$\delta_{\cdot,\cdot}$	Kronecker’s delta
$\delta(\cdot)$	Dirac’s delta
$\sim$	sequency transform operator
S	sequency space
$\cdot * \cdot$	convolution operator
ω_*	(unknown) optimal kernel parameter

1 Introduction

An often-cited property of the support vector machine (SVM) learning method is the existence of a unique solution. Another very desirable attribute, namely flexibility, is readily realised by the introduction of non-linear kernel methods. But herein lies a conflict. Although flexibility admits richness, it also introduces parameters, and thereby precludes uniqueness. Whether the parameter takes the form of a scaling vector, a scaling number, or the kernel itself, the fact remains that in the context of non-linear support vector machines there are uncountably many solutions. Unfortunately, the only way to determine the best solution is to build uncountably many kernels. This is, of course, intractable.

However, when framed in the context of reproducing kernel Hilbert spaces, it has been shown by Girosi (1998) that the choice of kernel and parameters control the nature and degree of regularisation that is imposed on the solution. A related issue is that the so-called curse of dimensionality often turns out not to have the detrimental effect that is predicted. Some recent machine learning research has focused on finding cogent explanations for this phenomenon. Belkin and Niyogi (2004) argue that a possible reason is that the data lie on a sub-manifold, embedded in the input space. Indeed, data with a large number of variables may lie entirely in a much smaller-dimensional manifold. Knowledge pertaining to the structure of the manifold can be used to guide the choice of parameters, and thus the nature and degree of regularisation. Such realisations lead to a more considered approach: that is to ascertain, *a priori*, properties of the space wherein the data lie. Although there may still exist infinitely many solutions, the range of an empirical search could then at least be focused upon subsets of parameters rather than all possible choices of parameters.

We propose a principled way of reducing the infinite parameter search space to an exhaustive and finite one. Our approach is motivated by sampling theory, where the main goal is to establish equivalence relations between data sequence spaces and kernel function spaces. To this end, we employ perhaps the most elementary function space from sampling theory, namely the simply connected and zero-centred Paley-Wiener reproducing kernel Hilbert space, more commonly referred to by engineers as baseband-limited signals. For a given class of data, we show how to estimate, *a priori*, a suitable kernel and parameter subspace. Smale and Zhou (2004) have also studied the application of sampling theory and reproducing kernel Hilbert spaces to learning theory. They consider the least squares loss regression problem and construct probability estimates for the sampling error. The work reported here adds to the rather small amount of literature on this under-explored topic.

The remainder of this paper is structured as follows. In Section 2, the data class under consideration and its corresponding reproducing kernel Hilbert space are constructed. Accordingly, some necessary signal theory concepts are introduced and discussed in Section 3, and exploited in Section 4. Finally, in Section 5, we report the best results to date on a popular hyperspectral image data set, confirming the power and utility of the approach.

2 Model Construction

Let $x_n \in X \subseteq \mathbb{R}^d, y_n \in \{\pm 1\}$, $n \in \mathbb{N}$, and consider the usual SVM classification problem

$$\min_{f \in \mathcal{H}} \frac{1}{2} \|\Gamma f\|^2 + C \sum_{n=1}^{N} |1 - y_n f(x_n)|_+ \,, \tag{1}$$

where f, the decision function to be determined in some Hilbert space $\mathcal{H}(X)$, is regularised by the operator $\Gamma\colon \mathcal{H} \mapsto \mathcal{F}$ that maps the input space to the desired feature space. The resulting learned decision function, implied by the representer theorem (Kimeldorf and Wahba, 1970), is the solution

$$f = \sum_{n=1}^{N} y_n \alpha_n k(x_n, \cdot) \,, \tag{2}$$

where k is a Mercer kernel (Mercer, 1909). Herewith, the classifier is defined by $\operatorname{sgn} f$. Our main contention is that before any effort is made to design the classifier, it is good practice, in a qualitative sense, to attempt to discern the properties of the underlying decision function. A natural preface, proposed here, is that the labelling function maps d-variate data to labels via $y\colon \mathbb{R}^d \supset X \mapsto \{\pm 1\}$, with

$$y(x) := \operatorname{sgn}\big(\varphi(x) + \varepsilon(x)\big), \tag{3}$$

where the noise is modelled by ε, and under the assumption that the information content φ, lies entirely within the space of Paley-Wiener functions over some multi-dimensional baseband region Ω^*, viz.

$$\Omega^* := \bigoplus_{r=1}^{d} \Omega_r^* := \bigoplus_{r=1}^{d} (-\omega_*^r \pi, \omega_*^r \pi)\,.$$

That is

$$\varphi \in PW_{\Omega^*} := \bigoplus_{r=1}^{d} \big\{\zeta \in L_2(X) : \operatorname{supp} \zeta^{\wedge} \subseteq \Omega_r^*\big\}\,, \tag{4}$$

with $\operatorname{supp} \zeta := \{x \in X : \zeta(x) \neq 0\}$, and where $\cdot^{\wedge}$ denotes Fourier transformation:

$$\zeta^{\wedge}(\omega) := \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \zeta(x) e^{-i\langle \omega, x\rangle}\, \mathrm{d}x.$$

The condition $\varphi \in PW_{\Omega^*}$ restricts the behaviour of the information content to functions of finite bandwidth around the origin.

Since the time of Hardy (1941), it has been known that the orthogonal function for the band-limited space $PW_{(-B\pi, B\pi)}$, with $B > 0$, is that function nowadays commonly known as the sinc kernel, defined by

$$\mathrm{sinc}_B(\cdot - x) := \frac{\sin\big(B\pi(\cdot - x)\big)}{B\pi(\cdot - x)} \quad .$$

Indeed, Higgins (1985) has suggested that the origins of this orthogonal system may well go as far back as Borel (1897). Although this kernel is familiar to signal theorists and engineers, it is a seemingly rare tool in machine learning. Kon, Raphael, and Williams (2005) make a brief mention of it, by way of an example, in their work on approximation estimates and statistical learning theory. Sugiyama and Müller (2002) use the sinc kernel, among other choices, to demonstrate that their generalisation bound for regression is stable with respect to kernel choice. It is perhaps less well known that, by virtue of the following three established results, the sinc kernel also lends itself to the regularised support vector classification setting.

Theorem 2.1 *(Self-consistency property, Smola, Schölkopf, and Müller, 1998.) Let the Mercer kernel defined by $k\colon X \times X \mapsto \mathbb{R}$, and the regularisation operator $\Gamma\colon \mathcal{H} \mapsto \mathcal{F}$, be such that $k(x,\xi) \equiv \big\langle (\Gamma k)(x), (\Gamma k)(\xi) \big\rangle_{\mathcal{F}}$. Then the SVM classification problem can be written*

$$\min_{f \in \mathcal{H}} \frac{1}{2} \|\Gamma f\|^2 + C \sum_{n=1}^{N} |1 - y_n f(x_n)|_+ \ ,$$

as earlier (Equation 1).

