Signal Theory for SVM Kernel Design with Applications to Parameter Estimation and Sequence Kernels

J. D. B. Nelson, R. I. Damper, S. R. Gunn and B. Guo

Information: Signals, Images, Systems (ISIS) Research Group
School of Electronics and Computer Science
University of Southampton
Southampton SO17 1BJ, UK

Abstract

Fourier-based regularisation is considered for the support vector machine (SVM) classification problem over absolutely integrable loss functions. We show that a principled and finite kernel hyper-parameter search space can be discerned a priori by using the sinc kernel. The method has been tested on two representative problems, deliberately chosen to be very different. First, simulations performed on a publicly-available hyperspectral image dataset reveal that the approach yields results that surpass state-of-the-art benchmarks. The method is then adapted to the recently-proposed max sequence kernel, which has previously been applied to speaker recognition (specifically text-independent verification) using the PolyVar corpus. Here, we apply our methods to text-dependent speaker identification using the BT Millar corpus. We show that this particular speaker-recognition problem gives rise (unlike earlier work) to a max kernel that is not sufficiently close to positive semi-definiteness for the SVM training algorithm to converge. To this end, we make adaptations to the max sequence kernel such that positive semi-definiteness, and so convergence, is guaranteed.

1 Introduction

Parameter choice is an open problem in support vector machine (SVM) learning. 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

[1] This work was supported by the Data and Information Fusion (DIF) Defence Technology Centre funded by the UK Ministry of Defence and managed by General Dynamics Limited and QinetiQ.

way to elicit 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 can be shown that the 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 [2] often turns out to be much less of a problem than expected. Some recent machine learning research has focused on finding cogent explanations for this phenomenon. Belkin and Niyogi [1] 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: i.e., 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. In fact, we propose principled assertions that reduce the infinite search space to a finite one. Ultimately, our philosophy is inspired by the discipline of 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 simple 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 base-band-limited signals. For a given class of data we show how to estimate, a priori, a suitable kernel and parameter subspace. Our method is evaluated on two representative problems, deliberately chosen to be quite different so as to exercise fully the approach. The first is a remote-sensing problem, i.e., classification of the pixels of a hyperspectral image using the popular, extensively-studied AVIRIS dataset. The second problem involves sequence data, i.e., text-dependent speaker identification using the BT Millar database.

The remainder of this paper is structured as follows. In Section 2, the data class and 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. In Section 5, we announce the best results to date on the AVIRIS hyperspectral image dataset. In Section 6, we adapt the approach to the application of a new sequence kernel family to the speaker-recognition problem. Although the results are some way short of the best reported in the literature on the Millar dataset, the work brings to light some important practical and theoretical issues surrounding the use of sequence kernels.

2 Model Construction

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

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

where f, the decision function to be determined, in some Hilbert space $\mathcal{H}(X)$, is regularised by the operator $T : \mathcal{H} \mapsto \mathcal{F}$. The resulting learned decision function, implied by the representor theorem [12], is the solution $f = \sum_{n \in \mathbb{N}} y_n \alpha_n k(x_n, \cdot)$, where k is a Mercer kernel [18]. Herewith, the classifier is defined by $\operatorname{sgn} f$. Our main contention is that before an effort is made to build 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 in this work, is that the labelling function maps d-variate data to labels via $y : \mathbb{R}^d \supset X \mapsto \{\pm 1\}$, with

$$y(x) := \operatorname{sgn}\big(\varphi(x) + \epsilon(x)\big), \tag{1}$$

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

$$\varphi \in PW_{\Omega^*} := \bigoplus_{r=1}^{d} \left\{ \zeta \in L_2(X) : \operatorname{supp} \zeta^\wedge \subseteq \Omega_r^* \right\}, \tag{2}$$

with $\operatorname{supp} \zeta := \{x \in X : \zeta(x) \neq 0\}$, and where $\cdot^\wedge$ denotes d-variate Fourier transformation. The condition $\varphi \in PW_{\Omega^*}$ restricts the behaviour of the information content to functions of finite bandwidth around the origin. Although this kernel is familiar to signal theorists and engineers, it is a seemingly rare tool in machine learning. 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 [23]) Let the Mercer kernel $k : X \times X \mapsto \mathbb{R}$, and the regularisation operator $T : \mathcal{H} \mapsto \mathcal{F}$, be such that $k(x, \xi) \equiv \big\langle (Tk)(x), (Tk)(\xi) \big\rangle_{\mathcal{F}}$. Then the SVM classification problem can be written*

