MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY
and
CENTER FOR BIOLOGICAL AND COMPUTATIONAL LEARNING
DEPARTMENT OF BRAIN AND COGNITIVE SCIENCES

A.I. Memo No. 1632 May 1998
C.B.C.L Paper No. 161

Notes on PCA, Regularization, Sparsity and Support Vector Machines

Tomaso Poggio and Federico Girosi

This publication can be retrieved by anonymous ftp to publications.ai.mit.edu.
The pathname for this publication is: ai-publications/1500-1999/AIM-1632.ps

Abstract

We derive a new representation for a function as a linear combination of local correlation kernels at optimal sparse locations and discuss its relation to PCA, regularization, sparsity principles and Support Vector Machines. We also discuss its Bayesian interpretation and justification.
We first review previous results for the approximation of a function from discrete data (Girosi, 1998) in the context of Vapnik's *feature space* and *dual representation* (Vapnik, 1995). We apply them to show 1) that a standard regularization functional with a stabilizer defined in terms of the correlation function induces a regression function in the span of the feature space of classical Principal Components and 2) that there exist a *dual* representations of the regression function in terms of a regularization network with a kernel equal to a generalized correlation function. We then describe the main observation of the paper: the dual representation in terms of the correlation function can be *sparsified* using the Support Vector Machines (Vapnik, 1982) technique and this operation is *equivalent* to sparsify a large dictionary of basis functions adapted to the task, using a variation of Basis Pursuit De-Noising (Chen, Donoho and Saunders, 1995; see also related work by Donahue and Geiger, 1994; Olshausen and Field, 1995; Lewicki and Sejnowski, 1998).

In all cases – regularization, SVM and BPD – we show that a bayesian approach justifies the choice of the the correlation function as kernel. In addition to extending the close relations between regularization, Support Vector Machines and sparsity, our work also illuminates and formalizes the LFA concept of Penev and Atick (1996). We discuss the relation between our results, which are about regression, and the different problem of pattern classification.

This report describes research done within the Center for Biological and Computational Learning in the Department of Brain and Cognitive Sciences and at the Artificial Intelligence Laboratory at the Massachusetts Institute of Technology. This research is sponsored by a grant from the National Science Foundation under contract ASC-9217041 (this award includes funds from ARPA provided under the HPCC program) and by MURI grant N00014-95-1-0600. Additional support is provided by Eastman Kodak Company, Daimler-Benz, Siemens Corporate Research, Inc. and AT&T.

1 Introduction

In supervised learning problems we are given a discrete data set $D_l \equiv \{(\mathbf{x}_i, z_i) \mapsto X \times Z\}_{i=1}^N$, obtained by sampling N times the set $X \times Z$ according to $P(\mathbf{x}, z)$. The goal of learning is to provide a deterministic function $f(\mathbf{x})$ which models the relationship between X and Z and thereby solves the associated *regression* problem.

A specific example that we will use throughout this paper is the regression problem of reconstructing a specific image f given its pixel values at discrete locations in the image plane. This paper focuses on a special version of this problem, in which prior information is available in terms of the correlation function of images of the same type as f.

We first reformulate known results to show that the classical Principal Component representation is associated with a regularization formulation of the problem, in which the stabilizer is defined in terms of the correlation function of an ensemble of functions f_α of the same type as the f of the regression problem. Principal Components thus correspond to a special case of the feature space of Vapnik (1995). Regularization provides another *dual representation* – in Vapnik's language – for the regression function in terms of a weighted sum of correlation kernels each centered at a data point $\mathbf{x}_i$. This dual representation contains a large number of terms if the number of data points is large (for instance all pixels in an image). Girosi's results show that it can be sparsified using the SVM formulation (Vapnik, 1995) and that this is equivalent to enforcing a sparsity constraint like in Chen, Donoho and Saunders (1995). Regularization, SVM and a special form of BPD have a Bayesian interpretation: we show that this equivalence (see Wahba, 1990) can be used to justify the use of the correlation function as the kernel.

We will also discuss how our regression results are related to corresponding classification tasks and how the kernels obtained for regression may be used for a pattern recognition problem in a SVM classifier, thus providing sparse features for a classification task.

We first give our reformulation of existing results and then describe our main observations. We assume that the reader is familiar with regularization, SVM techniques and sparsification algorithms (see Girosi, 1998).

2 Background

2.1 Reproducing Kernels and Regularization

Let us first summarize the basic results we will need from the theory of regularization. They are a special case of the technique discussed by Girosi (1998) and can also be found in Wahba (1990). Regularization techniques as developed by us to solve supervised learning problems (Poggio and Girosi, 1989; Girosi, Jones, Poggio, 1995) were limited to shift invariant stabilizers (tensor product and additive stabilizers are special exceptions, see Girosi et al. 1995): the underlying kernel $G(\mathbf{x}, \mathbf{y})$ was constrained to be $G(\mathbf{x}, \mathbf{y}) = G(\mathbf{x} - \mathbf{y})$, strongly limiting – in the language of Vapnik (1995)– the type of associated *feature* representations (the eigenfunctions of the associated integral operator are always Fourier basis functions). It is however possible to construct kernels of the general form $G(\mathbf{x}, \mathbf{y})$ (see Wahba, 1990; Girosi, 1998).