Theorem 2.2 *(Translation invariant kernels, Smola, Schölkopf, and Müller, 1998.) Consider a kernel, endowed with translation invariance, namely $k(x,\xi) = k(x-\xi)$, with the regularisation operator $\Gamma\colon \mathcal{H} \mapsto \mathcal{F}$, defined by*

$$\langle \Gamma f, \Gamma g \rangle_{\mathcal{F}} = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \frac{f^{\wedge}(\omega)\overline{g^{\wedge}(\omega)}}{k^{\wedge}(\omega)} \mathrm{d}\omega.$$

Then $k(x,\xi) \equiv \big\langle (\Gamma k)(x), (\Gamma k)(\xi) \big\rangle_{\mathcal{F}}$, and the self-consistency property of Theorem 2.1 is satisfied.

Corollary 2.3 *It follows from Theorem 2.2 and the work of Aronszajn (1950) on tensor products of reproducing kernels that the regularisation term from the SVM problem is*

$$\|\Gamma f\|_{\mathcal{F}}^2 = \frac{1}{(2\pi)^{d/2}} \prod_{r=1}^{d} \int_{\Omega_r^*} \frac{|f^\wedge(\omega)|^2}{k_r^\wedge(\omega^r)} \, \mathrm{d}\omega^r,$$

with $\omega := (\omega^r)_{r=1}^d$, *and that*

$$\frac{1}{k^\wedge(\omega)} = \left(\prod_{r=1}^{d} k_r^\wedge(\omega^r) \right)^{-1}$$

regularises the decision function f *by acting as a filter, in the signal analysis sense, on* $|f^\wedge|^2$.

The unique kernel associated with the reproducing kernel Hilbert space PW_{Ω^*} is the sinc kernel

$$k(x,\xi) := \prod_{r=1}^{d} k_{\omega_*^r}(x^r, \xi^r) := \prod_{r=1}^{d} \mathrm{sinc}_{\omega_*^r}(x^r - \xi^r). \tag{5}$$

Given the model (3), where the information content is embedded in the Paley-Wiener space (4), it is only sensible to constrain the decision function to the same Paley-Wiener space. From Corollary 2.3, it follows that in the Fourier domain the multiplicative filter that acts upon $|f^\wedge|^2$ is

$$\frac{1}{k^\wedge(\omega)} = \frac{1}{\chi_{\Omega^*}(\omega)} = \prod_{r=1}^{d} \frac{1}{\chi_{\Omega_r^*}(\omega^r)},$$

with the d-dimensional hypercuboid

$$\chi_\Omega(\omega) := \begin{cases} 1 & \text{if } \omega \in \Omega \\ 0 & \text{otherwise} \end{cases}. \tag{6}$$

In this case, since $k^\wedge \geq 0$ holds over $\mathbb{R}^d$, Bochner's theorem (Bochner, 1959) ensures that the sinc kernel is a Mercer kernel. The multiplicative filter regularises the decision function by penalising the frequency content of f on $\mathbb{R}\backslash\Omega^*$. The sinc kernel also keeps the content over Ω^* unaltered. These penalisation and preservation properties are, by definition, unique to the sinc kernel. Since Paley-Wiener spaces are closed under addition, the representer result (2) ensures that the decision function is restricted to PW_{Ω^*}.

Remark 2.4 *We now see that, in the context of our work, the non-regularised, higher-dimensional input space discussed by Belkin and Niyogi (2004) is $PW_{\mathbb{R}^d}$, and the sub-manifold is $PW_{\Omega^*} \subseteq PW_{\mathbb{R}^d}$. That is, in the frequency domain, the sub-manifold invoked by our work can be described as a hypercuboid centred on the origin, and the regularising operator is precisely the mapping $\Gamma\colon PW_{\mathbb{R}^d} \mapsto PW_{\Omega^*}$.*

We are now left with the problem of finding an optimal hyper-parameter set $\{\omega_*^r\}$, in the sense of the SVM problem. Before this is attempted, we propose a novel approach to elicit spectral properties of the labelling function that employs some recently-constructed tools from signal theory.

3 From Signal Theory to SVM Classification

Intuitively, the labelling function y of equation 3 can be understood as a piecewise-constant function that maps d-many real variables to positive or negative unity. It can, therefore, be treated as a square-wave function over d-variate space. To this end, we propose the use of sequency analysis as a means to elicit some properties of y and, consequently, the information content φ. Such properties will suggest how the decision function should be regularised. Before the analysis, it is instructive to introduce a family of functions that has the labelling function as a member.

Definition 3.1 *Let the* cal *and* sal *functions be defined by*

$$\begin{aligned}\mathrm{cal}_\omega(t) &:= \mathrm{sgn}\cos\omega t\,,\\ \mathrm{sal}_\omega(t) &:= \mathrm{sgn}\sin\omega t\,.\end{aligned}$$

Now, define the complex square-wave family as

$$\psi_\omega := \sqrt{\frac{\pi}{32}}\,(\mathrm{cal}_\omega + i\,\mathrm{sal}_\omega)\,.$$

This definition is consistent with the construction given by Elliot and Rao (1982), Hughes and Heron (1989), and Nelson (2002). However, this basis, and therefore the definition of sequency, differs from the more common Walsh-Hadamard analysis described elsewhere, such as Beer (1981). In particular, the Walsh-Hadamard system, defined over a dyadic grid, constitutes an orthogonal basis. On the other hand, the system employed here is defined over a denser, uniform grid and, as will be shown below, it forms a biorthogonal basis. As such, it can be used to analyse the spectral properties of functions over a more opaque domain. Now, consider the Möbius arithmetic function $\mu\colon \mathbb{N} \mapsto \{0, \pm 1\}$, given by

$$\mu(n) := \begin{cases} 1, & \text{if } n = 1 \\ (-1)^m, & \text{if } n \text{ is the product of } m \text{ distinct primes} \\ 0, & \text{otherwise} \end{cases},$$

which is employed here due to the utility afforded by the following result, taken from number theory:

Lemma 3.2 *(Möbius.) Let μ denote the Möbius function. Then, for $m \in \mathbb{N}$,*

$$\sum_{n|m} \mu(n) = \delta_{m,1}\,,$$

where $\delta_{\cdot,\cdot}$ denotes the Kronecker delta. The next result, outlined by Nelson (2001), enables us to express the labelling function in terms of the complex square-wave family.