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

Theorem 2.2 *(Translation invariant kernels, Smola, Schölkopf, and Müller [23]) Consider a kernel, endowed with translation invariance, namely $k(x,\xi)=k(x-\xi)$, with the regularisation operator $T:\mathcal{H}\mapsto\mathcal{F}$, defined by*

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

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

Corollary 2.1 *It follows from Theorem 2.2 that the regularisation term from the SVM problem is*

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

with $\omega:=(\omega^r)_{r=1}^d$, and that $\left(k^\wedge(\omega)\right)^{-1}=\left(\prod_{r=1}^d k_r^\wedge(\omega^r)\right)^{-1}$ regularises the decision function f by acting as a filter, in the signal analysis sense, on $\left|f^\wedge\right|^2$.

The unique kernel associated with the reproducing kernel Hilbert space PW_{Ω^*} is the sinc kernel $\prod_{r=1}^d \mathrm{sinc}_{\omega_*^r}(x^r-\xi^r)$. Given the model (1), where the information content is embedded in the Paley-Wiener space (2), it is only sensible to constrain the decision function to the same Paley-Wiener space. From Corollary 2.1, it follows that in the Fourier domain the multiplicative filter, which acts upon $\left|f^\wedge\right|^2$, is

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

with the d-dimensional hypercuboid

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

In this case, since $k^\wedge\geq 0$ holds over $\mathbb{R}^d$, Bochner's theorem 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 representor result ensures that the decision function is restricted to $PW_{\Omega^{*}}$.

Remark 2.1 *We now see that, in the context of our work, the non-regularised, higher dimensional input space discussed by Belkin and Niyogi [1] 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 $T\colon\ PW_{\mathbb{R}^d} \mapsto\ PW_{\Omega^{*}}$.*

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 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.

Let $\mathrm{cal}_{\omega}(t) := \mathrm{sgn}\cos\omega t$, and $\mathrm{sal}_{\omega}(t) := \mathrm{sgn}\sin\omega t$, and define the complex square-wave family as

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

This differs from the definition of the more common Walsh-Hadamard analysis described elsewhere. In particular, the system employed here is defined over a denser, uniform grid rather than over a dyadic 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. 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 3.1 *(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 [19], enables us to express the labelling function in terms of the complex square-wave family.

Proposition 3.1 *(Biorthogonal complex square-wave system, Nelson [19]) The biorthogonal dual of $\{\psi_n\}$ is*

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

We introduce the sequency transformation, $\cdot^{\sim}$, namely

$$f^{\sim}(\omega) = \int_{\mathbb{R}} f(t) \overline{\psi_\omega^*(t)}\, \mathrm{d}t. \tag{3}$$

From Proposition 3.1, it follows that y can be expanded as a superposition of square waves, viz.

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

Hence, the coefficients that express y in terms of the square-wave basis are found by performing the sequency transform of y. Recall from (1) that $\varphi \in PW_{\Omega^*}$, and, without loss of generality, $\epsilon \in PW_{\Omega^+}$. The linearity property of Paley-Wiener spaces gives rise to

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

We define the sequency function space S_Ω as

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

Now since $\operatorname{sgn} \varphi \in S_{\Omega^*} \Rightarrow \varphi \in PW_{\Omega^*}$, and $\operatorname{sgn} \epsilon \in S_{\Omega^+} \Rightarrow \epsilon \in PW_{\Omega^+}$, we can express the labelling function y, as a sequency-limited function,

$$\begin{aligned} & y = \operatorname{sgn}(\varphi + \epsilon) \in S_{\Omega^* \cup \Omega^+} \\ \text{with} \quad & y = \int_{\Omega^* \cup \Omega^+} y^\sim(\omega) \psi_\omega(\cdot) \, \mathrm{d}\omega, \end{aligned} \tag{4}$$

where $y^\sim$ can be computed via

$$\begin{aligned} y^\sim(\omega^r) &= \frac{1}{\sqrt{2\pi}} \sum_{m \in 4\mathbb{Z}+1} \frac{\mu(|m|)}{m} \int_{\mathbb{R}} y(x^r) e^{-i\omega x^r/m} \, \mathrm{d}x^r \\ &= \sum_{m \in 4\mathbb{Z}+1} \frac{\mu(|m|)}{m} y^\wedge\left(\frac{\omega^r}{m}\right), \end{aligned} \tag{5}$$

and 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 must be considered. Since a comprehensive treatment of irregular sampling issues is beyond the scope of this paper, we employ a simple strategy whereby the data are mapped to a uniform grid via nearest neighbour, constant interpolation.