Consider a positive definite function $K(\mathbf{x}, \mathbf{y})$. It is well known that K defines an integral operator with a complete system of orthogonal eigenfunctions that can be made orthonormal and ordered with decreasing eigenvalue[1] with positive λ_n

$$\int_{R^d} d\mathbf{y}\ K(\mathbf{x},\mathbf{y})\phi_n(\mathbf{y}) = \lambda_n\phi_n(\mathbf{x}) \tag{1}$$

and the following series representation that converges absolutely and uniformly:

$$K(\mathbf{x},\mathbf{y}) = \sum_{n=1}^{\infty} \lambda_n\phi_n(\mathbf{x})\phi_n(\mathbf{y}) \tag{2}$$

We now define a scalar product in the function space spanned by the system of ϕ_n and thus induced by K, as follows:

$$< f, g >_K = \sum \frac{1}{\lambda_n} < f, \phi_n >< g, \phi_n > \tag{3}$$

With this definition K defines a Reproducing Kernel Hilbert Space (RKHS) with a corresponding regularization functional:

$$H[f] = \frac{1}{N}\sum_{i=1}^{N}(z_i - f(\mathbf{x}_i))^2 + \gamma\|f\|_K^2 \tag{4}$$

where $\|f\|_K^2 = \sum_n \frac{<f,\phi_n>^2}{\lambda_n}$ is the norm in K induced by the scalar product defined earlier. Minimization of 4 yields the usual solution in terms of regularization networks

$$f(\mathbf{x}) \approx \sum_{i=1}^{N} a_i K(\mathbf{x}, \mathbf{x}_i), \tag{5}$$

solving the regression problem of estimating f from the discrete data $(\mathbf{x}_i, z_i)$. As we mentioned earlier, the specific example we have in mind is f an image and $\mathbf{x}$ a vector in the plane.

2.2 Vapnik's Feature Space and Regularization

The previous section implies that any positive definite kernel K induces a RKHS defined by the *feature vector* $\boldsymbol{\phi}(\mathbf{x}) = (\sqrt{\lambda_1}\phi_1(\mathbf{x}), \sqrt{\lambda_2}\phi_2(\mathbf{x}), \ldots, \sqrt{\lambda_n}\phi_n(\mathbf{x}), \ldots)$, with

$$\boldsymbol{\phi}(\mathbf{x}) \cdot \boldsymbol{\phi}(\mathbf{y}) = K(\mathbf{x}, \mathbf{y})$$

As a consequence, a function in the RKHS space spanned by the orthonormal features can be represented[2] as

$$f(\mathbf{x}) = \sum_{n=1}^{\infty} b_n\phi_n(\mathbf{x}) \tag{6}$$

and also approximated in terms of the dual representation (because of the underlying regularization principle of the previous section)

$$f(\mathbf{x}) = \sum_{i=1}^{N} a_i K(\mathbf{x}, \mathbf{x}_i). \tag{7}$$

[1] We use λ_n instead of the usual (for integral operators) $\frac{1}{\lambda_n}$

[2] In the discrete case $\mathbf{f} = \boldsymbol{\Phi}\mathbf{b}$

Instead of starting from a given K and derive the feature space we could start from any set of orthonormal functions ϕ_n – our features – with appropriate λ_n and construct a regularization kernel $K(\mathbf{x}, \mathbf{y})$ as

$$K(\mathbf{x}, \mathbf{y}) = \boldsymbol{\phi}(\mathbf{x}) \cdot \boldsymbol{\phi}(\mathbf{y}) \tag{8}$$

Remarks:

1. When the ϕ_n are a finite set, the λ_n can be arbitrary (finite) numbers, since there are no convergence problems. In particular they can all be equal to one. Of course, the choice of the ϕ_n defines the space of functions that can be represented accurately in terms of the features.

2. *All* translation-invariant stabilizers ($K(\mathbf{x}, \mathbf{x}_i) = K(\mathbf{x} - \mathbf{x}_i)$) correspond to Fourier eigenfunctions and only differ in the spectrum of the eigenvalues (for a Gaussian stabilizer the spectrum is Gaussian $\lambda_n = Ae^{(-n^2/2)}$ (for $\sigma = 1$)).

3. In standard regularization with translation invariant stabilizers and associated kernels, the common experience, often reported in the literature, is that the form of the kernel does not matter much. We conjecture that this may be because all translation invariant K induce the same type of ϕ_n features - the Fourier basis functions. Correlation functions which are not translation invariant can define instead quite different sets of features which are likely to have quite different effects.

2.3 PCA and Regularization

Until now we have considered the regression problem of estimating f from discrete data. In our example of image reconstruction f would map location $\mathbf{x}$ on the image plane to a real value – the image value at that location. A limit case of the regression problem is classification in which the range of f is $\{0, 1\}$. In our image example, classification corresponds to estimating the binary value of a pixel at a desired location from (binary) values at sparse locations in the (binary) image.

From now on, we will consider a special case of the regression-classification problem: we will assume that in addition to the training data – which are values of the underlying function f at discrete locations $\mathbf{x}_i$ – we also have information about the class of function to which f belongs. In particular, we will assume that the underlying correlation function is known. More formally, the given f is taken to belong to a set of functions f_α over which a probability distribution $P(\alpha)$ is defined. In our standard example of f being a specific image, the f_α are images of the same type, all aligned and registered, for instance images of faces. Then the correlation function of the random signal f – of which the f_α are realizations – is

$$R(\mathbf{x}, \mathbf{y}) = E[(f_\alpha(\mathbf{x}) f_\alpha(\mathbf{y})] \tag{9}$$

where $E[\cdot]$ denotes expectation with respect to $P(\alpha)$. In the following we will always assume that the average function is the null function: $E[f_\alpha(\mathbf{x})] = 0$.