Proposition 3.3 *(Biorthogonal complex square-wave system, Nelson 2001.) The biorthogonal dual of* $\{\psi_n\}$ *is*

$$\psi_n^*(t) := \frac{1}{\sqrt{2\pi}} \sum_{m \in 4\mathbb{Z}+1} m^{-1} \mu(|m|) e^{int/m}.$$

Proof We require $\langle \psi_n, \psi_j^* \rangle_{L_2(\mathbb{R})} = \delta_{n,j}$. Since the complex square wave ψ_n is periodic, it can be expanded as the Fourier series

$$\psi_n(t) = \frac{1}{\sqrt{2\pi}} \sum_{m \in 4\mathbb{Z}+1} \frac{1}{m} e^{imnt}.$$

Hence

$$\begin{aligned} \langle \psi_n, \psi_j^* \rangle_{L_2(\mathbb{R})} &= \int_{\mathbb{R}} \psi_n(t) \overline{\psi_j^*(t)} \, \mathrm{d}t \\ &= \frac{1}{2\pi} \sum_{m,\ell \in 4\mathbb{Z}+1} \frac{1}{m\ell} \mu(|\ell|) \int_{\mathbb{R}} e^{i(mn - j/\ell)t} \, \mathrm{d}t. \end{aligned}$$

The integral over $\mathbb{R}$ can be written as

$$\begin{aligned} \lim_{\tau \to 0} \int_{-\pi/\tau}^{\pi/\tau} e^{i(mn - j/\ell)x} \, \mathrm{d}x &= 2\pi \lim_{\tau \to 0} \left[\frac{1}{\tau} \operatorname{sinc}_{\tau^{-1}}(mn - j/\ell) \right] \\ &= 2\pi \delta(mn - j/\ell), \end{aligned}$$

where the $\delta(\cdot)$ denotes the Dirac delta generalised function, the non-zero values of which can be found by taking $mn = j/\ell$. For then

$$\langle \psi_n, \psi_j^* \rangle_{L_2(\mathbb{R})} = \delta(0) \frac{n}{j} \sum_m \sum_{\ell | j/n} \mu(|\ell|).$$

Lemma 3.2 implies

$$\sum_{\ell | j/n} \mu(|\ell|) = \begin{cases} 1, & \text{for } j = n \\ 0, & \text{otherwise} \end{cases}$$

and, hence, the non-zero values exist when $j = n$. Since $j = n$ implies that $m = 1/\ell$, it follows that the sum over m collapses to the sole term $m = 1$, and we have

$$\langle \psi_n, \psi_k^* \rangle_{L_2(\mathbb{R})} = \delta_{n,j} \delta(0)\,.$$

■

The discrepancy $\delta(0)$ occurs because, as Higgins (1996, p. 29), explains, "… the point evaluation functional is not properly defined on L_2 spaces". Now that the biorthonormal square-wave system has been established, we introduce the sequency transformation $\cdot^{\sim}$, namely

$$f^{\sim}(\omega) = \int_{\mathbb{R}} f(t) \overline{\psi_\omega^*(t)}\, \mathrm{d}t. \tag{7}$$

From Proposition 3.3, it follows that y can be expanded as a superposition of square waves,

$$y = \sum_{n \in \mathbb{Z}} \langle y, \psi_n^* \rangle_{L_2(\mathbb{R})} \psi_n.$$

Hence, the coefficients that express y in terms of the square-wave basis are found by performing the sequency transform of y. Recall from (3) that $\varphi \in PW_{\Omega^*}$, and, without loss of generality, $\varepsilon \in PW_{\Omega^+}$. The linearity property of Paley-Wiener spaces gives rise to

$$\varphi + \varepsilon \in PW_{\Omega^* \cup \Omega^+}.$$

We define the sequency function space S_Ω as

$$S_\Omega := \{\zeta \in L_2(X) : \operatorname{supp} \zeta^\sim \subseteq \Omega\},$$

Now, since $\varphi \in PW_{\Omega^*} \Rightarrow \operatorname{sgn}\varphi \in S_{\Omega^*}$, and $\varepsilon \in PW_{\Omega^+} \Rightarrow \operatorname{sgn}\varepsilon \in S_{\Omega^+}$, we can express the labelling function y as a sequency-limited function, $y = \operatorname{sgn}(\varphi+\varepsilon) \in S_{\Omega^* \cup \Omega^+}$, that is,

$$y = \int_{\Omega^* \cup \Omega^+} y^\sim(\omega)\psi_\omega(\cdot)\,\mathrm{d}\omega, \tag{8}$$

and where $y^\sim$ can be computed via

$$\begin{aligned} y^\sim(\omega^r) &= \frac{1}{\sqrt{2\pi}} \sum_{m\in 4\mathbb{Z}+1} \frac{\mu(|m|)}{m} \int_{\mathbb{R}} y(t) e^{-i\omega x^r/m}\,\mathrm{d}t \\ &= \sum_{m\in 4\mathbb{Z}+1} \frac{\mu(|m|)}{m} y^\wedge\left(\frac{\omega^r}{m}\right), \end{aligned} \tag{9}$$

where one (fast) Fourier transform is required to determine $y^\wedge(\omega^r)$, for each $r = 1,\ldots,d$. Since the samples x_n^r over which the Fourier transforms of $y^\wedge(\omega^r)$ are computed are typically non-uniformly distributed, the direct application of a Fourier transform is inappropriate. Instead, irregular sampling techniques, such as those discussed by Gröchenig (1993),must be considered. Since a comprehensive treatment of irregular sampling issues is beyond the scope of this work, we employ here a simple strategy whereby the data are mapped to a uniform grid via nearest neighbour, constant interpolation.

By definition, the information content of $(\varphi+\varepsilon)$ lies in the frequency baseband $\Omega^* = (-\omega_*\pi, \omega_*\pi)$. Analogously, the informative part of the labelling function $\operatorname{sgn}(\varphi+\varepsilon)$ lies inside some sequency baseband $\Omega^* = (-\omega_*\pi, \omega_*\pi)$.

Example 3.4 *Consider $y = \operatorname{sgn}\varphi$, where $\varphi(t) = \cos\omega_* t$, and $t \in \mathbb{R}$. Clearly, it follows that $\varphi \in PW_{(-\omega_*,\omega_*)}$, and*

$$y^{\sim}(\omega) = \delta(\omega - \omega_*) + \delta(\omega + \omega_*) \Rightarrow y \in S_{(-\omega_*,\omega_*)}.$$

In this case, ω_* *is estimated from* $y^{\sim}$*, and* $\mathrm{sinc}(\omega_* \cdot)$ *is chosen as the kernel.*

In practice, the approach taken to determine Ω^*, and hence the value of ω_*, is not straightforward unless we assume that $\Omega^* \cap \Omega^+ = \{\}$. However, in this section we have formulated the SVM classification problem in terms of a signal theory one, namely that of filter design, and in Section 4 we show how this avoids the necessity of unduly repeated implementation of computationally-expensive parameter estimators such as cross-validation.

4 Parameter Estimation

For each choice of the parameter set ω_*, there is a corresponding reproducing kernel Hilbert space $\mathcal{H}_*$, say. Commonly, the parameter set (or hyper-parameter) is chosen by estimating the performance of the SVM for each parameter value. The value of $\omega_* = \{\omega_*^r\}_1^d$ that yields the best performance is then chosen as the optimal parameter.

4.1 State-of-the-Art

Chapelle *et al.* (2002) describe several different ways to measure SVM performance. To facilitate the empirical comparisons drawn in Section 5, we consider perhaps the most straightforward measure, namely the validation error. Here, the data are split into two distinct sets. One set is used to train and the other to validate the SVM.

There also exist several ways to search for the optimal parameter, ω_*. Often misused, the phrase ‘exhaustive search’ has been adopted to describe an approach whereby the performance measure is computed over a finite number of parameters. In practice, however, the search can never be truly exhaustive. Either the range of parameters is too small, or the discretisation too large, or both. Various gradient-descent search methods have also been applied to SVM parameter optimisation. Common drawbacks of gradient methods include finding a suitable smoothing strategy for the performance measure, choosing a good first initial point, and bad convergence.