By definition, the information content of $\varphi + \epsilon$ lies in the frequency base-band $\Omega^* = (-\omega_*\pi, \omega_*\pi)$. The informative part of the labelling function $\operatorname{sgn}(\varphi + \epsilon)$ lies analogously in some sequency base-band $\Omega^* = (-\omega_*\pi, \omega_*\pi)$.

Example 3.1 *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* $\operatorname{sinc}(\omega_* \cdot)$ *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 computationally-expensive parameter estimators such as cross validation.

4 Parameter Estimation

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

There exist several different ways to measure SVM performance. To expedite the empirical comparisons drawn in Section 5, we shall consider perhaps the most straightforward measure, namely the validation error. Here, the data are split into two distinct sets. One 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 initial point for the search, and bad convergence. Unfortunately, the problems inherent in 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. [13] use semi-definite programming techniques to compute the kernel matrix. Debnath and Takahashi [6] 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. Guo et al. use measures of mutual information to guide parameter scaling [9].

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 that is inspired by the engineering discipline of filter design. Although this 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 (3), via the series of fast Fourier transforms (5), so as to discern the interval Ω^* from equation (4). 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 sinc kernel under the assumption that $\Omega^* \cap \Omega^+ = \{\}$. In practice, since each datum has finite length, the sequency transform (3) is taken over a finite domain T. From equation (5) 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 $*$ here 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 3.1 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. 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.

Of course, when the number of dimensions or maxima preclude an exhaustive search over the entire set, one may be compelled to compromise accuracy and either bound the search space, conduct a sparser search, or both. To this end, we effect a disciplined compromise between search sparsity and accuracy by the following construct.

Definition 4.1 *Define the sequency transform of y over the rth 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*

$$W_1(\kappa) := \{\omega_1^r\}_{r=1}^d,$$

and

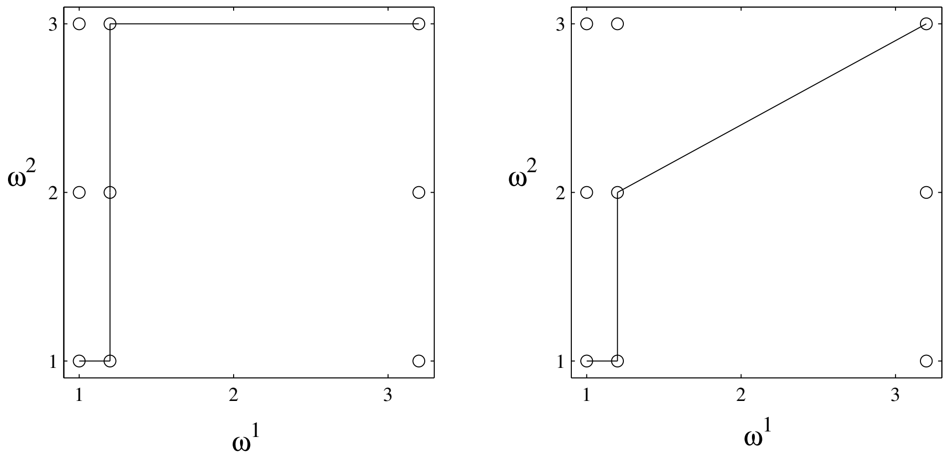

Fig. 1. The circles denote the location of the maxima over a 2-dimensional domain. The lines plot the searches $\{W_j(0)\}_{j=1}^5$ on the left and $\{W_j(0.2)\}_{j=1}^4$ on the right.

$$W_j(\kappa) := M_j^{\uparrow}(\kappa) \cup W_{j-1}(\kappa) \setminus M_j(\kappa),$$

with

$$M_j(\kappa) := \{\omega_{s_r}^r \in W_{j-1}(\kappa) \colon \omega_{s_r}^r - \min W_{j-1}(\kappa) < \kappa\},$$

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 4.1 *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 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. 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. Larger values of κ produce sparser search sets. Figure 1 depicts a simple example for two different values of κ. 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.

We next consider the application of our method to two representative and quite different problems. The first is a remote-sensing problem, i.e., classification of the pixels of a hyperspectral image using the popular, extensively-studied AVIRIS dataset. The second problem involves sequence data, i.e., text-dependent speaker identification using the BT Millar database.