The correlation function R is positive definite and thus induces a RKHS with the λ_n defined by the eigenvalue problem satisfied by R (Hilbert-Schmidt-Mercer theorems). It follows that R provides a "natural" kernel – among the many possible – for solving the regression-classification problem from discrete data for f (see section 4). It also provides the standard Principal Components representation for f in terms of a (in practice finite) set of M ϕ_n, $n = 1, \cdot, M$. The following points hold true:

1. There exists a regularization formulation corresponding to the PCA choice

$$H[f] = \sum_{i=1}^{N} (z_i - f(\mathbf{x}_i))^2 + \gamma \|f\|_R^2 \tag{10}$$

   where $\|f\|_R^2 = \sum_{n=1}^{M} \frac{<f,\phi_n>^2}{\lambda_n}$

2. The regression solution f is in the span of the ϕ_n and can be represented in terms of M Principal Components (with M finite or infinite) as

$$f(\mathbf{x}) = \sum_{n=1}^{M} b_n \phi_n(\mathbf{x}) \tag{11}$$

3. f can be represented in terms of a regularization network as

$$f(\mathbf{x}) = \sum_{i=1}^{N} a_i R(\mathbf{x}, \mathbf{x}_i) \tag{12}$$

   Notice that often only an estimate of R is available and that usually this estimate may be highly rank deficient (see appendix C). In these cases instead of R, one can use a regularization kernel $R^M(\mathbf{x}, \mathbf{y})$ defined as the natural approximation of R in the space of the available M Principal Components:

$$R^M(\mathbf{x}, \mathbf{y}) = \sum_{n=1}^{M} \lambda_n \phi_n(\mathbf{x}) \phi_n(\mathbf{y}) \tag{13}$$

In the following we will drop the superscript M in our notation.
The two representations are equivalent (under the same error criteria) when the number of principal components is chosen equal to be M. Notice that, unlike the global ϕ_n, the basis functions $R(\mathbf{x}, \mathbf{x}_i)$ are usually quite local: consider for instance the translation invariant case of natural images, where the ϕ_n are Fourier components, while the correlation is relatively short range.
Notice that in equation 13 one can assume that the only available *prior* knowledge is which M eigenfunctions are relevant. In the case of finite M we can then define several different regularization kernels all corresponding to the same PCA decomposition. The most natural kernel is simply the projection operator

$$P(\mathbf{x}, \mathbf{y}) = \sum_{n=1}^{M} \phi_n(\mathbf{x}) . \phi_n(\mathbf{y}) \tag{14}$$

P plays the role of the δ function in the space of the ϕ_n. It has an associated regularization formulation (with a stabilizer $\|f\|_P^2 = \sum_{n=1}^{M} < f, \phi_n >^2$). Thus f can be also represented as

$$f(\mathbf{x}) = \sum_{i=1}^{N} \tilde{a}_i P(\mathbf{x}, \mathbf{x}_i) \tag{15}$$

Following the spirit of a suggestion by Penev and Atick (1996), we can define *generalized correlation kernels* parametrized by d (for M finite) as

$$R_d(\mathbf{x}, \mathbf{y}) = \sum_{n=1}^{M} (\lambda_n)^d \phi_n(\mathbf{x}) \phi_n(\mathbf{y}) \tag{16}$$

of which $P = R_0$ and $R = R_1$ are special cases. It is not completely trivial to notice (following Penev and Atick, 1996) that d controls the locality of the kernel. In the shift-invariant case, for which the ϕ_n are Fourier basis functions, d acts as a filter: low-pass for increasing d and high-pass for decreasing d. Thus locality increases for decreasing d: for instance when R_d is a Radial Basis Gaussian function, d controls directly the effective σ of the Gaussian. The most interesting values of d range between 0 and 1: R_0, which is less smooth than R_1, plays the role of the δ function in the space spanned by the ϕ_n while R_d with negative d are similar to "derivatives" of the delta function (consider the example of band-limited functions for which the analog of the delta is the sinc function). For positive integer $d > 1$, R_d is a so called *iterated kernel*: for instance $R_2(\mathbf{x}, \mathbf{y}) = \int\ d\mathbf{z} R(\mathbf{x}, \mathbf{z}) R(\mathbf{z}, \mathbf{y})$, which indicates that positive d corresponds to integral operators (while negative d correspond to differential operators).

Remarks:

1. Given a set of ϕ_n the spectrum λ_n of the correlation function R depends on the specific $P(\alpha)$ since

$$\lambda_n = E[b_n^2] \tag{17}$$

   where the b_n are the coefficients of the expansion of the function f in the set of eigenfunctions ϕ_n.

2. The stabilizer in the form $\|f\|_{R_d}^2 = \sum_{n=1}^{M} \frac{<f,\phi_n>^2}{\lambda_n{}^d}$ has obvious smoothness properties (smoothness increases with d), since the eigenfunctions (ordered as usual) typically have increasing high-frequency content as n increases (theorems in the theory of integral equations, like in Courant and Hilbert, 1953, relate the number of nodes or zeros of the eigenfunctions to their index n).

3. Since regularization has a Bayesian interpretation (Kimeldorf and Wahba, 1971; Girosi, Jones and Poggio, 1995) we have now a probabilistic interpretation of PCA in terms of a prior probability on the space of functions f_α given by $P(f) = e^{(-\|f\|_R^2)}$ and a Gaussian model of the noise (see section **??** and Wahba, 1990).

4. The kernels P and R – and in general R_d – correspond to different prior probabilities (they are multivariate Gaussian priors with different covariances). They define, however, the same set of basis functions ϕ_n – Vapnik's features – and are therefore expected to behave in a similar way.

5. There is a relation between equation 12 and kriging (see Wahba, 1990).

6. The *sinc* function is the translation invariant correlation function of the set of one-dimensional band-limited functions with a flat Fourier spectrum up to f_c (and zero for higher frequencies). Perhaps more interestingly it also corresponds to the operator P of the band-limited functions with a given cut-off (and with any correlation function). The *sinc* function is a positive definite reproducing kernel with negative lobes.

7. The PCA representation equation 11 can be used to solve the problem of regression from N discrete data at locations $\mathbf{x}_l$ by computing the b_i from

$$f(\mathbf{x}_l) = \sum_{n=1}^{M} b_n \phi_n(\mathbf{x}_l) \quad l = 1, \cdot, N \tag{18}$$

In general, the equations can be solved if N is at least equal to M. Note that equation 12 can be used even when $N << M$ and equation 18 cannot be solved. The regularization representation equation 12 can be obtained solving for the a_n from the data (using $\mathbf{a} = (R + \gamma I)^{-1}\mathbf{y}$ with positive or zero γ).

