*LETTER* **Communicated by Christopher Williams**

# Kernel Fisher Discriminants for Outlier Detection

**Volker Roth**
*vroth@inf.ethz.ch*
*ETH Zurich, Institute of Computational Science,*
*CH-8092 Zurich, Switzerland*

**The problem of detecting atypical objects or outliers is one of the classical topics in (robust) statistics. Recently, it has been proposed to address this problem by means of one-class SVM classifiers. The method presented in this letter bridges the gap between kernelized one-class classification and gaussian density estimation in the induced feature space. Having established the exact relation between the two concepts, it is now possible to identify atypical objects by quantifying their deviations from the gaussian model. This model-based formalization of outliers overcomes the main conceptual shortcoming of most one-class approaches, which, in a strict sense, are unable to detect outliers, since the expected fraction of outliers has to be specified in advance. In order to overcome the inherent model selection problem of unsupervised kernel methods, a cross-validated likelihood criterion for selecting all free model parameters is applied. Experiments for detecting atypical objects in image databases effectively demonstrate the applicability of the proposed method in real-world scenarios.**

## 1 Introduction

A one-class-classifier attempts to find a separating boundary between a data set and the rest of the feature space. A natural application of such a classifier is estimating a contour line of the underlying data density for a certain quantile value. Such contour lines may be used to separate typical objects from atypical ones. Objects that look sufficiently atypical are often considered to be outliers, for which one rejects the hypothesis that they come from the same distribution as the majority of the objects. Thus, a useful application scenario would be to find a boundary that separates the jointly distributed objects from the outliers. Finding such a boundary defines a classification problem in which, however, usually only sufficiently many labeled samples from one class are available. In most practical problems, no labeled samples from the outlier class are available at all, and it is even unknown if any outliers are present. Since the contour lines of the data density often have a complicated form, highly nonlinear classification models are needed in such an outlier-detection scenario. Recently, it has been proposed to address this problem by exploiting the modeling power of kernel-based support vector machine (SVM) classifiers (see, e.g., Tax & Duin, 1999; Schölkopf, Williamson, Smola, & Shawe-Taylor, 2000). These one-class SVMs are able to infer highly nonlinear decision boundaries, although at the price of a severe model selection problem.

The approach of directly estimating a boundary, as opposed to first estimating the whole density, follows one of the main ideas in learning theory, which states that one should avoid solving an intermediate problem that is too hard. While this line of reasoning seems to be appealing from a theoretical point of view, it leads to a severe problem in practical applications: when it comes to detecting outliers, the restriction to estimating only a boundary makes it impossible to derive a formal characterization of outliers without prior assumptions on the expected fraction of outliers or even on their distribution. In practice, however, any such prior assumptions can hardly be justified. The fundamental problem of the one-class approach lies in the fact that outlier detection is a (partially) unsupervised task that has been squeezed into a classification framework. The missing part of information has been shifted to prior assumptions that require detailed information about the data source.

This letter aims at overcoming this problem by linking kernel-based one-class classifiers to gaussian density estimation in the induced feature space. Objects that have an unexpected high Mahalanobis distance to the sample mean are considered as atypical objects, or outliers. A particular Mahalanobis distance is considered to be unexpected if it is very unlikely to observe an object that far away from the mean vector in a random sample of a certain size. We formalize this concept in section 3 by way of fitting linear models in quantile-quantile plots. The main technical ingredient of our method is the one-class kernel Fisher discriminant classifier (OC-KFD), for which the relation to gaussian density estimation is shown. From the classification side, the OC-KFD-based model inherits both the modeling power of Mercer kernels and the simple complexity control mechanism of regularization techniques. Viewed as a function in the input space variables, the model can be viewed as a nonparametric density estimator. The explicit relation to gaussian density estimation in the kernel-induced feature space, however, makes it possible to formalize the notion of an atypical object by observing deviations from the gaussian model. Like any other kernel-based algorithm, however, the OC-KFD model contains some free parameters that control the complexity of the inference mechanism, and it is clear that deviations from gaussianity will heavily depend on the actual choice of these model parameters. In order to characterize outliers, it is thus necessary to select a suitable model in advance. This model selection problem is overcome by using a likelihood-based cross-validation framework for inferring the free parameters.

## 2 Gaussian Density Estimation and One-Class LDA

Let $X$ denote the $n \times d$ data matrix that contains the $n$ input vectors $\boldsymbol{x}_i \in \mathbb{R}^d$ as rows. It has been proposed to estimate a one-class decision boundary by separating the data set from the origin (Schölkopf et al., 2000), which effectively coincides with replicating all $\boldsymbol{x}_i$ with the opposite sign and separating $X$ and $-X$. Typically, a $\nu$-SVM classifier with a radial basis kernel function is used. The parameter $\nu$ upper-bounds the fraction of outliers in the data set and must be selected a priori. There are, however, no principled ways of choosing $\nu$ in a general (unsupervised) outlier-detection scenario. Such unsupervised scenarios are characterized by the lack of class labels that would assign the objects to either the typical class or the outlier class.

The method proposed here follows the same idea of separating the data from their negatively replicated counterparts. Instead of an SVM, however, a kernel Fisher discriminant (KFD) classifier is used (Mika, Rätsch, Weston, Schölkopf, & Müller, 1999; Roth & Steinhage, 2000). The latter has the advantage that is is closely related to gaussian density estimation in the induced feature space. By making this relation explicit, outliers can be identified without specifying the expected fraction of outliers in advance. We start with a linear discriminant analysis (LDA) model and then introduce kernels. The intuition behind this relation to gaussian density estimation is that discriminant analysis assumes a gaussian class-conditional data density.

Let $X_a = (X, -X)^\top$ denote the augmented ($2n \times d$) data matrix which also contains the negative samples $-\boldsymbol{x}_i$. Without loss of generality, we assume that the sample mean $\boldsymbol{\mu}_+ := \sum_i \boldsymbol{x}_i > 0$, so that the sample means of the positive data and the negative data differ: $\boldsymbol{\mu}_+ \neq \boldsymbol{\mu}_-$. Let us now assume that our data are realizations of a normally distributed random variable in $d$ dimensions: $\mathcal{X} \sim N_d(\boldsymbol{\mu}, \Sigma)$. Denoting by $X^c$ the centered data matrix, the estimator for $\Sigma$ takes the form

$$\hat{\Sigma} = (1/n) X^{c\top} X^c =: W.$$

The LDA solution $\boldsymbol{\beta}_*$ maximizes the between-class scatter $\boldsymbol{\beta}_*^\top B \boldsymbol{\beta}_*$ with $B = \boldsymbol{\mu}_+ \boldsymbol{\mu}_+^\top + \boldsymbol{\mu}_- \boldsymbol{\mu}_-^\top$ under the constraint on the within-class scatter $\boldsymbol{\beta}_*^\top W \boldsymbol{\beta}_* = 1$. Note that in our special case with $X_a = (X, -X)^\top$, the usual pooled within-class matrix $W$ simply reduces to the above-defined $W = (1/n) X^{c\top} X^c$. It is well known (see, e.g., Duda, Hart, & Stork, 2001) that the LDA solution (up to a scaling factor) can be found by minimizing a least-squares functional,

