Statistical Science
2006, Vol. 21, No. 3, 358–362
DOI: 10.1214/088342306000000501
Main article DOI: 10.1214/088342306000000493

Rejoinder

Javier M. Moguerza and Alberto Muñoz

1. INTRODUCTION

We are very grateful to the Executive Editors George Casella and Edward George for their active interest in our paper and for organizing this challenging discussion. We also thank all the discussants for their insightful and stimulating comments.

When we submitted the original manuscript in 2003, we were tempted to go for a more general paper on kernel methods. We decided to focus on support vector machines (SVMs), waiting for a mature development of the new and exciting ideas related to kernel methods, such as manifold learning and other related topics. We will refer to some of these methods below. Let us begin, first, with some general considerations.

Regarding the question of the dimensionality induced by the feature space, Hastie and Zhu remark in their comment that usual kernels do not automatically lead to infinite-dimensional feature spaces. They give a nice example that involves the radial (Gaussian) kernel function. This agrees with results in Keerthi and Lin [8], where an explanation of the performance of the Gaussian kernel is given when, according to the notation in the comment by Hastie and Zhu, $\gamma \to 0$ and λ is chosen in the appropriate way. In this case, the SVM classifier converges to a linear SVM classifier, and the effective dimension of the kernel is finite, agreeing with the empirical conclusion provided by the discussants.

We also agree with the assertions of some of the discussants regarding the probabilistic interpretability of the SVM output (the sign of some estimated function). Our comment was rather along the line of Sollich [18], who proposed to make Bayesian methods available for the support vector methodology, while leaving as much as possible of the standard SVM framework intact. This is not an easy task. In fact, as Bartlett, Jordan and McAuliffe remark, sparseness and the precise estimation of conditional probabilities are hard to reconcile.

Regarding the role of differentiability in SVMs (misplaced in the opinion of Bartlett, Jordan and McAuliffe), it is convenient to recall that the differentiable formulation of the SVM problem allows its solution by the use of standard Newton-type methods for convex optimization. Under the availability of second order derivatives (and this is the case for SVMs), these methods are known to be the most efficient ones for the solution of smooth problems.

We thank some of the discussants for turning the attention of the reader to general kernel methods. In particular, we appreciate the Bartlett, Jordan and McAuliffe effort to make clearer the potential impact of reproducing kernel Hilbert space (RKHS) methods. Regarding the origins of RKHS in statistics, for the sake of completeness, we strongly recommend reading the conversation with Emanuel Parzen in [14].

Given the history of SVMs, perfectly outlined by Wahba in the introduction of her comment, we do not like to think of SVMs as a "modest" variant of some standard statistical methodology (as suggested by Bartlett, Jordan and McAuliffe). Using a similar (a posteriori) reasoning, some strict mathematicians might think that RKHS methods in statistics are just a small variation on the general theory of Hilbert spaces. Of course, this is far from true. We rather think that the support vector methodology, followed closely by kernel methods, has been able to synthesize a variety of techniques from different fields, leading to a more unified framework for learning theory [5]. In addition, the geometrical viewpoint of SVMs allows new approaches to long-familiar problems, as illustrated in the next section.

2. KERNEL METHODS REVISITED

One interesting point regarding the geometrical interpretation of SVMs is that they have stirred the development of new techniques driven by the geometrical properties of the kernel. Some of these techniques have not so far been mentioned in the discussion. We now briefly describe two relevant examples.

2.1 One-Class SVMs

An example of a new method that has arisen from a geometrical point of view is one-class SVMs [16]. One-class SVMs deal with a problem related to estimating high density regions from data samples. The method computes a binary function that takes the value $+1$ in "small" regions that contain most data points and