8. The connection between the a_n of equation 12 and b_n of equation 11 is given by

$$b_k = \lambda_k \sum_{j}^{N} a_j \phi_k(\mathbf{x}_j) \tag{19}$$

9. As we will see later, estimates of the correlation function may often be possible from a sufficient set of examples, even in cases in which the dimensionality of the discretized space is very high.

10. Penev and Atick (1996) remarks that the object P corresponds to *local* features, similar to local receptive fields with compact support.

11. Wahba (1990) discusses the relation between regularization, RKHS and correlation functions of Gaussian processes. In particular, f in the RKHS defined by R and f a sample function from a zero-mean Gaussian stochastic process are not the same (when R has more than a finite number of non-zero eigenvalues).

3 Sparsification of the regularization representation and Support Vector Machines

Let us consider again the main scenario we sketched above: a space of functions is characterized probabilistically through its correlation function $R(\mathbf{x}, \mathbf{y})$. Any function f in the space can be represented in terms of the eigenvectors associated with R. An (approximate) representation of f in terms of a finite numbers of Principal Components is a natural compact approximation of f. It is natural to ask whether we could sparsify the N-terms dual representation of f in terms of a regularization network, that is the weighted sum of the kernels R centered at N data points. A natural way to sparsify

$$f(\mathbf{x}) = \sum_{i=1}^{N} a_i R(\mathbf{x}, \mathbf{x}_i) \tag{20}$$

is to use SVM regression (Vapnik, Golowich and Smola, 1997; and Vapnik, 1995) with the kernel R (or R^M, see later). As shown by Girosi (1998), this corresponds to minimizing –instead of the regularization functional 4 - the following functional

$$H[f] = \frac{1}{N}\sum_{i=1}^{N} \mid z_i - f(\mathbf{x}_i) \mid_\epsilon + \gamma \|f\|_R^2 \tag{21}$$

where the following *robust* error function has been defined instead of the usual L_2 norm on the data term:

$$\mid x \mid_\epsilon = \begin{cases} 0 & \text{if } \mid x \mid < \epsilon \\ \mid x \mid - \epsilon & \text{otherwise.} \end{cases} \tag{22}$$

The function that minimizes the functional in eq. (21) depends on a finite number of parameters, and has in our case the following form:

$$f(\mathbf{x}) = \sum_{i=1}^{N'} a_i R(\mathbf{x}, \mathbf{x}_i), \tag{23}$$

where the coefficients a_i are now found by solving a quadratic programming problem. Notice that the sum in equation (23) runs only up to N', where $N' \leq N$. The reason is that, due to the nature of this QP problem, only a "small" number of coefficients will be different from zero, and the data points associated to them are called *support vectors* (in many cases $N' << N$). Thus we can *sparsify the regularization representation of a function* by using the correlation function as the regularization kernel in equation 21.

We now invoke a result in Girosi (1998, section 5) to claim that *the result of minimizing equation 21 is the same as of sparsifying the overcomplete dictionary of $R(\mathbf{x}, \mathbf{x}_i)$ using Basis Pursuit De-noising* (Chen, Donoho and Saunders, 1995) (see also the sparsification approaches of Olshausen and Field, 1996, and Lewicki and Sejnowski, 1998). The proof consists of applying the Girosi version of the Chen-Donoho cost functional to sparsify equation 20, leading to the minimization of the following functional with respect to the coefficients a_i

$$E[\mathbf{a}] = \frac{1}{2}\|f(\mathbf{x}) - \sum_{i=1}^{N} a_i R(\mathbf{x}; \mathbf{x}_i)\|_R^2 + \epsilon \|\mathbf{a}\|_{L_1} \tag{24}$$

The solution of equation 24 is the same as the solution of minimizing 21, which is given by equation 23. Thus a solution equivalent to the SVM solution – in which only a subset of the data points has non-zero coefficients, the so-called support vectors – can be obtained simply by enforcing a sparsity constraint in an approximation scheme of the standard regularization form with R being the correlation matrix

$$f(\mathbf{x}) = \sum_{i=1}^{N} a_i R(\mathbf{x}, \mathbf{x}_i)$$

a sparse representation is sought among a "large" number of possibly local and task-dependent features $R(\mathbf{x}, \mathbf{x}_i)$.

Notice that the framework of sparsification (and the equivalent SVM) allows us to consider a dictionary of overcomplete basis functions and in particular of R not only at multiple locations but also at multiple scales. A natural way to define such a dictionary is to consider, instead of R, R_d for several different values of d and, of course, at many locations (for instance at each pixel in an image). In this case (see appendix) we minimize the sparsification functional to select appropriate sparse scales and locations, yielding *a sparse, multi-scale* representation

$$f(\mathbf{x}) = \sum_{i,d}^{N',D'} a_{i,d} R_d(\mathbf{x}, \mathbf{x}_i), \tag{25}$$

Remarks:

1. Basis Pursuit Denoising provides only a suboptimally sparse representation from a dictionary (because it uses $\|\mathbf{a}\|_{L_1}$ instead of $\|\mathbf{a}\|_{L_0}$ in equation 24) but it probably has good generalization (because in the form of equation 24 it is equivalent to SVM).

2. The form of the solution - a superposition of kernels - does not depend on the form of the norm involved in the data term, as observed earlier by Girosi, Caprile and Poggio (1990). In particular, it is the same for the standard L_2 norm and for the robust norm defined by Vapnik.

3. Our approach of sparsifying the representations of f in terms of the generalized correlation kernel R_d is a principled way to achieve the sparsification proposed by Penev and Atick (1996).