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \| \boldsymbol{y}_a - X_a \boldsymbol{\beta} \|^2, \tag{2.1}$$

where $\boldsymbol{y}_a = (2, \ldots, 2, -2, \ldots, -2)^\top$ denotes a $2n$-indicator vector for class membership in class + or −. In (Hastie, Buja, & Tibshirani, 1995), a slightly more general form of the problem is described where the above functional is minimized under a constrained on $\boldsymbol{\beta}$, which in the simplest case amounts to adding a term $\gamma\boldsymbol{\beta}^\top\boldsymbol{\beta}$ to the functional. Such a ridge regression model assumes a penalized total covariance of the form $T = (1/(2n)) \cdot X_a^\top X_a + \gamma I = (1/n) \cdot X^\top X + \gamma I$. Defining an $n$-vector of ones $\boldsymbol{y} = (1, \ldots, 1)^\top$, the solution vector $\hat{\boldsymbol{\beta}}$ reads

$$\hat{\boldsymbol{\beta}} = \left(X_a^\top X_a + \gamma I\right)^{-1} X_a^\top \boldsymbol{y}_a = (X^\top X + \gamma I)^{-1} X^\top \boldsymbol{y}. \tag{2.2}$$

An appropriate scaling factor is defined in terms of the quantity

$$s^2 = (1/n) \cdot \boldsymbol{y}^\top \hat{\boldsymbol{y}} = (1/n) \cdot \boldsymbol{y}^\top X\hat{\boldsymbol{\beta}}, \tag{2.3}$$

which leads us to the correctly scaled LDA vector $\boldsymbol{\beta}_* = s^{-1}(1 - s^2)^{-1/2}\hat{\boldsymbol{\beta}}$ that fulfills the normalization condition $\boldsymbol{\beta}_*^\top W \boldsymbol{\beta}_* = 1$.

One further derives from Hastie et al. (1995) that the mean vector of $X$, projected onto the one-dimensional LDA subspace, has the coordinate value $m_+ = s(1 - s^2)^{-1/2}$, and that the Mahalanobis distance from a vector $\boldsymbol{x}$ to the sample mean $\boldsymbol{\mu}_+$ is the sum of the squared Euclidean distance in the projected space and an orthogonal distance term:

$$\begin{aligned} D(\boldsymbol{x}, \boldsymbol{\mu}_+) &= (\boldsymbol{\beta}_*^\top \boldsymbol{x} - m_+)^2 + D_\perp \\ \text{with } D_\perp &= -(1 - s^2)(\boldsymbol{\beta}_*^\top \boldsymbol{x})^2 + \boldsymbol{x}^\top T^{-1} \boldsymbol{x}. \end{aligned} \tag{2.4}$$

While in the standard LDA setting, all discriminative information is contained in the first term, we have to add the orthogonal term $D_\perp$ to establish the link to density estimation. Note, however, that it is the term $D_\perp$ that makes the density estimation model essentially different from OC classification: while the latter considers only distances in the direction of the projection vector $\boldsymbol{\beta}$, the true density model also takes into account the distances in the orthogonal subspace.

Since the assumption $\mathcal{X} \sim N_d(\boldsymbol{\mu}, \Sigma)$ is very restrictive, we propose to relax it by assuming that we have found a suitable transformation of our input data $\boldsymbol{\phi} : \mathbb{R}^d \mapsto \mathbb{R}^p,\ \boldsymbol{x} \mapsto \boldsymbol{\phi}(\boldsymbol{x})$, such that the transformed data are gaussian in $p$ dimensions. If the transformation is carried out implicitly by introducing a Mercer kernel $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$, we arrive at an equivalent problem in terms of the kernel matrix $K = \Phi\Phi^\top$ and the expansion coefficients $\boldsymbol{\alpha}$:

$$\hat{\boldsymbol{\alpha}} = (K + \gamma I)^{-1} \boldsymbol{y}. \tag{2.5}$$

From Schölkopf et al. (1999) it follows that the mapped vectors can be represented in $\mathbb{R}^n$ as $\boldsymbol{\phi}(\boldsymbol{x}) = K^{-1/2}\boldsymbol{k}(\boldsymbol{x})$, with $\boldsymbol{k}(\boldsymbol{x}) = (k(\boldsymbol{x}, \boldsymbol{x}_1), \ldots, k(\boldsymbol{x}, \boldsymbol{x}_n))^\top$.[1]

Finally we derive the following form of the Mahalanobis distances, which again consists of the Euclidean distance in the classification subspace plus an orthogonal term,

$$D(\boldsymbol{x}, \boldsymbol{\mu}_+) = (\boldsymbol{\alpha}_*^\top \boldsymbol{k}(\boldsymbol{x}) - m_+)^2 - (1 - s^2)(\boldsymbol{\alpha}_*^\top \boldsymbol{k}(\boldsymbol{x}))^2 + n\Omega(\boldsymbol{x}), \tag{2.6}$$

where $\boldsymbol{\alpha}_* = s^{-1}(1 - s^2)^{-1/2}\hat{\boldsymbol{\alpha}}$ and $\Omega(\boldsymbol{x}) = \boldsymbol{k}^\top(\boldsymbol{x})(K + \gamma I)^{-1}K^{-1}\boldsymbol{k}(\boldsymbol{x})$.

Equation 2.6 establishes the desired link between OC-KFD and gaussian density estimation, since for our outlier detection mechanism, only Mahalanobis distances are needed. While it seems to be rather complicated to estimate a density by the above procedure, the main benefit over directly estimating the mean and the covariance lies in the inherent complexity regulation properties of ridge regression. Such a complexity control mechanism is of particular importance in highly nonlinear kernel models. Moreover, for ridge regression models, it is possible to analytically calculate the effective degrees of freedom, a quantity that will be of particular interest when it comes to detecting outliers.