Unfortunately, the inherent problems of any search-based method are exacerbated in an exponential manner as the number of parameters increases linearly, and when using a one-against-one strategy for example, in a combinatorial manner as the number of classes increases linearly. Only a few authors have attempted automatic estimation of the optimal hyper-parameter set. Lanckriet *et al.* (2004) use semi-definite programming techniques to compute the kernel matrix. Debnath and Takahashi (2004) attempt to make a link between the eigenvalues of the features and the optimal Gaussian parameter. However, their work relies almost entirely on empirical evidence and qualitative remarks. Wang *et al.* (2003) argue that the Gaussian parameter should be chosen with respect to a Fisher-discriminant-based measure. Guo *et al.* use mutual information theory to guide parameter selection (Guo *et al.*, 2005a) and parameter scaling (Guo *et al.*, 2005b).

4.2 Sinc Parameter Estimation

We propose a principled means to estimate a search space wherein the optimal parameter lies. Rather than blindly searching for a set of parameters by induction alone, we follow an approach inspired by the engineering discipline of filter design, catalogued by such works as Oppenheim and Schafer (1989). Although filter design is sometimes glibly described as ‘more of an art than a science’, it has a successful theoretical and practical history that arguably stretches further back than statistical machine learning. Not only does signal theory suggest parameters *a priori*, it can also (via spectral analysis) aid the interpretation of the underlying properties of a particular solution.

Our approach is to compute the sequency transform (7), via the series of fast Fourier transforms (9), in order to discern the interval Ω^*, from Equation (8). For a d-variate space $\Omega = \bigoplus_1^d \Omega_r$, we require d-many sequency transforms. When $\Omega_r^* = (-\omega_*^r\pi, \omega_*^r\pi)$ has been established, we use the estimate ω_*^r to construct the kernel described by (5) under the earlier-mentioned assumption that $\Omega^* \cap \Omega^+ = \{\}$.

4.2.1 Sinc Parameter Search Space

In practice, since each datum has finite length, the sequency transform (7) is taken over a finite domain T. From Equations (6) and (9) and the convolution theorem, this is equivalent to computing

$$(\chi_T y)^{\sim}(\omega) = \frac{T}{2\pi} \sum_{m \in 4\mathbb{Z}+1} \frac{\mu(|m|)}{m} \left(\mathrm{sinc}_T * y^{\wedge}\right)\left(\frac{\omega}{m}\right),$$

where $*$ denotes the convolution operator. Consequently, like the finite Fourier transform, the finite sequency transform is subject to so-called sinc ringing effects. Notwithstanding such artifacts, the sequency components can still be estimated. The shifted Dirac generalised functions found in the idealised and trivial Example 3.4 above are replaced by shifted sinc functions in the finite case. It follows that only the locations of the local maxima of $|y^{\sim}|$ should be considered as candidates for ω_*. Since y is necessarily restricted to a discrete and finite domain, the sequency spectrum is smooth and cannot take the same value at every point. Hence, only finitely many maxima will exist. This simple and intuitive argument serves to reduce an exhaustive but theoretically infinite search to an exhaustive, finite search.

A simple practical example, similar to the analytical Example 3.4, is given in Figure 1. We can see that both the Fourier and sequency transforms yield the correct maxima at 0.4 Hz. However, in this case the Fourier transform also gives rise to strong maxima at the 3rd and 5th harmonics of 0.4 Hz. By expanding the signal as a Fourier series, it is easily seen that, in general, there will be harmonics at $n \in (4\mathbb{Z}+1)$ times the fundamental frequency of 0.4 Hz.

[Figure 1 about here.]

For a one-dimensional problem, one merely tests the performance of the SVM by setting the parameter value to each local maximum of the sequency spectrum. To keep track of values that have or have not been tested and to ensure an orderly approach, one could, for example, conduct the search by first choosing the maximum that is located closest to the zero sequency, then work outwards to the second closest, and so on. To consider the generalisation to the d-dimensional case, it is helpful to consider the following construct.

Definition 4.1 *The sequence* $\{\omega_p\}_{p=1}^{P}$ *is defined as the set that contains the locations of the local maxima of* $|y^{\sim}(\omega)|$*, ordered such that* $\left\|\omega_p\right\|_2 \leq \left\|\omega_{p+1}\right\|_2$*, for all* $p = 1, \ldots, P$*.*

Herewith, the d-dimensional search would take place over the ordered finite set $\{\omega_p\}_{p=1}^{P}$.

4.2.2 Family of Search Strategies

Of course, when the number of dimensions or maxima preclude an exhaustive search over the entire set $\{\omega_p\}_{p=1}^{P}$, one may be compelled to compromise accuracy and either bound the search space, conduct a sparser search, or both. For example, a d-dimensional data set with m-many maxima in each dimension would have a total number of m^d maxima. For large d, an exhaustive search over all the maxima would be intractable. With this in mind, the construct from Definition 4.1 is modified.

Definition 4.2 *Define the sequency transform of y over the r-th variate* x^r*, by* $y^{\sim}(\omega^r)$*. The sequence* $\{\omega_p^r\}_{p=1}^{P_r}$ *is defined as the set that contains the locations of the local maxima of* $|y^{\sim}(\omega^r)|$*, ordered such that* $\omega_p^r \leq \omega_{p+1}^r$*, for all* $p_r = 1, \ldots, P_r$*. Furthermore, define the sets*

$$\begin{array}{rrcl} & W_1(\kappa) & := & \{\omega_1^r\}_{r=1}^d, \\ \textit{and} & W_j(\kappa) & := & M_j^{\uparrow}(\kappa) \cup W_{j-1}(\kappa) \setminus M_j(\kappa), \\ \textit{with} & M_j(\kappa) & := & \{\omega_{s_r}^r \in W_{j-1}(\kappa) \colon \omega_{s_r}^r - \min W_{j-1}(\kappa) < \kappa\}, \end{array}$$

and where the set operator $\cdot^{\uparrow}$ *is defined as* $M_j^{\uparrow} \colon M_j = \{\omega_{s_r}^r\} \mapsto \{\omega_{s_r+1}^r\}$.

Example 4.3 *Consider the set* $W_1(0) := \{\omega_1^r\}_{r=1}^3$, *with* $\omega_1^1 < \omega_1^2 < \omega_2^1 < \omega_1^3$. *It then follows that* $M_2(0) = \{\omega_1^1\}$, $M_2^{\uparrow}(0) = \{\omega_2^1\}$, *and* $W_2(0) = \{\omega_2^1, \omega_1^2, \omega_1^3\}$. *Likewise, we have* $W_3(0) = \{\omega_2^1, \omega_2^2, \omega_1^3\}$, *and* $W_4(0) = \{\omega_3^1, \omega_2^2, \omega_1^3\}$.