4. Though the representations of a function f in terms of R_d are all equivalent, independent of d, in the standard regularization case, we expect that they will have in general different properties after sparsification.

4 Bayesian interpretation and why R is the kernel of choice

Consider

$$\min_{f \in \mathcal{H}} H[f] = \sum_{i=1}^{N} (y_i - f(\mathbf{x}_i))^2 + \gamma \|f\|_K^2$$

In the standard bayesian interpretation of RN (see for instance (see Girosi et al., 1995) the data term is a model of the noise and the stabilizer is a prior on the regression function f. Informally the equation follows from a MAP estimate of

$$P(f/\mathbf{y}) \propto P(\mathbf{y}/f)P(f)$$

To see the argument in more detail, let us assume that the data y_i are affected by additive independent gaussian noise processes, i.e. $y_i = f(x_i) + \epsilon_i$ with $E[\epsilon_i \epsilon_j] = 2\delta_{i,j}$

$$P(\mathbf{y}/f) \propto \exp(-\sum_i (y_i - f(x_i))^2)$$

and

$$P(f) \propto \exp(-\|f\|_R^2) = \exp\left(-\sum_{n=1}^{M} \frac{c_n^2}{\lambda_n}\right)$$

where $M < \infty$

$$f(\mathbf{x}) = \sum_{n=1}^{M} c_n \phi_n(\mathbf{x}).$$

Thus the stabilizer measures the Malahanobis distance of f from the mean of f_α. To see this, let us represent f in any complete orthonormal basis ψ_i as the vector $\mathbf{f}_i =< f, \psi_i >$. We assume that the data are zero-mean in the sense that $E[f_\alpha(\mathbf{x})] = 0$ (obviously the data can always be processed to satisfy this condition). Then we know that if $P(f)$ is Gaussian then

$$P(f) \propto \exp(-\mathbf{f}^T(\Sigma)^{-1}\mathbf{f})$$

and $\mathbf{f}^T(\Sigma)^{-1}\mathbf{f}$ is the Malahanobis distance of $\mathbf{f}$ from its mean (the origin). $P(f)$ is therefore a multivariate Gaussian with zero mean in the Hilbert space of functions defined by R and spanned by the ϕ_n, that is the space related to Principal Components.
Remarks:

1. Notice that for SVM the prior is the same Gaussian prior but the model of the noise is different and is NOT gaussian additive as in RN (see Pontil et al., 1998). The same is true for BPD, given the equivalence between SVM and BPD.

2. Thus also for SVM (regression) and BPD the prior $P(f)$ gives a probability measure to f in terms of the Malahanobis distance in the Hilbert space defined by R and identical to the space of the Principal Components.

3. There is a natural probabilistic interpretation of the data term (see Girosi et al., 1995). As we have mentioned, in the case of standard regularization, the data term norm (a L_2 norm) corresponds to a Gaussian model of the noise, that is the conditional probability of the data z_i given the function is a Gaussian. Other norms can be interpreted as shown by Girosi, Caprile and Poggio (1990) in probabilistic terms as different models of the noise. Pontil et al. (1998) have derived the noise model corresponding to Vapnik's ϵ insensitive norm.

4.1 Why R is the kernel of choice.

Assume that the problem is to estimate f from sparse data y_i at location $\mathbf{x}_i$. From the previous description it is clear that choosing a kernel K is equivalent to assuming a Gaussian prior on f with covariance equal to K. Thus an empirical estimate of the correlation function associated with a function f should be used, whenever available. Notice that in the Bayesian interpretation a Gaussian prior is *assumed* in regularization as well as in SVM (and in the equivalent BPD formulation). Thus when empirical data are available on the statistics of the family of functions f_α one should check that $P(f)$ is Gaussian and make it zero-mean. Then an empirical estimate of the correlation function $E[(f_\alpha(\mathbf{x})f_\alpha(\mathbf{y})]$ can be used as the kernel.
The relation between positive definite kernels and correlation functions R of Gaussian random processes is characterized in details in Wahba (1990), Theorem 5.2.

5 Conclusions

We know from Wahba (1990) and Girosi (1998) that, given a positive definite K,

1. a regularization functional can be defined
2. a function in the RKHS can be represented in terms of the non-linear *features* provided by the orthonormal eigenfunctions of K and *also* in terms of a linear combination of the kernels K evaluated at sparse points
3. the data term in the regularization functional can be modified to yield a SVM formulation
4. minimizing the SVM functional is the same as sparsifying the regularization representation, that is the dictionary of $K(\mathbf{x}, \mathbf{x}_i)$.

Here we consider the case in which the kernel K is a very special "object" – the correlation function $R(\mathbf{x}, \mathbf{y})$ and justify this choice in terms of the Bayesian interpretation of regularization, SVM and BPD . We focus on its role in regression (function approximation from sparse data)[3]. We show that

1. a function can be represented either by the Principal Components induced by the associated correlation function or in a dual way by the regularization solution - a weighted sum of correlation kernels evaluated at N data points.
2. the representation in terms of the correlation kernel can be sparsified using the SVM technique or, in a completely equivalent way, by using the basis pursuit denoising technique on the dictionary of $R(\mathbf{x}, \mathbf{x}_i)$. Notice that this representation is not only compact (see Chen, Donoho and Saunders, 1995) but it is also likely to achieve good generalization, (since the SVM cost functional implements Vapnik's theory of risk minimization).