## 3 Detecting Outliers

Let us assume that the model is completely specified: both the kernel function $k(\cdot, \cdot)$ and the regularization parameter $\gamma$ are fixed. The central lemma that helps us to detect outliers can be found in most statistical textbooks:

**Lemma 1.** *Let $\mathcal{X}$ be a gaussian random variable $\mathcal{X} \sim N_d(\boldsymbol{\mu}, \Sigma)$. Then $\Delta := (\mathcal{X} - \boldsymbol{\mu})^\top \Sigma^{-1}(\mathcal{X} - \boldsymbol{\mu})$ follows a $\chi^2$- distribution on $d$ degrees of freedom.*

For the penalized regression models, it might be more appropriate to use the effective degrees of freedom $df$ instead of $d$ in the above lemma. In the case of one-class LDA with ridge penalties, we can easily estimate it as $df = trace(X(X^\top X + \gamma I)^{-1}X^\top)$ (Moody, 1992), which for a kernel model translates into

$$df = trace(K(K + \gamma I)^{-1}). \tag{3.1}$$

[1] With a slight abuse of notation, we denote both the map and its empirical counterpart by $\boldsymbol{\phi}(\boldsymbol{x})$.

The intuitive interpretation of the quantity *df* is the following: denoting by $V$ the matrix of eigenvectors of $K$ and by $\{\lambda_i\}_{i=1}^n$ the corresponding eigenvalues, the fitted values $\hat{\boldsymbol{y}}$ read

$$\hat{\boldsymbol{y}} = V \mathrm{diag}\{\underbrace{\lambda_i/(\lambda_i+\gamma)}_{=:\delta_i}\} V^\top \boldsymbol{y}. \tag{3.2}$$

It follows that compared to the unpenalized case, where all eigenvectors $\boldsymbol{v}_i$ are constantly weighted by 1, the contribution of the $i$th eigenvector $\boldsymbol{v}_i$ is downweighted by a factor $\delta_i/1 = \delta_i$. If the ordered eigenvalues decrease rapidly, however, the values $\delta_i$ are either close to zero or close to one, and *df* determines the number of terms that are essentially different from zero. A similar interpretation is possible for the orthogonal distance term in equation 2.6.

From lemma 1, we conclude that if the data are well described by a gaussian model in the kernel feature space, the observed Mahalanobis distances should look like a sample from a $\chi^2$-distribution with *df* degrees of freedom. A graphical way to test this hypothesis is to plot the observed quantiles against the theoretical $\chi^2$ quantiles, which in the ideal case gives a straight line. Such a quantile-quantile plot is constructed as follows. Let $\Delta_{(i)}$ denote the observed Mahalanobis distances ordered from lowest to highest, and $p_i$ the cumulative proportion before each $\Delta_{(i)}$ given by $p_i = (i - 1/2)/n$. Further, let $z_i = F^{-1}p_i$ denote the theoretical quantile at position $p_i$, where $F$ is the cumulative $\chi^2$-distribution function. The quantile-quantile plot is then obtained by plotting $\Delta_{(i)}$ against $z_i$.

Deviations from linearity can be formalized by fitting a linear model on the observed quantiles and calculating confidence intervals around the fit. Observations falling outside the confidence interval are then treated as outliers. A potential problem of this approach is that the outliers themselves heavily influence the quantile-quantile fit. In order to overcome this problem, the use of robust fitting procedures has been proposed in the literature (see, e.g., Huber, 1981). In the experiments below we use an M-estimator with Huber loss function.

For estimating confidence intervals around the fit, we use the standard formula (see, e.g., Fox, 1997; Kendall & Stuart, 1977),

$$\sigma(\Delta_{(i)}) = b \cdot (\chi^2(z_i))^{-1}\sqrt{(p_i(1-p_i))/n}, \tag{3.3}$$

which can be intuitively understood as the product of the slope $b$ and the standard error of the quantiles. A $100(1-\varepsilon)\%$ envelope around the fit is then defined as $\Delta_{(i)} \pm z_{\varepsilon/2}\sigma(\Delta_{(i)})$ where $z_{\varepsilon/2}$ is the $1-(1-\varepsilon)/2$ quantile of the standard normal distribution.

The choice of the confidence level $\varepsilon$ is somewhat arbitrary, and from a conceptual viewpoint, one might even argue that the problem of specifying one free parameter (i.e., the expected fraction of outliers) has simply been transferred into the problem of specifying another one. In practice, however, selecting $\varepsilon$ is a much more intuitive procedure than guessing the fraction of outliers. Whereas the latter requires problem-specific prior knowledge, which is hardly available in practice, the former depends on only the variance of a linear model fit. Thus, $\varepsilon$ can be specified in a problem-independent way. Note that a $100(1-\varepsilon)\%$ envelope defines a relative confidence criterion. Since we identify all objects outside the envelope as outliers and remove them from the model (see algorithm 1), it might be plausible to set $\varepsilon \leftarrow \varepsilon/n$, which defines an absolute criterion.

As described above, we use a robust fitting procedure for the linear quantile fit. To further diminish the influence of the outliers, the iterative exclusion and refitting procedure presented in algorithm 1 has been shown to be very successful.

**Algorithm 1:** Iterative Outlier Detection (Given-Estimated Mahalanobis Distances)

```
repeat
    Fit a robust line into the quantile plot.
    Compute the $100(1-(\varepsilon/n))\%$-envelope.
    Among the objects within the upper quartile range of Mahalanobis
    distances, remove the one with the highest positive deviation from
    the upper envelope.
until no further outliers are present.
```

The OC-KFD model detects outliers by measuring deviations from gaussianity. The reader should notice, however, that for kernel maps, which transform the input data into a higher-Dimensional space, a severe modeling problem occurs: in a strict sense, the gaussian assumption in the feature space will always be violated, since the transformed data lie on a $d$-dimensional submanifold. For regularized kernel models, the effective dimension of the feature space (measured by $df$) can be much lower than the original feature space dimension. If the chosen model parameters induce a feature space where $df$ does not exceed the input space dimension $d$, the gaussian model might still provide a plausible data description. We conclude that the user should be alarmed if the chosen model has $df \gg d$. In such a case, the proposed outlier detection mechanism might produce unreliable results, since one expects large deviations from gaussianity anyway. The experiments presented in section 5, on the other hand, demonstrate that if a model with $df \approx d$ is selected, the OC-KFD approach successfully overcomes the problem of specifying the fraction of outliers in advance, which seems to be inherent in the $\nu$-SVMs.

## 4 Model Selection

In our model, the data are first mapped into some feature space in which a gaussian model is fitted. Mahalanobis distances to the mean of this gaussian are computed by evaluating equation 2.6. The feature space mapping is implicitly defined by the kernel function, for which we assume that it is parameterized by a kernel parameter $\sigma$. For selecting all free parameters in equation 2.6, we are thus left with the problem of selecting $\boldsymbol{\theta} = (\sigma, \gamma)^\top$.