5 Application to Hyperspectral Imagery

The airborne visual and infrared imaging system (AVIRIS) hyperspectral image data comprises intensity information over 224 co-terminous electromagnetic spectral bands, ranging from 0.4 to 2.5 μm. AVIRIS data facilitate myriad applications including land resource management, mineral exploitation, and environmental monitoring. The large number of variables, and classes, makes the dataset ideal for demonstrating the utility of our sinc kernel approach and search strategy. Furthermore, there exists the free, publicly-available AVIRIS dataset [14] that has been widely used by several research groups to benchmark various hyperspectral image classification techniques. The AVIRIS dataset consists of a single 'datacube', i.e., it does not comprise sequential data.

In the hyperspectral images, 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 rth spectral band. Each pixel belongs to one of seventeen different classes of ground vegetation. Previous work on the dataset has considered four-, sixteen- and seventeen-class problems. For a fair comparison to be drawn between our results and others, we follow the same sampling and validation technique as used in previous research on the AVIRIS data. That is, 20% of the original data is randomly chosen as training data, and the remaining 80% is held out as the test data. The resulting validation measure is simply the percentage of incorrect classifications on the test data. Figure 2 shows the sequency spectra $y^{\sim}$ taken from the four-class AVIRIS problem.

Table 1 compares results using the proposed sinc methods and the best results found by previous researchers using the same sampling and validation regime. Gualtieri and Cromp [8] tested several orders of polynomial SVM kernels over 5 trials and found that the degree-7 kernel performed the best [2]. We can see that the SVM approach holds a significant advantage over the Bayesian method used by Tadjudin [24] and Landgrebe [15].

The sinc-based search strategy implemented here is the sparse hyper-parameter search space $\{W_j(0.05)\}_{j=1}^{5}$ from Definition 4.1. All of the sinc kernel results represent the average, taken over 10 trials. The mean standard error was below 0.2% for the four-class problem, and below 0.1% for the sixteen- and seventeen-class

[2] We have been unable to replicate Gualtieri and Cromp's result of 4.1% error rate; our result is 4.7%, in line with Du [7].

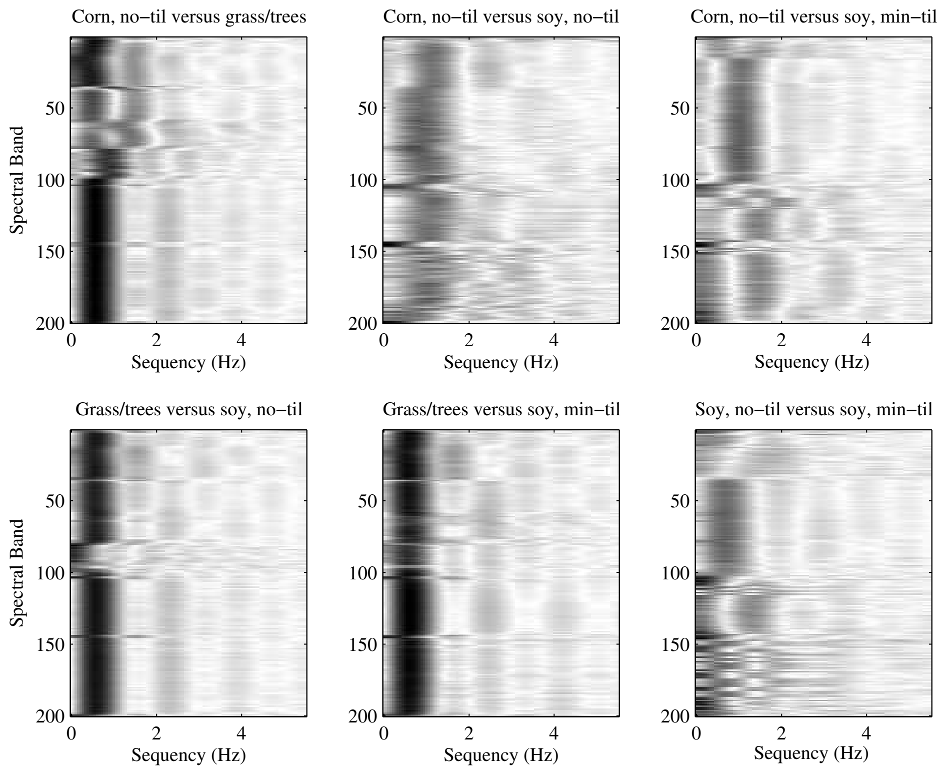

Fig. 2. Sequency spectra $|y^{\sim}|$ for the four-class AVIRIS problem. Darker tones indicate higher magnitude.

problems. The sinc methods appear to be superior to the state-of-the-art in the four-class problem. For the sixteen- and seventeen-class subsets, the sinc methods comfortably surpass all previous published results [3].