In our case SVM can be therefore regarded as a "sparse" version of a regularization network with a kernel derived directly from the correlation function. The regression problem we consider is a problem of signal reconstruction; is very different from the problem of pattern classification (see Appendix). Following the spirit of Penev and Atick, the same sparsified kernels computed for regression may be used with SVM classifiers – in the same way in which principal components are often used – effectively representing a choice of sparse feature from an appropriate large dictionary of basis functions (provided by the $R(\mathbf{x}, \mathbf{x}_i)$).
Correlation functions that are shift invariant are not very interesting from the point of view of the representations discussed here: they all correspond to the same set of Fourier features. Of course, the correlation function corresponding to a large set of images of different scenes and objects will be translation and scale invariant (see Penev and Atick, 1996 and references therein). Properly aligned images of objects of the same type (such as for instance faces or people, see Papageorgiou et al., 1998) instead yield correlation functions which are not shift invariant (see Sirovich and Kirby, 1988; see also Turk and Pentland, 1990). The associated ϕ_n features capture information about the category of objects. They are however global. The correlation kernels $R(\mathbf{x}, \mathbf{x}_i)$ (or the corresponding $P(\mathbf{x}, \mathbf{x}_i)$), instead yield local "features", which can be sparsified

[3] Atick and Penev were probably the first to study the correlation function R in the context or regression.

and thereby simultaneously optimized for generalization. Results from experiments in progress are promising.
It is suggestive to speculate that cortex may use machinery to align and normalize visual inputs so that dictionaries of object specific features can be learned without being affected by arbitrary translations and scalings. At earlier stages of the visual system, however, one may expect from our results that translation invariant correlation functions associated with non-aligned images of different types will determine basis functions similar to local Fourier components. It is interesting to speculate that the correlation functions associated with images at different scales may be learned separately, providing receptive fields at multiple resolutions.

Acknowledgments We would like to thank Mike Oren, Amnon Shashua, Alessandro Verri.

A Multiple scales correlation kernels for regularization, sparsification and SVM

A.1 Multiple scales and classical regularization

Let us consider here the multiple scale generalized correlation R_d of equation 16. Let us assume that $f(\mathbf{x}) = f_1(\mathbf{x}) + f_2(\mathbf{x})$, where f_1 and f_2 represent the components of f at two different scales (the generalization to the case of more than two scales is cumbersome but possible). The functional to be minimized is

$$H[f] = \sum_{i=1}^{N}(z_i - f_1(\mathbf{x}_i))^2 + \eta \sum_{i=1}^{N}(z_i - f_1(\mathbf{x}_i) - f_2(\mathbf{x}_i))^2 + \gamma_1 \|f_1\|^2_{R_1} + \gamma_2 \|f_2\|^2_{R_2}$$

where η is a positive, small number. The underlying idea is that f_1 is a coarse approximation to the data at one scale, while f_2 is a refinement at a finer scale (f_2 approximates the residuals of f_1).

A.2 Multiple scales and sparsification and SVM

The sparsity functional of Chen et al. can be used to choose a sparse subset from the dictionary of basis functions $R_d(\mathbf{x}, \mathbf{x}_i)$, with i and d ranging over a "large" set of locations and scales.
One possible way of obtaining equation 25 from the SVM technique is the following. We assume that $f(\mathbf{x}) = f_1(\mathbf{x}) + f_2(\mathbf{x})$, where f_1 and f_2 represent the components of f at two different scales (the generalization to the case of more than two scale is immediate). The functional to be minimized is

$$H[f] = \sum_{i=1}^{N} \mid z_i - f_1(\mathbf{x}_i) \mid_{\epsilon_1} + \eta \sum_{i=1}^{N} \mid z_i - f_1(\mathbf{x}_i) - f_2(\mathbf{x}_i) \mid_{\epsilon_2} + \gamma_1 \|f_1\|^2_{R_1} + \gamma_2 \|f_2\|^2_{R_2}$$

where $\epsilon_1 > \epsilon_2$, and a η is a positive, small number. The final result is

$$f(\mathbf{x}) = \sum_{i,d}^{N',D'} a_{i,d} R_d(\mathbf{x}, \mathbf{x}_i), \tag{26}$$

B Principal Components under Regularization and Sparsification

As we discussed, minimization of the regularization functional 10 defines a regression function f that is in the span of the features space of the Principal Components (that is the eigenfunctions of R). As we mentioned the solution f of the regression problem can be represented in terms of ϕ_n *and equivalently* in terms of $R(\mathbf{x}, \mathbf{x}_i)$. It is interesting to look at the solution when it is expressed in terms of the principal components. We do this in the case in which we have an infinite number of data points, which corresponds to the case in which we actually know the function we want to approximate. Therefore, we plug the PC representation

$$f(\mathbf{x}) = \sum_{n=1}^{M} b_n \phi_n(\mathbf{x}) \tag{27}$$

in the regularization functional:

$$H[f] = \|g(\mathbf{x}) - f(\mathbf{x}\|_{L_2}^2 + \gamma \|f\|_R^2 \tag{28}$$

where we denoted by g the function we want to approximate. The solution of the minimization problem is

$$b_n = \left(1 + \frac{\gamma}{\lambda_n}\right)^{-1} < \phi_n, g(\mathbf{x}) > . \tag{29}$$

Notice that for $\gamma = 0$ we have the usual solution: b_n is simply the result of projecting the target function $g(\mathbf{x})$ on the principal component ϕ_n. For $\gamma > 0$, the effect of regularization is to decrease all b_n by a factor which depends on the corresponding eigenvalue of the correlation matrix.
It is interesting to compare these regularization solutions to the sparsification and SVM solution (which are the same). We consider the minimization with respect to b_n of

$$E[\mathbf{b}] = \frac{1}{2}\|g(\mathbf{x}) - f(\mathbf{x}\|_{\mathcal{H}}^2 + \epsilon \|\mathbf{b}\|_{L_1} \tag{30}$$