The idea is now to select $\boldsymbol{\theta}$ by maximizing the cross-validated likelihood. From a theoretical viewpoint, the cross-validated (CV) likelihood framework is appealing, since in van der Laan, Dudoit, and Keles (2004), the CV likelihood selector has been shown to asymptotically perform as well as the optimal benchmark selector, which characterizes the best possible model (in terms of Kullback-Leibler divergence to the true distribution) contained in the parametric family.

For kernels that map into a space with dimension $p > n$, however, two problems arise: (1) the subspace spanned by the mapped samples varies with different sample sizes, and (2) not the whole feature space is accessible for vectors in the input space. As a consequence, it is difficult to find a proper normalization of the gaussian density in the induced feature space. This problem can be easily avoided by considering the likelihood in the input space rather than in the feature space; that is, we are looking for a properly normalized density model $p(\boldsymbol{x}|\cdot)$ in some bounded subset $\mathbb{S} \subset \mathbb{R}^d$ such that the contour lines of $p(\boldsymbol{x}|\cdot)$ and the gaussian model in the feature space have the same shape: $p(\boldsymbol{x}_i|\cdot) = p(\boldsymbol{x}_j|\cdot) \Leftrightarrow p(\boldsymbol{\phi}(\boldsymbol{x}_i)|\cdot) = p(\boldsymbol{\phi}(\boldsymbol{x}_j)|\cdot)$.[2] Denoting by $X_n = \{\boldsymbol{x}_i\}_{i=1}^n$ a sample from $p(\boldsymbol{x})$ from which the kernel matrix $K$ is built, a natural input space model is

$$p_n(\boldsymbol{x}|X_n, \boldsymbol{\theta}) = Z^{-1} \exp\left\{-\frac{1}{2} D(\boldsymbol{x};\ X_n, \boldsymbol{\theta})\right\}, \text{ with } Z = \int_{\mathbb{S}} p_n(\boldsymbol{x}|X_n, \boldsymbol{\theta})\, d\boldsymbol{x}, \tag{4.1}$$

where $D(\boldsymbol{x};\ X_n, \boldsymbol{\theta})$ denotes the (parameterized) Mahalanobis distances, equation 2.6, of a gaussian model in the feature space.

Note that this density model in the input space has the same functional form as our gaussian model in the feature space, except for the different normalization constant $Z$. Only the interpretation of the models is different: the input space model is viewed as a function in $\boldsymbol{x}$, whereas the feature space model is treated as a function in $\boldsymbol{\phi}(\boldsymbol{x})$. The former can be viewed as a nonparametric density estimator (note that for RBF kernels, the functional

[2] In order to guarantee integrability, we assume that the input density has a bounded support. Since in practice we have to approximate the integral by sampling anyway, this assumption does not limit the applicability of the proposed method.

form of the Mahalanobis distances in the exponent of equation 4.1 is closely related to a Parzen-window estimator). The feature-space model, on the other hand, defines a parametric density. Having selected the maximum likelihood model in the input space, the parametric form of the corresponding feature space model is then used for detecting atypical objects.

Computing this constant $Z$ in equation 4.1 requires us to solve a normalization integral over the $d$-dimensional space $\mathbb{S}$. Since in general this integral is not analytically tractable for nonlinear kernel models, we propose to approximate $Z$ by a Monte Carlo sampling method. In our experiments, for instance, the VEGAS algorithm (Lepage, 1980), which implements a mixed importance-stratified sampling approach, was a reasonable method for up to 15 input dimensions. The term *reasonable* here is not of a qualitative nature, but refers solely to the time needed for approximating the integral with a sufficient precision. For the ten-dimensional examples presented in the next section, for instance, the sampling takes approximately 1 minute on a standard PC. The choice of the subset $\mathbb{S}$ on which the numerical integration takes place is somewhat arbitrary, but choosing $\mathbb{S}$ to be the smallest hypercube including all training data has been shown to be a reasonable strategy in the experiments.

## 5 Experiments

**5.1 Detecting Outliers in Face Databases.** In a first experiment the performance of the proposed method is demonstrated for an outlier detection task in the field of face recognition. The Olivetti face database[3] contains 10 different images of each of 40 distinct subjects, taken under different lighting conditions and at different facial expressions and facial details (glasses/no glasses). None of the subjects, however, wears sunglasses. All the images are taken against a homogeneous background with the subjects in an upright, frontal position. In this experiment, we additionally corrupted the data set by including two images in which we artificially changed normal glasses to "sunglasses," as can be seen in Figure 1. The goal is to demonstrate that the proposed method is able to identify these two atypical images without any problem-dependent prior assumptions on the number of outliers or on their distribution.

In order to exclude illumination differences, the images are standardized by subtracting the mean intensity and normalizing to unit variance. Each of the 402 images is then represented as a ten-dimensional vector that contains the projections onto the leading 10 eigenfaces (eigenfaces are simply the eigenvectors of the images treated as pixel-wise vectorial objects). From these vectors, a RBF kernel of the form $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2/\sigma)$ is built. In order to guarantee the reproducibility of the results, both the data

[3] See http://www.uk.research.att.com/facedatabase.html.

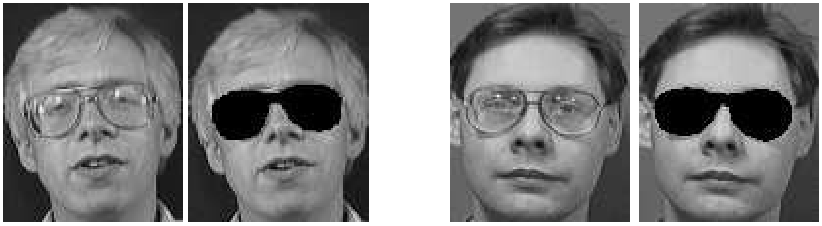

Figure 1: Original and corrupted images with in-painted "sunglasses."

set and an R-script for computing the OC-KFD model can be downloaded from www.inf.ethz.ch/personal/vroth/OC-KFD/index.html.

Whereas our visual impression suggests that the "sunglasses" images might be easy to detect, the automated identification of these outliers based on the eigenface representation is not trivial: due to the high heterogeneity of the database, these images do not exhibit extremal coordinate values in the directions of the leading principal components. For the first principal direction, for example, the coordinate values of the two outlier images are $-0.34$ and $-0.52$, whereas the values of all images range from $-1.5$ to $+1.1$. Another indicator of the difficulty of the problem is that a one-class SVM has severe problems to identify the sunglasses images, as will be shown.