The set $\{W_j(\kappa)\}_j$ is a subset of points that lie in the set of all sequency maxima. Larger values of κ result in sparser search spaces. Figure 2 depicts a simple 2-dimensional example for two different values of κ. It can be seen that the search space traces a path between the maximum closest to the origin to the one furthest away. It is constructed such that a search over this subspace is not unduly influenced by the sequency spectrum of any one particular dimension relative to the other $(d-1)$ dimensions. Equivalently, it assumes that the spectral bandwidth of the noise, or information, does not change too much from one dimension to another. Herewith lies a useful compromise between accuracy and sparsity. The result is a family of search spaces parameterised by κ, which should be chosen in accordance with the computational resources available.

[Figure 2 about here.]

Even when the number of dimensions is greater than one, many researchers follow the orthodox strategy of searching for a universal, or scalar, parameter that is constant with respect to dimension. In fact, we can use our framework to develop this search method and consider a situation where some subset of the variables suffer an undue level of noise such that it is difficult, or impossible, to make reliable estimates of the individual parameters. In this case, a somewhat more rudimentary approach is to use all of the variables to bound a single universal scalar parameter estimate $\omega_*^r = \omega_*$, for all $r = 1, \ldots, d$, by

$$\arg \inf_r \sup_{\omega^r} |y^{\sim}(\omega^r)| \leq \omega_* \leq \arg \sup_r \sup_{\omega^r} |y^{\sim}(\omega^r)| \,. \tag{10}$$

A grid search can then be employed inside this interval.

If the search strategy of Definition 4.1 is deemed too computationally costly, then Definition 4.2 offers a trade-off between SVM optimisation times and coarser searches. Furthermore, if this is also deemed to take an unacceptable amount of computational time then Inequality (10) can be used to search for a parameter that is uniform over all dimensions.

4.3 Summary of Method

We can now summarise our method. Given the training data $x_n \in X \subseteq \mathbb{R}^d$ and training labels, $y_n \in \{\pm 1\}$, we proceed as follows:

- Use nearest-neighbour constant interpolation to derive a regularly-sampled labelling function $y(x)$.
- Use Equation (9) to perform a sequency transform in each dimension. (The infinite sum of fast Fourier transforms needs to be truncated at the user's discretion.)
- Find the absolute maxima of the sequency transform in each dimension.
- Use search methods from Section 4.2.2 to train and test the SVM performance for the hyper-parameter candidates using the sinc kernel.
- Choose the best performing hyper-parameter.

5 Application to Hyperspectral Imagery

In this section, we illustrate the efficacy of the approach on a well-studied problem, namely classification of regions of vegetation in a remotely-sensed hyperspectral image.

5.1 Data and Approach

The airborne visual and infrared imaging system (AVIRIS) remotely senses hyperspectral image data comprising intensity information over 224 co-terminous electromagnetic spectral bands, ranging from 0.4 to 2.5 μm. AVIRIS data facilitate myriad applications including resource management, mineral exploitation, environmental monitoring (Landgrebe, 2002), and detection of military targets (Nothard *et al.*, 2003). The large number of variables and classes make the data set ideal for demonstrating the utility of our sinc kernel approach and search strategy. Furthermore, there exists a free and publicly-available AVIRIS data set that has been used by several research groups to benchmark various hyperspectral image classification techniques. It can be downloaded from `ftp://ftp.ecn.purdue.edu/biehl/MultiSpec/` (last accessed 25 November 2005). The following simulations make use of these data.

In the hyperspectral image context, each pixel is described by a single data point, $x_n \in \mathbb{R}^d$. Each element x_n^r, represents the intensity value of pixel n, in the r-th spectral band. Each pixel belongs to one of 17 different classes of ground vegetation. Previous work on the data set has considered 4-, 16-, and 17-class problems. Tadjudin (1998) gives specific details of the pixel and spectral band subsets used. Figure 3 shows the sequency spectra $|y^{\sim}|$ taken from the 4-class AVIRIS problem. In this case, it can clearly be seen by inspection in the top-left plot that the bands 99–148 and 150–200 have remarkably similar spectra. It follows that their maxima, depicted in Figure 4, all fall on very similar points. Moreover, several other such congruences are apparent. Consequently, the search strategy constructed in Definition 4.2 is appropriate.

[Figure 3 about here.]

[Figure 4 about here.]

5.2 Simulation Results

For a fair comparison to be drawn between our results and others, we follow the same sampling and validation technique used in previous research on the AVIRIS data. That is, 20% of the original data are randomly chosen as training data, and the remaining 80% are held out as the testing data. The sampling of training data was repeated 10 times to allow an estimate of the sampling error to be made. The resulting validation measure is simply the percentage of incorrect classifications on the testing data.

The sinc-based search strategies implemented are the bounded scalar search described by Inequality (10) and the sparse hyper-parameter search space $\{W_j(0.05)\}_{j=1}^5$ from Definition 4.2. Figures 5 and 6 show how the validation accuracy varies with respect to the universal scalar parameter $1/\omega_*$, using the search strategy defined by Inequality (10). Note that the optimal scalar value lies within the estimated parameter bounds predicted by Inequality (10). Although the range of variation of accuracy is small, the reader is reminded that we are classifying many thousands of pixels, so that the number of degrees of freedom is very high. In these circumstances, even apparently quite small differences can be enormously significant, as the error bars on the figures confirm.

[Figure 5 about here.]

[Figure 6 about here.]

Table 1 draws a comparison between the proposed sinc methods and the best results found by previous researchers, as well as some comparative results of our own using different kernels. Gualtieri and Cromp (1998) tested several orders of polynomial SVM kernels over 5 trials for the 4-class problem (but just 1 trial for the 16-class problem) and found that the degree-7 kernel performed the best. The entry in the table for the 4-class problem of 4.1% error is the average over the 5 trials. Du (2004) also used a degree-7 polynomial kernel and obtained an apparently poorer error rate of 4.5%. We do not know whether this was for multiple trials or not; if it was, we do not know if this figure is the average or best. Our results for the average over 10 trials for the 4-class problem using a 7th order polynomial closely match those of Du (2004), yet fall some 0.6 percentage points short of the figure reported by Gualtieri and Cromp (1998) for the same method. It seems unlikely that a difference of this magnitude could be due to sampling error (since the standard error of the mean for our 10 trials was just 0.13 percentage points for the 4-class problem). Concerning the SVM approach in general, we can see that this performs significantly better than the Bayesian method used by Tadjudin (1998) and Landgrebe (2002).

[Table 1 about here.]

All of the sinc kernel results represent the average, taken over 10 trials. The mean standard error was below 0.2 percentage points for the 4-class problem, and below 0.1 percentage points for the 16- and 17-class problems. The sinc methods appear to be comparable to the state-of-the-art in the 4-class problem if Gualtieri and Cromp (1998) is taken as the basis of comparison but superior if our replication of the degree-7 polynomial kernel is taken as the reference. For the 16- and 17-class subsets, the sinc kernel SVM clearly surpasses all previous results. Generally, the search based on Definition 4.2 yields slightly better performance than that based on Inequality (10).