6 Application to Speaker Recognition

Using a support vector machine (SVM) sequence kernel approach, Campbell et al. [3] recently obtained text-independent speaker recognition results outperforming the traditional Gaussian mixture model method [21]. However, their design precludes the kernel trick and thus limits kernel choice. With this issue in mind, Mariéthoz and Bengio [16] not only proposed a similar design that admitted the kernel trick but also proposed their max kernel method. In their study of text-independent speaker verification, they found that the max kernel incurred the least error on the popular PolyVar [4] telephone corpus.

[3] Although Guo et al. [10] report an apparently lower error rate of $< 10\%$ on the 16-class AVIRIS problem using a polynomial kernel, this was for a different training and testing regime in which the data were split 50:50.

Table 1
AVIRIS classification: State-of-the-art

Source	Penalty	Method	Error (%)
FOUR-CLASS PROBLEM			
Section 4, Definition 4.1	∞	**Sinc SVM, sparse search**	**3.9**
Gualtieri and Cromp [8] (5 trials)	1000	SVM poly. kernel, degree 7	4.1
Du [7]	1000	SVM poly. kernel, degree 7	4.5
This work	1000	SVM poly. kernel, degree 7	4.7
This work	∞	Gaussian RBF kernel	4.9
Tadjudin [24], Landgrebe [15]	1000	Bayesian discrim. analysis	6.5
Du	1000	Gaussian RBF kernel	7.9
SIXTEEN-CLASS PROBLEM			
Section 4, Definition 4.1	∞	**Sinc SVM, sparse search**	**10.9**
Gualtieri and Cromp (1 trial)	1000	SVM poly. kernel, degree 7	12.7
SEVENTEEN-CLASS PROBLEM			
Section 4, Definition 4.1	∞	**Sinc SVM, sparse search**	**11.3**
This work	1000	SVM poly. kernel, degree 7	15.1
Tadjudin and Landgrebe	1000	Bayesian discrim. analysis	17.1

Inspired by this approach to text-independent speaker verification, we have explored its application to text-dependent speaker identification[4]. Our reasons for so doing were to avoid mere duplication of Mariéthoz and Bengio's work, prior familiarity with this form of the speaker-recognition problem, and immediate access to a particular corpus specifically designed for digit recognition, namely the BT Millar corpus [20,17]. However, our simulations realised in practice a problem that Mariéthoz and Bengio had only noted in theory, namely that the resulting kernel is not guaranteed to be positive semi-definite (PSD). Moreover, the positive semi-indefiniteness of all the standard kernel choices that we attempted to use in the max kernel paradigm meant that the SVM training algorithm failed to converge to a solution. The specific reason that we encountered this problem whereas Mariéthoz and Bengio did not is unknown. However, the two problems (text-independent verification vs. text-independent identification) are quite different and involve very

[4] In *verification*, the task is to verify or deny that a speaker is who he/she claims to be; in *identification*, the task is to identify from the speech signal one out of a number of possible speakers.

different training and testing regimes on different datasets. Thus, potential reasons are not hard to find.

When presented with non-PSD kernels, the standard approach is simply to add a constant term to the diagonal of the kernel matrix [22]. We propose a generalisation of this technique, and hence facilitate the implementation of the sinc kernel, along with some of the other standard kernels.

6.1 Sequence Kernels

In the context of support vector machine classification, the speaker recognition problem gives rise to the following sequence kernel formulation. The nth utterance X_n of some corpus comprises a sequence of T_n many frames $\{x_{n,t}\}_{t=1}^{T_n}$, where each frame $x_{n,t}$ contains d many cepstral coefficients. Hence, a kernel K must be designed such that $K(X_n, X_m)\colon \mathbb{R}^{d\times T_n} \times \mathbb{R}^{d\times T_m} \mapsto \mathbb{R}$. Such a kernel is known as a sequence kernel because it must act on an ordered set of vectors.