In a first step of the proposed procedure, the free model parameters are selected by maximizing the cross-validated likelihood. A simple two-fold cross-validation scheme is used: the data set is randomly split into a training set and a test set of equal size, the model is built from the training set (including the numerical solution of the normalization integral), and finally the likelihood is evaluated on the test set. This procedure is repeated for different values of $(\sigma, \gamma)$. In order to simplify the selection process, we kept $\gamma = 10^{-4}$ fixed and varied only $\sigma$. Both the test likelihood and the corresponding model complexity measured in terms of the effective degrees of freedom ($df$) are plotted in Figure 2 as a function of the (natural) logarithm of $\sigma$. One can clearly identify both an overfitting and an underfitting regime, separated by a broad plateau of models with similarly high likelihood. The $df$-curve, however, shows a similar plateau, indicating that all these models have comparable complexity. This observation suggests that the results should be rather insensitive to variations of $\sigma$ over values contained in this plateau. This suggestion is indeed confirmed by the results in Figures 3 and 4, where we compared the quantile-quantile plots for different parameter values (marked as I to IV in Figure 2). The plots for models II and III look very similar, and in both cases, two objects clearly fall outside a $100(1 - 0.1/n)$%-envelope around the linear fit. Outside the plateau, the number of objects considered as outliers drastically increases in the overfitting regime (model I, $\sigma$ too small), or decreases to zero in the underfitting regime (model IV, $\sigma$ too large). The upper-right panel in Figure 3

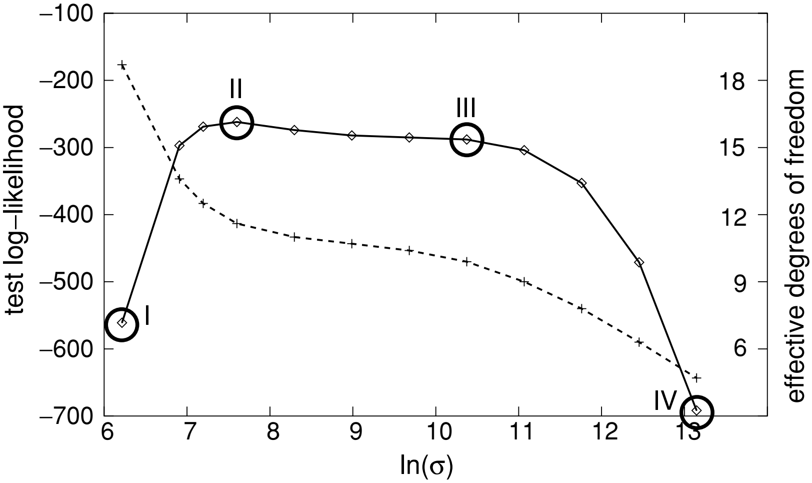

Figure 2: Selecting the kernel width $\sigma$ by cross-validated likelihood (solid line). The dashed line shows the corresponding effective degrees of freedom (*df*).

shows that the two outliers found by the maximum-likelihood model II are indeed the two sunglasses images. Furthermore, we observe that the quantile plot, after removing the two outliers by way of algorithm 1, appears to be gaussian-like (bottom right panel). Despite the potential problems of fitting gaussian models in kernel-induced feature spaces, this observation may be explained by the similar degrees of freedom in the input space and the kernel space: the plateau of the likelihood curve in Figure 2 corresponds to approximately 10 to 11 effective degrees of freedom.

In spite of the fact that the width of the maximum-likelihood RBF kernel is relatively large ($\sigma = 2250$), the kernelized model is still different from a standard linear model. Repeating the model selection experiment with a linear kernel for different values of the regularization parameter $\gamma$, the highest test likelihood is found for a model with $df = 6.5$ degrees of freedom. A reliable identification of the two sunglass images, however, is not possible with this model: one image falls clearly within the envelope, and the other only slightly exceeds the upper envelope.

In order to compare the results with standard approaches to outlier detection, a one-class $\nu$-SVM with RBF kernel is trained on the same data set. The main practical problem with the $\nu$-SVM is the lack of a plausible selection criterion for both the $\sigma$- and the $\nu$-parameter. Taking into account the conceptual similarity of the SVM approach and the proposed OC-KFD method, we decided to use the maximum likelihood kernel emerging from the above model selection procedure (model II in Figure 2). The choice of the $\nu$-parameter that upper-bounds the fraction of outliers turned out to

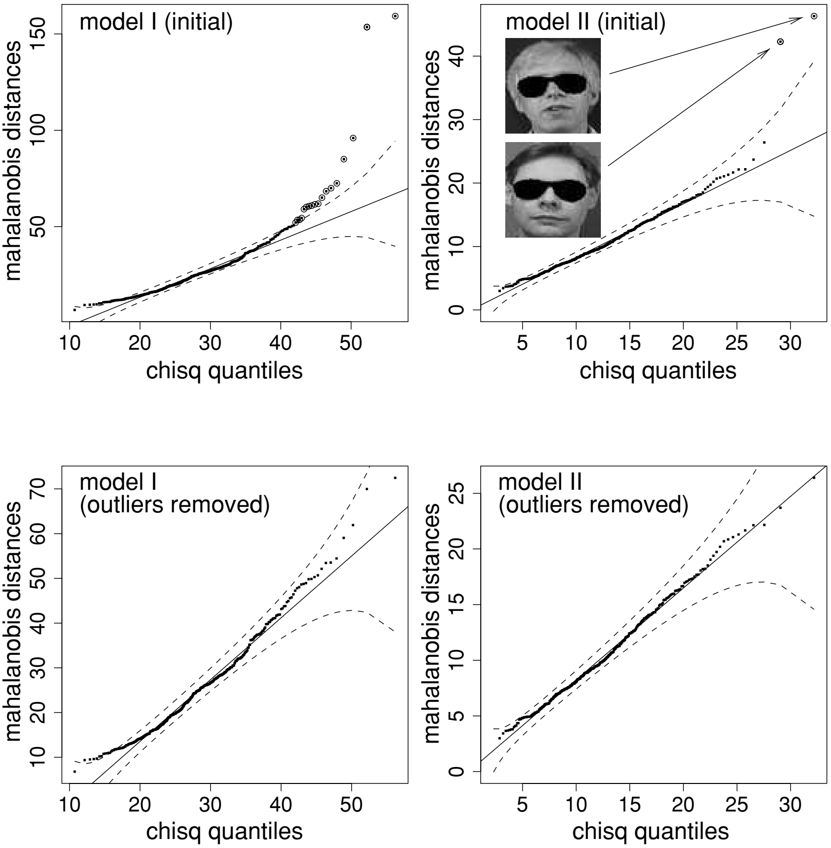

Figure 3: Quantile-quantile plots for different modes. (Left column) Overfitting model I from Figure 2, initial qq-plot (top) and final plot after subsequently removing the outliers (bottom). (Right column) Optimal model II.