We conjecture that a more comprehensive filter design, or noise estimation, strategy may enhance the performance of our approach. Since the constant-interpolation technique used here is somewhat crude, a more rigorous treatment of the irregular-sampling problem should be considered. The penalty term C from the SVM problem has been fixed such that no training errors are allowed. The effect that $C < \infty$ has on the optimal parameter has not been addressed here. Such examination is beyond the scope of this work and is left for possible future consideration.

6 Conclusion

We have shown that the SVM classification machine learning problem can be tackled in the context of signal theory. The interconnection between Paley-Wiener spaces and the sinc kernel has been exploited to form an explicit relationship between our information model and the sinc kernel hyper-parameter. By employing some recent work on sequency analysis, it has been shown that the nature of the model can be discerned. Driven by this theory, a finite hyper-parameter search space was realised. Moreover, by introducing further assumptions, we have shown that the compromise between computational effort and search space sparseness can be managed sensibly. Finally, the approach achieves the best results so far on the much-studied AVIRIS remote-sensing data set.

References

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, **68**(3), 337–404.

Beer, T. (1981). Walsh transforms. *American Journal of Physics*, **49**(5), 466–472.

Belkin, M. and Niyogi, P. (2004). Semi-supervised learning on Riemannian manifolds. *Machine Learning*, **56**(1–3), 209–239.

Bochner, S. (1959). *Lectures on Fourier Integrals*. Princeton University Press, Princeton, NJ.

Borel, E. (1897). Sur l'interpolation. *Comptes Rendus de l'Acadmie des Sciences*, **124**, 673–676.

Chapelle, O., Vapnik, V., Bousquet, O., and Mukherjee, S. (2002). Choosing multiple parameters for support vector machines. *Machine Learning*, **46**(1–3), 131–159.

Debnath, R. and Takahashi, H. (2004). Analyzing the behaviour of distribution of data in the feature space of SVM with Gaussian kernel. *Neural Information Processing Letters*, **5**(3), 41–48.

Du, P. (2004). *Self Adaptive Support Vector Machines and Automatic Feature Selection*. MSc thesis, McMaster University, Hamilton, Ontario.

Elliot, D. F. and Rao, K. R. (1982). *Fast Transforms: Algorithms, Analyses, Applications*. Academic Press, Orlando, FL.

Girosi, F. (1998). An equivalence between sparse approximation and support vector machines. *Neural Computation*, **10**(6), 1455–1480.

Gröchenig, K. (1993). Irregular sampling of wavelet and short-time Fourier transforms. *Constructive Approximation*, **9**(2-3), 283–297.

Gualtieri, J. and Cromp, R. (1998). Support vector machines for hyperspectral remote sensing classification. In *Proceeding of the 27th AIPR Workshop: Advances in Computer Assisted Recognition*, pages 121–132, Washington DC.

Guo, B., Damper, R., Gunn, S., and Nelson, J. (2005a). Adaptive band selection for hyperspectral image fusion using mutual information. *Proceedings of the Eighth International Conference on Information Fusion, Philadelphia, PA, no pagination (CD-ROM).*

Guo, B., Damper, R., Gunn, S., and Nelson, J. (2005b). Hyperspectral image fusion using spectrally weighted kernels. *Proceedings of the Eighth International Conference on Information Fusion, Philadelphia, PA, no pagination (CD-ROM).*

Hardy, G. H. (1941). Notes on special systems of orthogonal functions, IV: The orthogonal functions of Whittaker's cardinal series. *Proceedings of the Cambridge Philosophical Society*, **37**, 331–348.

Higgins, J. R. (1985). Five short stories about the cardinal series. *Bulletin of the American Mathematical Society*, **12**(1), 45–89.

Higgins, J. R. (1996). *Sampling Theory in Fourier and Signal Analysis: Foundations*. Clarendon Press, Oxford, UK.

Hughes, R. D. and Heron, M. L. (1989). Approximate Fourier transform using square waves. *Proceedings of the IEE*, **136**(4), 223–288.

Kimeldorf, G. and Wahba, G. (1970). A correspondence between Bayesian estimation of stochastic processes and smoothing by splines. *Annals of Mathematical Statistics*, **41**(2), 495–502.

Kon, M. A., Raphael, L. A., and Williams, D. A. (2005). Extending Girosi's approximation estimates for functions in Sobolev spaces via statistical learning theory. *Journal of Analysis and Applications*, **3**(2), 67–90.

Lanckriet, G. R. G., Cristianini, N., Bartlett, P., El Ghaoui, L., and Jordan, M. I. (2004). Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, **5**(1), 27–72.

Landgrebe, D. (2002). Hyperspectral image data analysis as a high dimensional signal processing problem. *IEEE Signal Processing Magazine*, **19**(1), 17–28.

Mercer, J. (1909). Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society of London*, **A-209**, 415–446.

Nelson, J. D. B. (2001). *The Construction of Some Riesz Basis Families and their Application to Coefficient Quantization, Sampling Theory, and Wavelet Analysis*. PhD thesis, Anglia Polytechnic University, Cambridge, UK.

Nelson, J. D. B. (2002). A multi-channel multi-sampling rate theorem. *International Journal of Sampling Theory and its Applications*, **2**(1), 83–96.

Nothard, J. M., Kent, N. M., West, C. E., Wood, J., and Oxford, W. J. (2003). Full system modelling for hyperspectral target detection and identification. *Proceedings of the SPIE Algorithms and Technologies for Multispectral, Hyperspectral and Ultraspectral Imagery IX. (S. S. Shen and P. E. Lewis, eds.)*, **5093**, 37–44.

Oppenheim, A. V. and Schafer, R. W. (1989). *Discrete-Time Signal Processing*. Prentice Hall, Englewood Cliffs, NJ.

Smale, S. and Zhou, D. X. (2004). Shannon sampling and function reconstruction from point values. *Bulletin of the American Mathematical Society*, **41**(3), 279–305.

Smola, A. J., Schölkopf, B., and Müller, K.-R. (1998). The connection between regularization operators and support vector kernels. *Neural Networks*, **11**(4), 637–649.

Sugiyama, M. and Müller, K.-R. (2002). The subspace information criterion for infinite dimensional hypothesis spaces. *Journal of Machine Learning Research*, **3**(5), 323–359.

Tadjudin, S. (1998). *Classification of High Dimensional Data with Limited Training Samples*. PhD thesis, School of Electrical Engineering and Computer Science, Purdue University, West Lafayette, IN.

Wang, W. J., Xu, Z. B., Lu, W. Z., and Zhang, X. Y. (2003). Determination of the spread parameter in the Gaussian kernel for classification and regression. *Neurocomputing*, **55**(1), 643–663.