Perhaps the simplest design is the mean kernel, which is realised by constructing a kernel $k\colon \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$ for each frame of speech and taking the mean value over all possible combinations:

$$K(X_n, X_m) = \frac{1}{T_n T_m} \sum_{t=1}^{T_n} \sum_{s=1}^{T_m} k(x_{n,t}, x_{m,s}) \ . \tag{6}$$

An important feature of this kernel is that it is guaranteed to be positive semi-definite. To see this we rewrite it as

$$K(X_n, X_m) = \left\langle \frac{1}{T_n} \sum_{t=1}^{T_n} \phi(x_{n,t}), \frac{1}{T_m} \sum_{t=1}^{T_m} \phi(x_{m,t}) \right\rangle,$$

with $k(x, z) = \langle \phi(x), \phi(z) \rangle$, and note that any matrix that can be written as a Gram matrix of linearly independent vectors is positive definite.

6.2 The Max Kernel

Mariéthoz and Bengio [16] note a clear theoretical drawback of this approach. It does not necessarily make sense to compare all the frames of one utterance with all the frames of another utterance. In particular, when viewed as a similarity measure, one would expect the kernel to give a maximum result for two identical utterances.

The mean kernel does not guarantee this. Mariéthoz and Bengio offer the following simple counter-example.

Example 6.1 *Given a multi-frame utterance X_n, consider the one-frame utterance $X_m = \{x_{n,t_*}\}$ with $x_{n,t_*} = \operatorname*{argmax}_t k(x_{n,t}, x_{n,s})$. It follows that:*

$$K(X_n, X_m) \geq K(X_n, X_n) .$$

Motivated by this counter-example, Mariéthoz and Bengio construct their max kernel:

$$K(X_n, X_m) := \frac{1}{T_n} \sum_{t=1}^{T_n} \max_s k(x_{n,t}, x_{m,s}) + \frac{1}{T_m} \sum_{s=1}^{T_m} \max_t k(x_{n,t}, x_{m,s}). \tag{7}$$

The max kernel ensures that only the closest matching frames are included in the computation of the kernel. Although this kernel is no longer guaranteed to satisfy Mercer's conditions, Mariéthoz and Bengio nonetheless found it to be positive semi-definite in practice when applied to text-independent speaker verification using the PolyVar corpus. By contrast, when we applied this method to text-dependent speaker recognition using the BT Millar corpus, we found that the max kernel always resulted in a positive semi-indefinite training matrix. Moreover, the SVM quadratic optimiser failed to converge for any choice of the kernel k. We do not know the specific reason for this, other than that our work has many differences to that of Mariéthoz and Bengio, any of which could potentially be the cause.

6.3 From Indefiniteness to Definiteness

Although, to the authors' knowledge, non-PSD kernels rarely arise in speaker-recognition problems, they do occur in the context of protein classification problems. By far the most common approach to deal with positive semi-indefinite kernels is simply to add a constant term to the diagonal of the kernel matrix so as to obtain a PSD kernel [22]. Since a matrix is PSD if and only if all eigenvalues are non-negative, it suffices to perform the kernel modification sometimes referred to as the diagonal-shift kernel:

$$K_* = K + \lambda I,$$

with $\lambda := \min(\lambda_n, 0)$ and where λ_n is the smallest eigenvalue of K. However, the resulting diagonal-shift kernel may well be far away, in some sense, from the original kernel. This is certainly the case if the smallest eigenvalue of K has a large magnitude. A less common method is to find the nearest PSD kernel, namely the so-called positive approximant.

Definition 6.1 *Let $\mathcal{S}_+$ denote the space of all positive semi-definite matrices. Define the the positive approximant of a matrix $K \in \mathbb{R}^{d \times d}$ by*

$$K_+ := \operatorname*{argmin}_{S \in \mathcal{S}_+} \|K - S\| .$$

It turns out that by recovering the following result from Higham [11], we can find the positive approximant, in the Frobenius sense, uniquely and analytically.

Theorem 6.1 *(Higham) Let $K = K^T \in \mathbb{R}^{d \times d}$, have the polar decomposition $K = UH$, with $U^T U = I$, and $H = H^T \in \mathcal{S}_+$. Then*

$$K_+ = \frac{K + H}{2}$$

is the unique positive approximant of K with respect to the Frobenius norm $\|\cdot\|_F$.