be even more problematic: for different $\nu$-values, the SVM model identifies roughly $\nu \cdot 402$ outliers in the data set (cf. Figure 5). Note that in the $\nu$-SVM model, the quantity $(\nu \cdot 402)$, where 402 is the size of the data set, provides an upper bound for the number of outliers. The observed almost linear increase of the detected outliers means, however, that the SVM model does not provide us with a plausible characterization of outliers. We basically "see" as many outliers as we have specified in advance by choosing $\nu$. Furthermore, it is interesting to see that the SVM model has problems to identify the two sunglasses images: for $\nu = 0.0102$, the SVM detects two outliers, which, however, do *not* correspond to the desired sunglass images (see the right panel of Figure 5). To find the sunglasses images within the

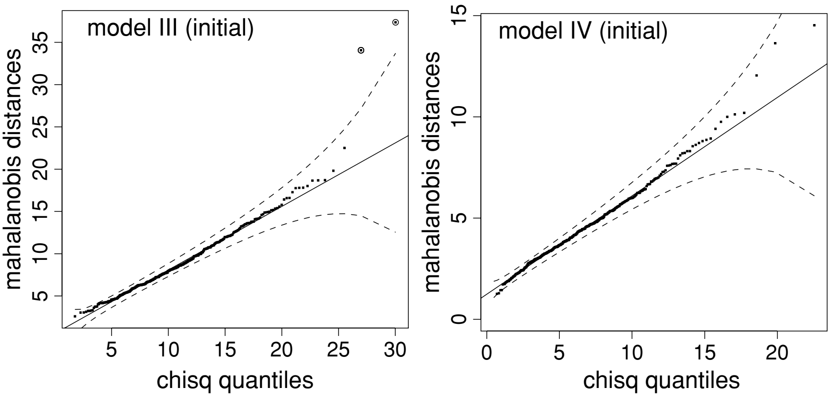

Figure 4: Quantile-quantile plots. (Left) Slightly suboptimal model III. (Right) Underfitting model IV.

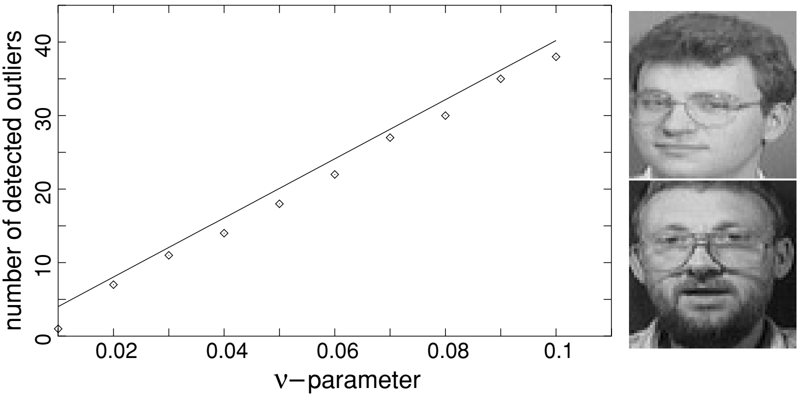

Figure 5: One-class SVM results. (Left) Number of detected outliers versus $\nu$-parameter. The solid line defines the theoretical upper bound $\nu \cdot 402$. (Right) The two outliers identified with $\nu = 0.0102$.

outlier set, we have to increase $\nu$ to 0.026, which “produces” nine outliers in total.

One might argue that the observed correlation between the $\nu$ parameter and the number of identified outliers would be an artifact of using a kernel width that was selected for a different method (i.e., OC-KFD instead of $\nu$-SVM). When the SVM experiments are repeated with both a 10 times smaller width ($\sigma = 225$) and a 10 times larger width ($\sigma = 22{,}500$), one observes the same almost linear dependency of the number of outliers on $\nu$. The problem of identifying the sunglass images remains too: in all tested cases, the most

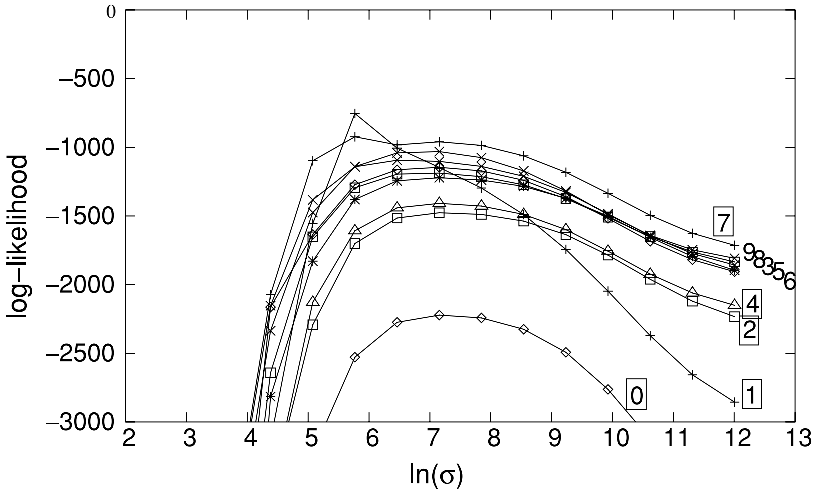

Figure 6: USPS data set. Cross-validated likelihoods as a function of the kernel parameter $\sigma$. Each curve corresponds to a separate digit.

dominant outlier is not a sunglass image. This observation indicates that, at least in this example, the problems of the $\nu$-SVM are rather independent of the used kernel.

**5.2 Handwritten Digits from the USPS Database.** In a second experiment, the proposed method is applied to the USPS database of handwritten digits. The data are divided into a training set and a test set, each consisting of 16 × 16 gray-value images of handwritten digits from postal codes. It is well known that the test data set contains many outliers. The problem of detecting these outlier images has been studied before in the context of one-class SVMs (see Schölkopf & Smola, 2002). Whereas for the face data set, we used the unsupervised eigenface method (i.e., PCA) to derive a low-dimensional data representation, in this case we are given class labels, which allow us to employ a supervised projection method, such as LDA. For 10 classes, LDA produces a data description in a 9-dimensional subspace. For actually computing the projection, a penalized LDA model (Hastie et al., 1995) was fitted to the training set. Given the trained LDA model, the test set vectors were projected on the subspace spanned by the nine LDA vectors. To each of the classes, we then fitted an OC-KFD outlier detection model. The test-likelihood curves for the individual classes are depicted in Figure 6. For most of the classes, the likelihood attains a maximum around $\sigma \approx 1300$. The classes 1 and 7 require a slightly more complex model with $\sigma \approx 350$. The maximum likelihood models correspond to