In this case the solution is

$$b_n = | < g, \phi_n > -\gamma_n \epsilon |_+ + | < g, \phi_n > +\gamma_n \epsilon |_- \tag{31}$$

where $|x|_+$ ($|x|_-$) is equal to x when x is positive (negative) and equal to 0 otherwise. The dependency of b_n on $< g, \phi_n >$ is plotted in figure (1). In this case, if the principal component n has a projection which is too small it is simply not used. The non-zero coefficients are shrunk by a factor that depends on the sparsification parameter ϵ (and correspondingly on the ϵ insensitive norm of SVM) and on the eigenvalue of the correlation matrix.
Finally, we consider a very different problem: we perform *exact sparsification with respect to the average reconstruction error over the space of "images"* f_α rather than with respect to a single image. We follow Girosi (1998) and minimize an appropriate functional H, that is

$$\min_{\boldsymbol{\xi}} H[\boldsymbol{\xi}] = \frac{1}{2}E[\|f_\alpha - \sum_n < f_\alpha, \phi_n > \phi_n(\mathbf{x})\xi_n\|^2] + \epsilon \sum_n \xi_n \tag{32}$$

where ξ_n are binary random variables with values in $\{0, 1\}$, $E[\cdot]$ denotes the expectation with respect to $P(\alpha)$ and the ϕ_n are the eigenfunctions of R. In this simple case of orthonormal ϕ_n we find that ($\theta(x)$ is 1 if $x > 0$ and 0 otherwise)

$$\xi_n = \theta(\lambda_n - \epsilon)$$

Thus only those Principal Components are chosen that correspond to eigenvalues larger that the sparsity and SVM parameter ϵ.

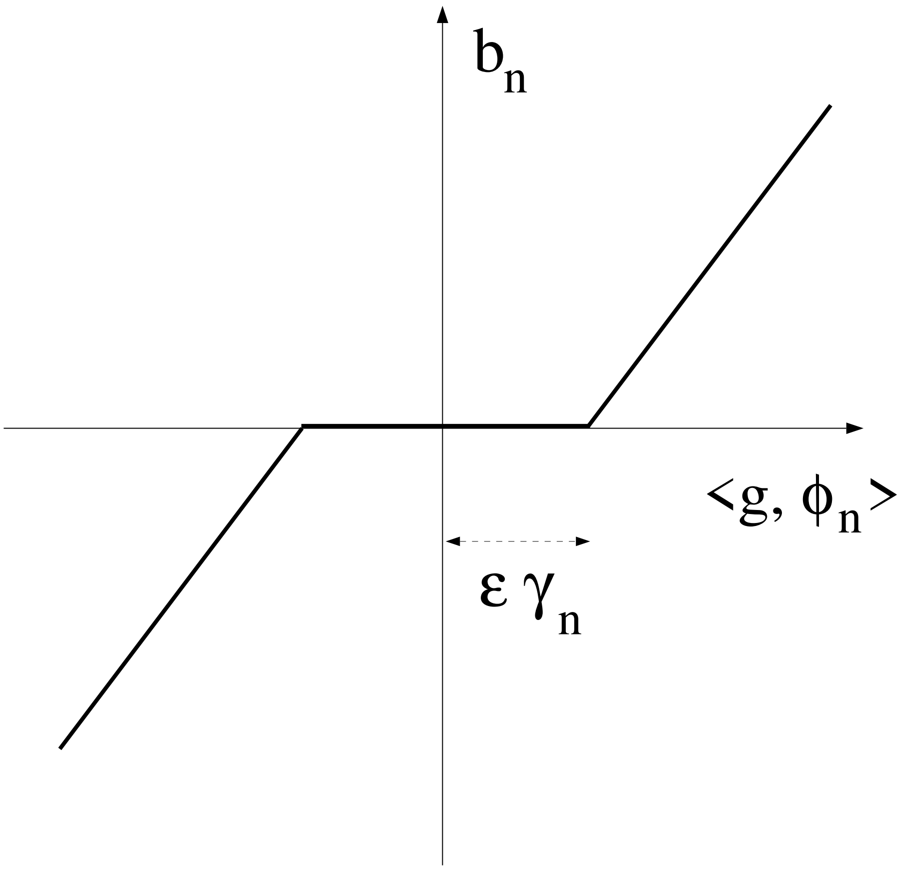

Figure 1: The value of the coefficient b_n as a function of the projection $< g, \phi_n >$ of the data g on the principal component ϕ_n.

C The discrete case

It is often believed that the estimation from data of the correlation function R is impossible or very difficult, because of the 'difficulty of sampling the space f_α. In practice, however, the functions f_α (as well as ϕ_n) must be represented as vectors in a finite dimensional space, albeit of possibly very high dimensionality V. Thus, though the correlation matrix $R = FF^T$ with F a VxQ matrix with columns $\mathbf{f}_\alpha$ may be highly rank deficient, it is possible to obtain useful estimates of it in terms of its M Principal Components, where $M \leq T$ with T being the number of the observations $\mathbf{f}_\alpha$, even when $T << V$. The best technique is to compute the Singular Value Decomposition of F, that is $F = UDV^T$ where the columns of U are the eigenvectors of FF^T, the columns of V are the eigenvectors of F^TF and the diagonal matrix D contains the singular values. Thus an estimate of R can be obtained as the R^M of equation 13.
Assume that values of an unknown vector (say an image) $\mathbf{f}$ are given at a discrete set of points and that an estimate of the underlying correlation matrix is available as R^M. Then $\mathbf{f}$ can be reconstructed either as

$$\mathbf{f}_x = \sum_{n=1}^{M} b_i \phi_{n,x} \tag{33}$$

or as

$$\mathbf{f}_x = \sum_{i=1}^{N'} a_i R^M_{x,x_i} \tag{34}$$

D Pattern Classification

Our standard example in the paper is the problem of image reconstruction from sparse pixel values. This is very different from the problem of classifying images, for instance classifying whether an image is an image of a face or not. To see this consider the spaces involved:

1. *Image reconstruction.* In the case of image reconstruction we would like to approximate the map

   $$f : \mathcal{R}^2 \mapsto \mathcal{R}$$

   from its values at sparse points in $\mathcal{R}^2$. The equivalent problem of binary pixel classification synthesizes a map

   $$f : \mathcal{R}^2 \mapsto \{0, 1\}.$$

   For solving this problem we could use any positive definite function $K(\mathbf{x}, \mathbf{y})$, such as the Gaussian.