List of Figures

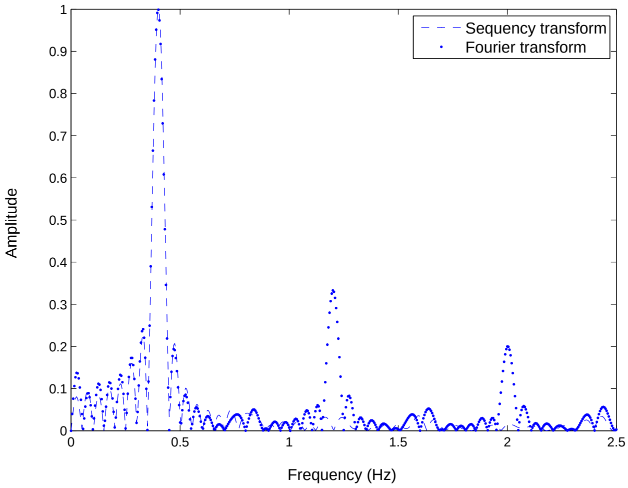
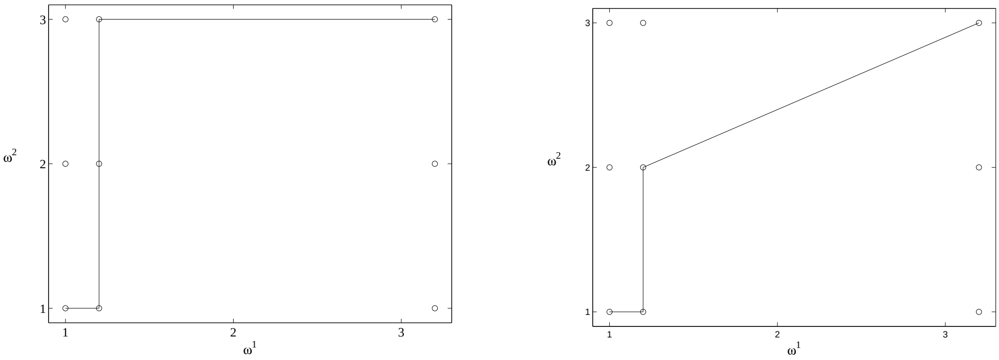


Figure 1: Fourier and sequency spectra for the finite 1-D signal $f(x) = \operatorname{sgn}\left(\sin(0.4 \times 2\pi x)\right)$.



(a) $\kappa = 0$

(b) $\kappa = 0.2$

Figure 2: Search space for two different values of κ. The circles denote the location of the maxima over a 2-dimensional domain. The lines plot the searches (a) $\{W_j(0)\}_{j=1}^{5}$ and (b) $\{W_j(0.2)\}_{j=1}^{4}$.

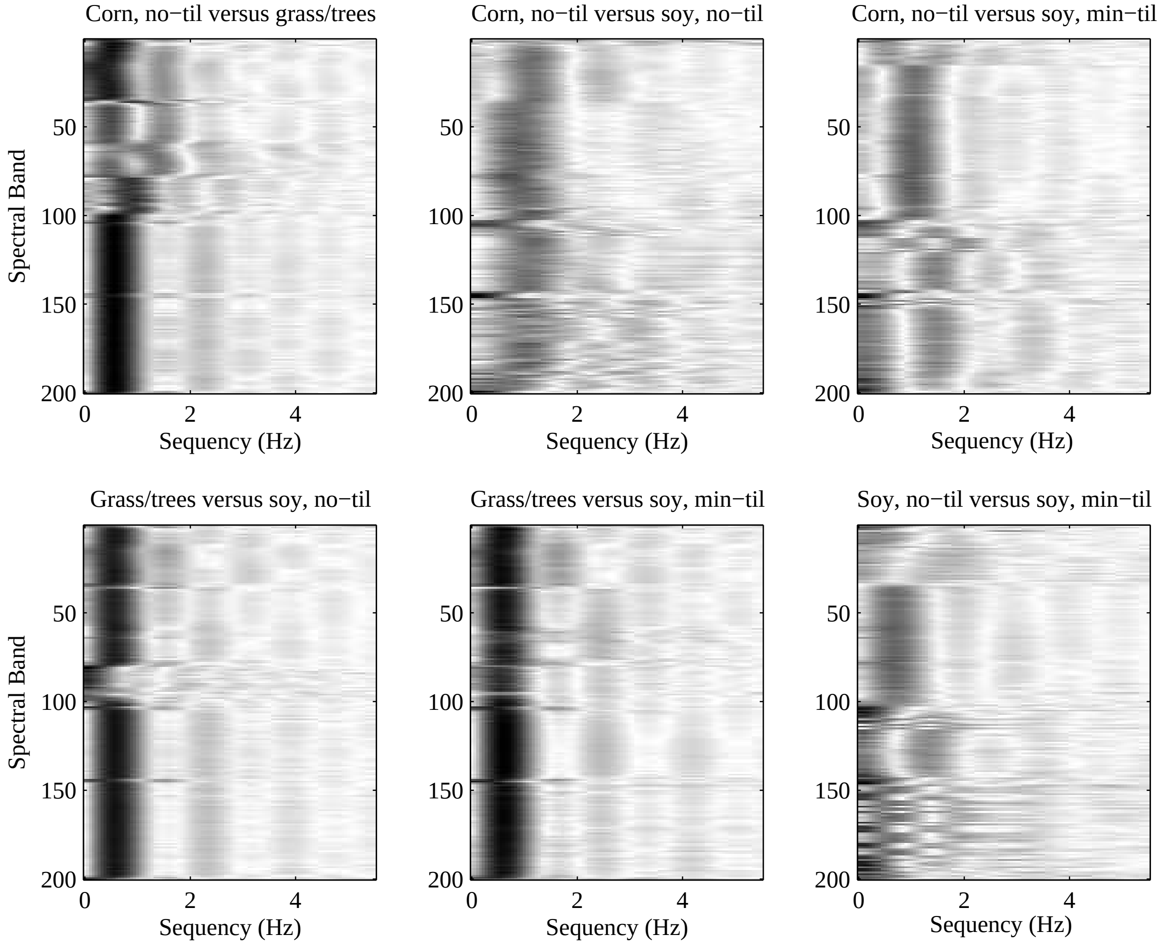


Figure 3: Sequency spectra $|y^{\sim}|$ for the 4-class AVIRIS problem. Darker tones indicate higher magnitude.