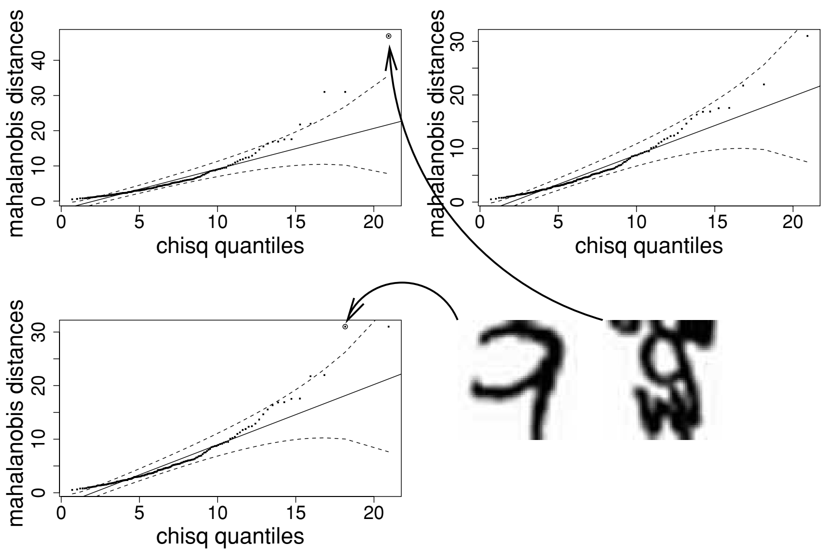

Figure 7: Outlier detection for the digit 9. The iteration terminates after excluding two images (top left → bottom left → top right panel).

approximately 9 to 11 effective degrees of freedom in the kernel space, which is not too far from the input space dimensionality.

Once the optimal kernel parameters are selected, the outliers are detected by iteratively excluding the object with the highest deviation from the upper envelope around the linear quantile fit (cf. algorithm 1). For the digit 9, the individual steps of this iteration are depicted in Figure 7. The iteration terminates after excluding two outlier images. All remaining typical images with high Mahalanobis distances fall within the envelope (top right panel).

For the combined set of outliers for all digits, Figure 8 shows the first 18 outlier images, ordered according to their deviations from the quantile fits. Many European-style 1s and "crossed" 7s are successfully identified as atypical objects in this collection of U.S. postal codes. Moreover, some almost unreadable digits are detected, like the "0," which has the form of a horizontal bar (middle row), or the "9" in the bottom row.

In order to compare the results with a standard technique, a one-class $\nu$-SVM was also trained on the data. As in the previous experiment, the width of the RBF kernel was set to the maximum-likelihood value identified by the above model selection procedure. For the digit 9, the dependency of the number of identified outliers on the $\nu$ parameter is depicted in Figure 9. The almost linear dependency again emphasizes the problem that the SVM approach does not provide us with a meaningful characterization of outliers. Rather, one "sees" (roughly) as many outliers as specified in advance by

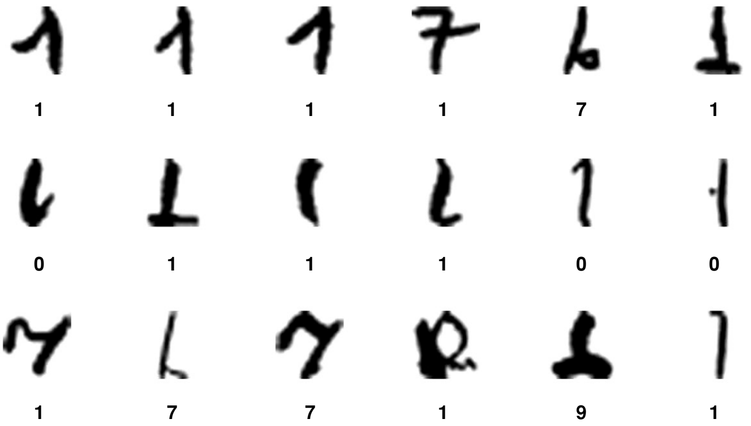

Figure 8: The first 18 detected outliers in the U.S. Postal Service test data set, ordered according to decreasing deviation from the quantile fits. The caption below each image shows the label provided by the database.

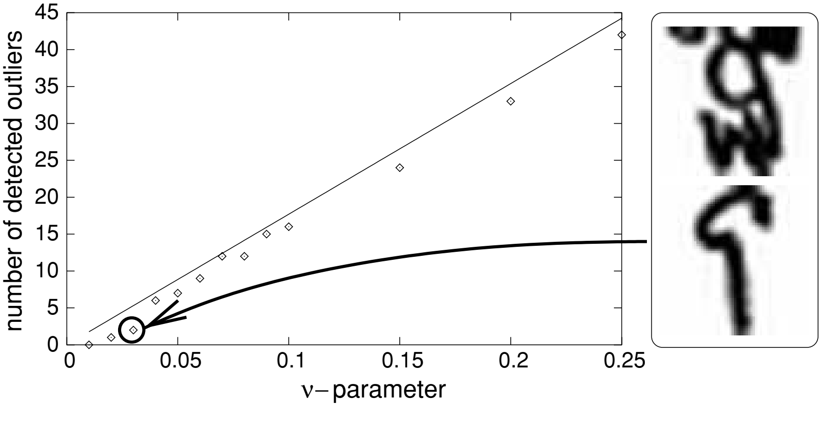

Figure 9: One-class SVM results for the digit 9. (Left) Number of detected outliers as a function of the $\nu$-parameter. The solid line defines the theoretical upper bound $\nu \cdot 177$. (Right) The two outliers identified with $\nu = 0.03$.

choosing $\nu$. When setting $\nu = 0.03$, the SVM identifies two outliers, which equals the number of outliers identified in the OC-KFD experiment. The two outlier images are depicted in the right panel. Comparing the results with that of the OC-KFD approach, we see that both methods identified the almost unreadable 9 as the dominating outlier.

**5.3 Some Implementation Details.** Presumably the easiest way of implementing the model is to carry out an eigenvalue decomposition of $K$. Both the the effective degrees of freedom $df = \sum_i \lambda_i/(\lambda_i + \gamma)$ and the Mahalanobis distances in equation 2.6 can then be derived easily from this decomposition. For practical use, consider the pseudo-code presented in algorithm 2. A complete R-script for computing the OC-KFD model can be downloaded from www.inf.ethz.ch/personal/vroth/OC-KFD/index.html.