Use of the positive approximant is sometimes referred to as 'de-noising', since it is equivalent to replacing the negative eigenvalues of the original kernel matrix with zeros. In our simulations, we have found that the positive approximant K_+ performs less effectively than the diagonal-shift kernel K_*. However, we propose the kernel

$$K_\beta = \beta K_* + (1 - \beta) K_+, \qquad 0 \le \beta \le 1, \tag{8}$$

and have found that it can perform better than K_* for $\beta \neq 1$. Moreover, thanks to the isometry $(a_{n,m}) \mapsto [a_{nm}]$, between the space of all N-by-M matrices and NM sized vectors, induced by the Frobenius norm $\|\cdot\|_F$, our kernel K_β also has a geometric interpretation as illustrated in Figure 3. Our kernel family defines a straight line between the diagonal-shift kernel and the nearest PSD kernel with respect to the Frobenius metric. Since PSD matrices are closed under addition, and multiplication by scalars, K_β defines a line that exists entirely within the space of all PSD matrices. The parameter β determines how close the modified kernel is to the original kernel. It acts as a trade-off between closeness to the positive approximant and the diagonal-shift kernel.

6.4 Simulations

The text-dependent British Telecom Millar corpus [20,17] comprises high quality microphone recordings, downsampled from 20 kHz sampling rate to 8 kHz, with

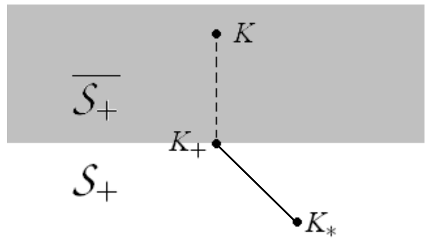

Fig. 3. Geometrical representation of the proposed modified kernel. The black solid line illustrates the kernel family $\beta K_* + (1-\beta)K_+$. The shaded area distinguishes the positive semi-indefinite space $\overline{\mathcal{S}_+}$ from the positive semi-definite space $\mathcal{S}_+$.

16-bit resolution. Speech data were collected from 46 male and 14 female English speakers over five sessions and a period of three months. The speakers were all required to utter the digits "one" to "nine", "zero", "nought", and "oh" five times per session. We follow a similar procedure to that outlined by Damper and Higgins [5] and use the first 10 utterances of the words "seven" and "nine" for each speaker as the training data and the remaining 15 sessions of the words "seven" and "nine" as the test data.

The max method (7) was tested for classification accuracy over the polynomial $\langle x, z\rangle^{\gamma}$, exponential $\exp\left(\|x-z\|\,\gamma^{-1}\right)$, and sinc kernel $\prod \mathrm{sinc}_{\gamma}(x-z)$. These max kernels were tested over $\beta = 0, 0.1, \ldots, 1$. The best results are tabulated in Table 2 alongside the results obtained with $\beta = 1$. We see that the polynomial and sinc kernels benefit from our generalisation of the diagonal-shift sequence kernel. Results for the mean method (6) are also given in Table 3 for completeness. They confirm that, overall, the best accuracy is realised with the sinc kernel and max method [5]. For the mean method, best results are obtained for the exponential kernel, but the sinc kernel is only a little worse.

7 Conclusions

We have argued that the SVM classification machine learning problem can profitably be tackled in the context of signal theory. The interrelation 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, the nature of the model can be discerned. Consequently, a finite hyper-parameter search space was realised.

[5] We make no claim here that our results are competitive with the best in the literature on the BT Millar database. In fact, Damper and Higgins [5] obtained 100% correct identification for "seven" and "nine" with added noise.

Table 2
Speaker recognition results for the max kernel on the BT Millar corpus. Best results (shown in bold) for the two words tested were obtained with the sinc kernel.

Kernel	γ	β	Error (%)
sinc	1.6	1	3.00
sinc	1.6	0.7	**2.78**
exp	10	1	3.22
polynomial	1	1	18.77
polynomial	1	0.5	16.22
polynomial	2	1	16.89
polynomial	2	0.2	14.11
polynomial	3	1	18.44
polynomial	3	0.1	15.44

(a) “seven”

Kernel	γ	β	Error (%)
sinc	3	1	**5.08**
sinc	3	0.7	5.31
exp	10	1	6.89
exp	10	0.5	6.33
polynomial	1	1	15.60
polynomial	1	0.5	13.56
polynomial	2	1	13.22
polynomial	2	0.1	11.75
polynomial	3	1	11.86
polynomial	3	0.1	11.17

(b) “nine”

Table 3
Speaker recognition results for the mean kernel on the BT Millar corpus. Best results (shown in bold) for the two words tested were obtained with the exponential kernel, but are poorer than those for the sinc kernel.