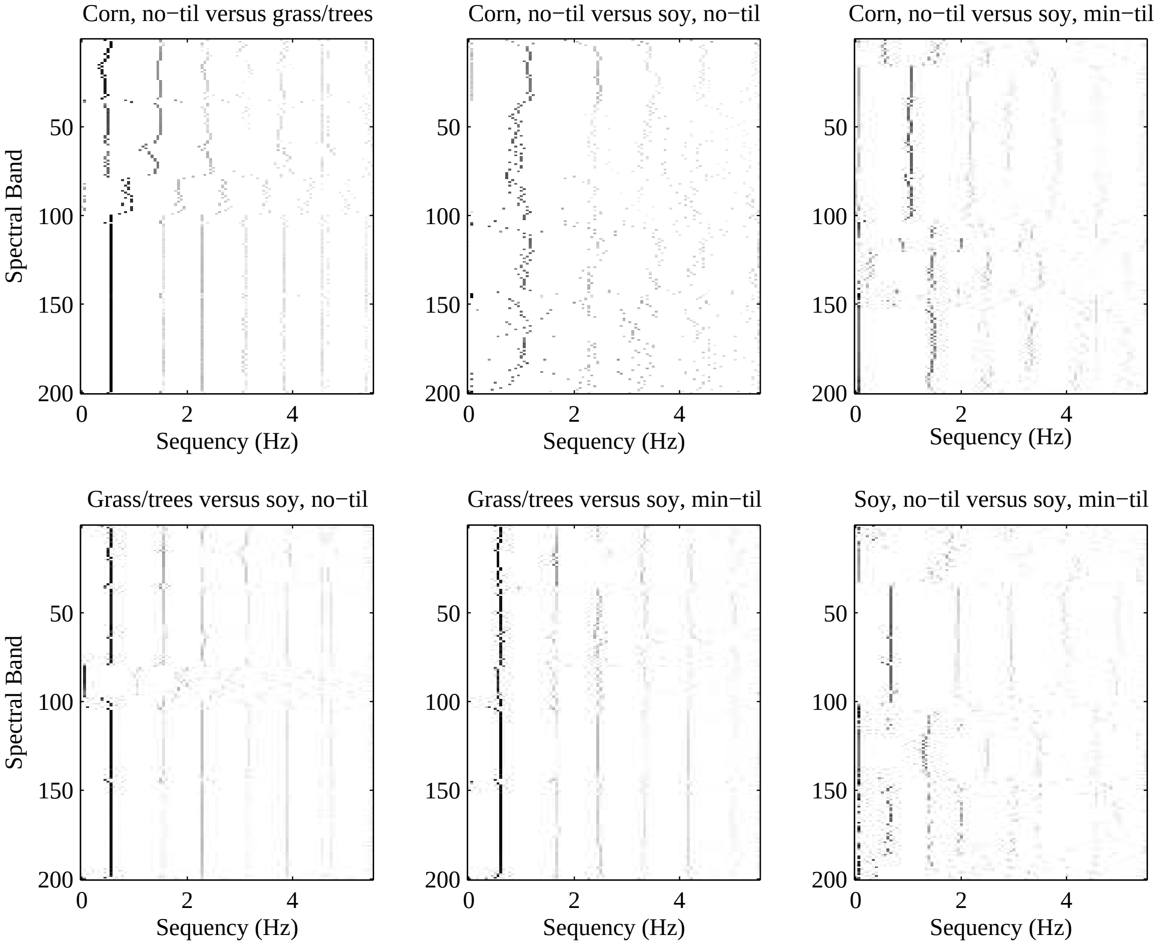


Figure 4: Sequency spectra maxima for the 4-class AVIRIS problem. Darker tones indicate higher magnitude.

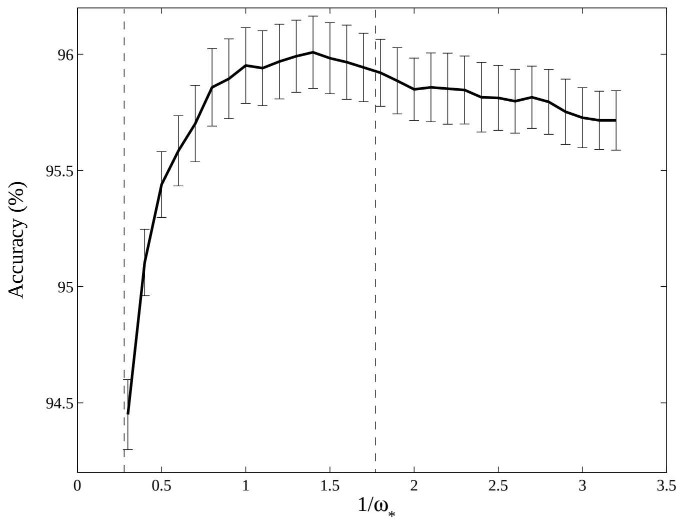


Figure 5: Accuracy with respect to the sinc kernel parameter over the bounded scalar search defined by Inequality (10) for the 4-class problem. The dotted, vertical lines indicate the estimated bounds for the optimal parameter. The error bars denote the standard mean error over 10 trials with different partitions of training and test data.

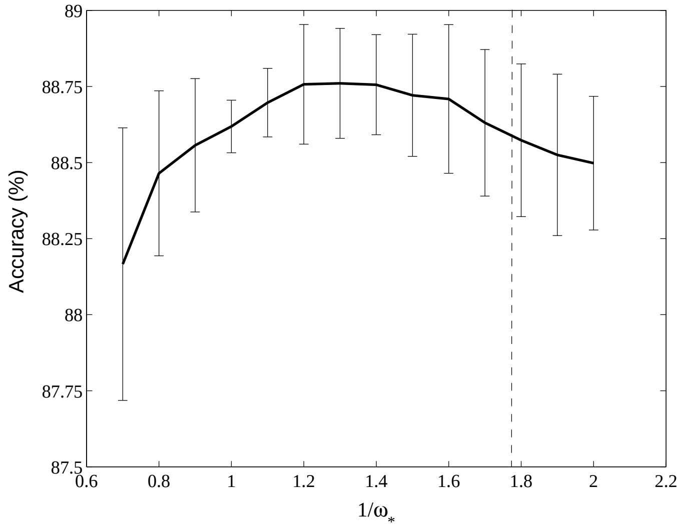


Figure 6: Accuracy with respect to the sinc kernel parameter over the bounded scalar search defined by Inequality (10) for the 16-class problem. The dotted, vertical lines indicate the estimated bounds for the optimal parameter. The error bars denote the standard mean error over 10 trials with different partitions of training and test data.

List of Tables

Table 1: AVIRIS classification: state-of-the-art results compared to results using the sinc kernel developed in this paper (shown in bold).

Source	Penalty	Trials	Method	Error (%)
4-class problem				
Sect. 4.2.2, Definition 4.2	∞	10	**Sinc SVM, sparse search**	**3.9**
Sect. 4.2, Inequality (10)	∞	10	**Sinc SVM, bounded search**	**4.0**
Gualtieri & Cromp (1998)	1000	5	SVM poly. kernel, degree-7	4.1
Du (2004)	1000	?	SVM poly. kernel, degree-7	4.5
This work	1000	10	SVM poly. kernel, degree-7	4.7
This work	∞	10	Gaussian RBF kernel	4.9
Tadjudin (1998); Landgrebe (2002)	1000	10	Bayesian discrim. analysis	6.5
Du (2004)	1000	?	Gaussian RBF kernel	7.9
16-class problem				
Sect. 4.2.2, Definition 4.2	∞	10	**Sinc SVM, sparse search**	**10.9**
Sect. 4.2, Inequality (10)	∞	10	**Sinc SVM, bounded search**	**11.2**
Gualtieri & Cromp (1998)	1000	1	SVM poly. kernel, degree-7	12.7
17-class problem				
Sect. 4.2.2, Definition 4.2	∞	10	**Sinc SVM, sparse search**	**11.3**
Sect. 4.2, Inequality (10)	∞	10	**Sinc SVM, bounded search**	**12.2**
This work	1000	10	SVM poly. kernel, degree-7	15.1
Tadjudin (1998); Landgrebe (2002)	1000	10	Bayesian discrim. analysis	17.1