2. *Pattern classification.* For pattern classification the problem is quite different: we have several images, which are vectors of N components (pixels) and each image is associated with a binary label. The goal is to learn the map

$$g : \mathcal{R}^N \mapsto \{0, 1\}$$

If I replace $\{0, 1\}$ with $\mathcal{R}$ we have the corresponding (difficult !) regression problem.

The two problems are quite different. They are somewhat related however in the special case of problem (1) that we consider in this paper. In this case we have to solve the regression (or possibly binary regression) problem for pixels of an image $f(\mathbf{x})$ but we also know the generalized correlation function $R(\mathbf{x}, \mathbf{y})$ of the set of similar images $f_\alpha(\mathbf{x})$. As we discussed, R provides a "natural" choice for the regression kernel K. An estimate of R is given by FF^T. In problem (2) the input space consist of vectors $\mathbf{f}(\alpha)$ that may be related to the functions f_α of problem (1) by defining each component indexed by $\mathbf{x}$ of $\mathbf{f}(\alpha)$ as $f_\alpha(\mathbf{x})$. One way to solve problem (2) (classification or even regression) is to use regularization or SVM with a kernel $K(\alpha, \beta)$ equal to the dot product, that is $K(\alpha, \beta) =< f_\alpha, f_\beta >$. The corresponding matrix needed from the data is then $F^T F$. Obviously, the $Q \times Q$ matrix $F^T F$ and the $N \times N$ matrix FF^T are closely related[4]. Notice that in practice it is very difficult if not impossible to estimate empirically the correlation function in a classification problem: that is equivalent to estimate the sufficient (Gaussian) statistics characterizing the classification functions (in our example on the images and not of the images as in the regression case).

References

[1] D. Beymer, A. Shashua, and T. Poggio. Example based image analysis and synthesis. A.I. Memo No. 1431, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1993.

[2] B.E. Boser, I.M. Guyon, and V.N. Vapnik. A training algorithm for optimal margin classifier. In *Proc. 5th ACM Workshop on Computational Learning Theory*, pages 144–152, Pittsburgh, PA, July 1992.

[3] S. Chen, , D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. Technical Report 479, Department of Statistics, Stanford University, May 1995.

[4] R.R. Coifman and M.V. Wickerhauser. Entropy-based algorithms for best-basis selection. *IEEE Transactions on Information Theory*, 38:713–718, 1992.

[5] C. Cortes and V. Vapnik. Support vector networks. *Machine Learning*, 20:1–25, 1995.

[6] R. Courant and D. Hilbert. *Methods of mathematical physics. Vol. 1.* Interscience, London, England, 1953.

[7] M. Donahue and D. Geiger. Sparse representations for image decompositionn. Technical report, Siemens Corporate Research, Princeton, NJ, 1994.

[8] F. Girosi. An equivalence between sparse approximation and Support Vector Machines. *Neural Computation*, 10(6):1455–1480, 1998.

[4]Consider the Singular Value Decomposition of F, that is $F = UDV^T$ where the columns of U are the eigenvectors of FF^T, the columns of V are the eigenvectors of $F^T F$ and the diagonal matrix D contains the singular values.

[9] F. Girosi, T. Poggio, and B. Caprile. Extensions of a theory of networks for approximation and learning: outliers and negative examples. In R. Lippmann, J. Moody, and D. Touretzky, editors, *Advances in Neural information processings systems 3*, San Mateo, CA, 1991. Morgan Kaufmann Publishers.

[10] G.F. Harpur and R.W. Prager. Development of low entropy coding in a recurrent network. *Network*, 7:277–284, 1996.

[11] G.S. Kimeldorf and G. Wahba. A correspondence between Bayesan estimation on stochastic processes and smoothing by splines. *Ann. Math. Statist.*, 41(2):495–502, 1971.

[12] M. Lewicki and T. Sejnowski. Learning nonlinear overcomplete representations for efficient coding. In *Advances in Neural and Information Processing Systems 10.* MIT Press, 1998.

[13] S. Mallat and Z. Zhang. Matching Pursuit in a time-frequency dictionary. *IEEE Transactions on Signal Processing*, 41:3397–3415, 1993.

[14] B.A. Olshausen and D.J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

[15] C. Papageorgiou, M. Oren, and T. Poggio. A general framework for object detection. In *Proceedings of the International Conference on Computer Vision*, Bombay, India, January 1998.

[16] P. S. Penev and J. J. Atick. Local feature analysis: A general statistical theory for object representation. *Neural Systems*, 7:477, 500, 1996.

[17] T. Poggio and F. Girosi. A Sparse Representation for Function Approximation. *Neural Computation*, 10(6), 1998.

[18] M. Pontil, S. Mukherjee, and F. Girosi. On a novel class of loss functions for robust estimation. A.I. Memo, MIT Artificial Intelligence Laboratory, 1998. (in preparation).

[19] L. Sirovich and M. Kirby. A low dimensional procedure for identifying human faces. *Journal of Optical Society A*, 4:519, 1987.

[20] V. Vapnik. *The Nature of Statistical Learning Theory.* Springer, New York, 1995.

[21] V. Vapnik, S.E. Golowich, and A. Smola. Support vector method for function approximation, regression estimation, and signal processing. In M. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processings Systems 9*, pages 281–287, Cambridge, MA, 1997. The MIT Press.

[22] V. N. Vapnik. *Estimation of Dependences Based on Empirical Data.* Springer-Verlag, Berlin, 1982.

[23] G. Wahba. *Splines Models for Observational Data.* Series in Applied Mathematics, Vol. 59, SIAM, Philadelphia, 1990.