Kernel	γ	Error (%)
exp	10	**4.00**
polynomial	1	23.00
polynomial	2	10.00
polynomial	3	9.94
sinc	1.6	4.63

(a) “seven”

Kernel	γ	Error (%)
exp	10	**5.99**
polynomial	1	17.29
polynomial	2	9.97
polynomial	3	9.15
sinc	1.6	6.33

(b) “nine”

Moreover, by introducing further assumptions, we have shown that the compromise between computational effort and search space sparseness can be managed sensibly.

The approach has been applied to two very different problems: hyper-spectral image classification using the AVIRIS dataset and text-dependent speaker identification using the BT Millar database. The former problem involves a single (static) datacube whereas the latter requires appropriate handling of sequential (dynamic) speech data. Applied to the much-studied AVIRIS dataset, we achieve the best results so far published.

The approach can also be adapted for our newly-constructed sequence kernel family. The main conclusion from our work on speaker identification is that the max kernel yields superior performance to the mean kernel, so vindicating Mariéthoz and Bengio's method. One of the important features of their method is that one can "plug in" any kernel. However, we found that this method did not converge without modification because the kernel matrix was not guaranteed to be positive semi-definite. A diagonal shift was, therefore, employed to promote convergence. It was found that a linear combination of the positive approximant K_+ and the diagonal-shift kernel K_*, as in equation (8), generally performed slightly better than K_* alone at the expense of having to search for a good value of the weighting coefficient β.

References

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

[2] R. Bellman. *Adaptive Control Processes: A Guided Tour*. Princeton University Press, Princeton, NJ, 1961.

[3] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, and P. A. Torres-Carrasquillo. Support vector machines for speaker and language recognition. *Computer Speech and Language*, 20:210–229, 2006.

[4] G. Chollet, J.-L. Cochard, A. Constantinescu, C. Jaboulet, and P. Langlais. Swiss French PolyPhone and PolyVar: Telephone speech databases to model inter- and intra-speaker variability. Technical Report IDIAP-RR 01, IDIAP, Martigny, Switzerland, 1996. Available at `ftp://www.idiap.ch/pub/reports/1996/rr96-01.ps.gz`.

[5] R. I. Damper and J. E. Higgins. Improving speaker identification in noise by subband processing and decision fusion. *Pattern Recognition Letters*, 24(13):2167–2173, 2003.

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

[7] P. Du. Self adaptive support vector machines and automatic feature selection. MSc thesis, McMaster University, Hamilton, Ontario, Canada, 2004.

[8] J. Gualtieri and R. Cromp. Support vector machines for hyperspectral remote sensing classification. In *Proceedings of the 27th AIPR Workshop on Advances in Computer Assisted Recognition*, pages 121–132, Washington DC, 1998.

[9] B. Guo, R. Damper, S. Gunn, and J. Nelson. Hyperspectral image fusion using spectrally weighted kernels. In *Proceedings of the 8th International Conference on Information Fusion*, Philadelphia, PA, 2005. paper B2-1, no pagination, Proceedings on CD-ROM.

[10] B. Guo, S. R. Gunn, R. I. Damper, and J. D. B. Nelson. Band selection for hyperspectral image classification using mutual information. *IEEE Geoscience and Remote Sensing Letters*, 3(4):522–526, 2006.

[11] N. J. Higham. Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra and Applications*, 103:103–118, 1988.

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

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

[14] D. Landgrebe. AVIRIS data. `ftp.ecn.purdue.edu`. last accessed 25/11/05.

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

[16] J. Mariéthoz and S. Bengio. A max kernel for text-independent speaker verification systems. In *Second International Workshop on Multimodal User Authentication*, 2006. no pagination, available at `http://mmua.cs.ucsb.edu/`.

[17] H. Melin. Databases for speaker recognition: Activities in COST250 Working Group 2. In *Proceedings of COST Workshop on Speaker Recognition in Telephony*, pages 8–15, Rome, Italy, 1999.

[18] J. Mercer. Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of Royal Society of London*, 209(A456):415–446, 1909.

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

[20] M. Pawlewski and S. Downey. Channel effects in speaker recognition. In *Proceedings of COST Workshop on Applications of Speaker Recognition Techniques in Telephony*, pages 39–46, Vigo, Spain, 1996.

[21] D. A. Reynolds and R. C. Rose. Robust text-independent speaker identification using Gaussian mixture models. *IEEE Transactions on Speech and Audio Processing*, 3(1):72–83, 1995.

[22] H. Saigo, J. P. Vert, N. Ueda, and T. Akutsu. Protein homology detection using string alignment kernels. *Bioinformatics*, 20(11):1682–1689, 2004.

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

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