**Algorithm 2:** OC-KFD

```
$l_{\max} \leftarrow -\infty$
for $\theta$ on a specified grid do
  Split data into two subsets $X^{\text{train}}$ and $X^{\text{test}}$ of size $n/2$.
  Compute kernel matrix $K(X^{\text{train}}, \sigma)$, its eigenvectors $V$, and eigenvalues $\{\lambda_i\}$.
  Compute $\hat{\boldsymbol{\alpha}} = V \text{diag}\{1/(\lambda_i + \gamma)\} V^\top \boldsymbol{y}$.
  Compute normalization integral $Z$ by Monte Carlo sampling (see equation 4.1).
  Compute Mahalanobis distances by equations 2.6 and 3.2, and evaluate log likelihood on test set: $l(X^{\text{test}}|\boldsymbol{\theta}) = \sum_i -(1/2)D(x_i \in X^{\text{test}}|X^{\text{train}}, \boldsymbol{\theta}) - (n/2)\ln(Z)$.
  if $l(X^{\text{test}}|\boldsymbol{\theta}) > l_{\max}$ then
    $l_{\max} = l(X^{\text{test}}|\boldsymbol{\theta}),\ \boldsymbol{\theta}_{opt} = \boldsymbol{\theta}$.
  end if
end for
  Given $\boldsymbol{\theta}_{opt}$, compute $K(X, \sigma_{opt})$, $V$, $\{\lambda_i\}$.
  Compute Mahalanobis distances and $df$ (see equations 2.6 and 3.2).
  Detect outliers using algorithm 1.
```

## 6 Conclusion

Detecting outliers by way of one-class classifiers aims at finding a boundary that separates typical objects in a data sample from the atypical ones. Standard approaches of this kind suffer from the problem that they require prior knowledge about the expected fraction of outliers. For the purpose of outlier detection, however, the availability of such prior information seems to be an unrealistic (or even contradictory) assumption. The method proposed in this article overcomes this shortcoming by using a one-class KFD classifier directly related to gaussian density estimation in the induced feature space. The model benefits from both the built-in classification method and the explicit parametric density model in the feature space: from the former, it inherits the simple complexity regulation mechanism based on only two free parameters. Moreover, within the classification framework, it is possible to quantify the model complexity in terms of the effective degrees of freedom *df*. The gaussian density model, on the other hand, makes it possible to derive a formal description of atypical objects by way of hypothesis testing: Mahalanobis distances are expected to follow a $\chi^2$ distribution in *df* dimensions, and deviations from this distribution can be quantified by confidence intervals around a fitted line in a quantile-quantile plot.

Since the density model is parameterized by both the kernel function and the regularization constant, it is necessary to select these free parameters before the outlier detection phase. This parameter selection is achieved by observing the cross-validated likelihood for different parameter values and choosing the parameters that maximize this quantity. The theoretical motivation for this selection process follows from van der Laan et. al. (2004), where it has been shown that the cross-validation selector asymptotically performs as well as the so-called benchmark selector, which selects the best model contained in the model family.

The experiments on detecting outliers in image databases effectively demonstrate that the proposed method is able to detect atypical objects without problem-specific prior assumptions on the expected fraction of outliers. This property constitutes a significant practical advantage over the traditional $\nu$-SVM approach. The latter does not provide a plausible characterization of outliers. One "detects" (roughly) as many outliers as one has specified in advance by choosing $\nu$. Prior knowledge about $\nu$, on the other hand, will be hardly available in general outlier-detection scenarios.

In particular, the presented experiments demonstrate that the whole processing pipeline, consisting of model selection by cross-validated likelihood, fitting linear quantile-quantile models, and detecting outliers by considering confidence intervals around the fit, works very well in practical applications with reasonably small input dimensions. For input dimensions $\gg$15, the numerical solution of the normalization integral becomes rather time-consuming when using the VEGAS algorithm. Evaluating the usefulness of more sophisticated sampling models like Markov chain Monte Carlo methods for this particular task will be the subject of future work.

## Acknowledgments

I thank the referees who helped to improve this letter. Special thanks go to Tilman Lange, Mikio Braun, and Joachim M. Buhmann for helpful discussions and suggestions.

## References

Duda, R., Hart, P., & Stork, D. (2001). *Pattern classification*. Hoboken, NJ: Wiley.

Fox, J. (1997). *Applied regression, linear models, and related methods*. Thousand Oaks, CA: Sage.

Hastie, T., Buja, A., & Tibshirani, R. (1995). Penalized discriminant analysis. *Annals of Statistics*, *23*, 73–102.

Huber, P. (1981). *Robust statistics*. Hoboken, NJ: Wiley.

Kendall, M., & Stuart, A. (1977). *The advanced theory of statistics* (Vol. 1). New York: Macmillan.

Lepage, G. (1980). *Vegas: An adaptive multidimensional integration program* (Tech. Rep. CLNS-80/447). Ithaca, NY: Cornell University.

Mika, S., Rätsch, G., Weston, J., Schölkopf, B., & Müller, K.-R. (1999). Fisher discriminant analysis with kernels. In Y.-H. Hu, J. Larsen, E. Wilson, & S. Douglas (Eds.), *Neural networks for signal processing IX* (pp. 41–48). Piscataway, NJ: IEEE.

Moody, J. (1992). The effective number of parameters: An analysis of generalization and regularization in nonlinear learning systems. In J. Moody, S. Hanson, & R. Lippmann (Eds.), *Advances in neural information processing systems, 4* (pp. 847–854). Cambridge, MA: MIT Press.

Roth, V., & Steinhage, V. (2000). Nonlinear discriminant analysis using kernel functions. In S. Solla, T. Leen, & K.-R. Müller (Eds.), *Advances in neural information processing systems, 12* (pp. 568–574). Cambridge, MA: MIT Press.

Schölkopf, B., Mika, S., Burges, C., Knirsch, P., Müller, K.-R., Rätsch, G., & Smola, A. (1999). Input space vs. feature space in kernel-based methods. *IEEE Trans. Neural Networks*, *10*(5), 1000–1017.

Schölkopf, B., & Smola, A. (2002). *Learning with kernels*. Cambridge, MA: MIT Press.

Schölkopf, B., Williamson, R., Smola, A., & Shawe-Taylor, J. (2000). SV estimation of a distribution's support. In S. Solla, T. Leen, & K.-R. Müller (Eds.), *Advances in neural information processing systems, 12* (pp. 582–588). Cambridge, MA: MIT Press.

Tax, D., & Duin, R. (1999). Support vector data description. *Pattern Recognition Letters*, *20*(11–13), 1191–1199.

van der Laan, M., Dudoit, S., & Keles, S. (2004). Asymptotic optimality of likelihood-based cross-validation. *Statistical Applications in Genetics and Molecular Biology*, *3*(1), art. 4.

Received January 27, 2005; accepted August 25, 2005.