Journal of Machine Learning Research 17 (2016) 1-33 Submitted 6/15; Revised 9/16; Published 9/16

Minimum Density Hyperplanes

Nicos G. Pavlidis N.PAVLIDIS@LANCASTER.AC.UK
Department of Management Science
Lancaster University
Lancaster, LA1 4YX, UK

David P. Hofmeyr D.HOFMEYR@LANCASTER.AC.UK
Department of Mathematics and Statistics
Lancaster University
Lancaster, LA1 4YF, UK

Sotiris K. Tasoulis S.TASOULIS@LJMU.AC.UK
Department of Applied Mathematics
Liverpool John Moores University,
Liverpool, L3 3AF, UK

Editor: Andreas Krause

Abstract

Associating distinct groups of objects (clusters) with contiguous regions of high probability density (high-density clusters), is central to many statistical and machine learning approaches to the classification of unlabelled data. We propose a novel hyperplane classifier for clustering and semi-supervised classification which is motivated by this objective. The proposed *minimum density hyperplane* minimises the integral of the empirical probability density function along it, thereby avoiding intersection with high density clusters. We show that the minimum density and the maximum margin hyperplanes are asymptotically equivalent, thus linking this approach to maximum margin clustering and semi-supervised support vector classifiers. We propose a projection pursuit formulation of the associated optimisation problem which allows us to find minimum density hyperplanes efficiently in practice, and evaluate its performance on a range of benchmark data sets. The proposed approach is found to be very competitive with state of the art methods for clustering and semi-supervised classification.

Keywords: low-density separation, high-density clusters, clustering, semi-supervised classification, projection pursuit

1. Introduction

We study the fundamental learning problem: *Given a random sample from an unknown probability distribution with no, or partial label information, identify a separating hyperplane that avoids splitting any of the distinct groups (clusters) present in the sample.* We adopt the cluster definition given by Hartigan (1975, chap. 11), in which a *high-density cluster* is defined as a maximally connected component of the level set of the probability density function, $p(\mathbf{x})$, at level $c \geqslant 0$,

$$\mathrm{lev}_c p(\mathbf{x}) = \left\{ \mathbf{x} \in \mathbb{R}^d \,\middle|\, p(\mathbf{x}) > c \right\}.$$

An important advantage of this approach over other methods is that it is well founded from a statistical perspective, in the sense that a well-defined population quantity is being estimated.

However, since $p(\mathbf{x})$ is typically unknown, detecting high-density clusters necessarily involves estimates of this function, and standard approaches to nonparametric density estimation are reliable only in low dimensions. A number of existing *density clustering* algorithms approximate the level sets of the empirical density through a union of spheres around points whose estimated density exceeds a user-defined threshold (Walther, 1997; Cuevas et al., 2000, 2001; Rinaldo and Wasserman, 2010). The choice of this threshold affects both the shape and number of detected clusters, while an appropriate threshold is typically not known in advance. The performance of these methods deteriorates sharply as dimensionality increases, unless the clusters are assumed to be clearly discernible (Rinaldo and Wasserman, 2010). An alternative is to consider the more specific problem of allocating observations to clusters, which shifts the focus to local properties of the density, rather than its global approximation. The central idea underlying such methods is that if a pair of observations belong to the same cluster they must be connected through a path traversing only high-density regions. Graph theory is a natural choice to address this type of problem. Azzalini and Torelli (2007); Stuetzle and Nugent (2010) and Menardi and Azzalini (2014) have recently proposed algorithms based on this approach. Even these approaches however are limited to problems of low dimensionality by the standards of current applications (Menardi and Azzalini, 2014).

An equivalent formulation of the density clustering problem is to assume that clusters are separated through contiguous regions of low probability density; known as the *low-density separation* assumption. In both clustering and semi-supervised classification, identifying the hyperplane with the maximum margin is considered a direct implementation of the low-density separation approach. Motivated by the success of support vector machines (SVMs) in classification, maximum margin clustering (MMC) (Xu et al., 2004), seeks the maximum margin hyperplane to perform a binary partition (bi-partition) of unlabelled data. MMC can be equivalently viewed as seeking the binary labelling of the data sample that will maximise the margin of an SVM estimated using the assigned labels.

In a plethora of applications data can be collected cheaply and automatically, while labelling observations is a manual task that can be performed for a small proportion of the data only. Semi-supervised classifiers attempt to exploit the abundant unlabelled data to improve the generalisation error over using only the scarce labelled examples. Unlabelled data provide additional information about the marginal density, $p(\mathbf{x})$, but this is beneficial only insofar as it improves the inference of the class conditional density, $p(\mathbf{x}|y)$. Semi-supervised classification relies on the assumption that a relationship between $p(\mathbf{x})$ and $p(\mathbf{x}|y)$ exists. The most frequently assumed relationship is that high-density clusters are associated with a single class (cluster assumption), or equivalently that class boundaries pass through low-density regions (low-density separation assumption). The most widely used semi-supervised classifier based on the low-density separation assumption is the semi-supervised support vector machine (S^3VM) (Vapnik and Sterin, 1977; Joachims, 1999; Chapelle and Zien, 2005). S^3VMs implement the low-density separation assumption by partitioning the data according to the maximum margin hyperplane with respect to both labelled and unlabelled data.

Encouraging theoretical results for semi-supervised classification have been obtained under the cluster assumption. If $p(\mathbf{x})$ is a mixture of class conditional distributions, Castelli and Cover (1995, 1996) have shown that the generalisation error will be reduced exponentially in the number of labelled examples if the mixture is identifiable. More recently, Singh et al. (2009) showed that the mixture components can be identified if $p(\mathbf{x})$ is a mixture of a finite number of smooth density functions, and the separation between mixture components is large. Rigollet (2007) considers the cluster assumption in a nonparametric setting, that is in terms of density level sets, and shows that the generalisation error of a semi-supervised classifier decreases exponentially given a sufficiently large number of unlabelled data. However, the cluster assumption is difficult to verify with a limited number of labelled examples. Furthermore, the algorithms proposed by Rigollet (2007) and Singh et al. (2009) are difficult to implement efficiently even if the cluster assumption holds. This renders them impractical for real-world problems (Ji et al., 2012).

Although intuitive, the claim that maximising the margin over (labelled and) unlabelled data is equivalent to identifying the hyperplane that goes through regions with the lowest possible probability density has received surprisingly little attention. The work of Ben-David et al. (2009) is the only attempt we are aware of to theoretically investigate this claim. Ben-David et al. (2009) quantify the notion of a low-density separator by defining the *density on a hyperplane*, as the integral of the probability density function along the hyperplane. They study the existence of universally consistent algorithms to compute the hyperplane with minimum density. The maximum hard margin classifier is shown to be consistent only in one dimensional problems. In higher dimensions only a soft-margin algorithm is a consistent estimator of the minimum density hyperplane. Ben-David et al. (2009) do not provide an algorithm to compute low density hyperplanes.

This paper introduces a novel approach to clustering and semi-supervised classification which directly identifies low-density hyperplanes in the finite sample setting. In this approach the density on a hyperplane criterion proposed by Ben-David et al. (2009) is directly minimised with respect to a kernel density estimator that employs isotropic Gaussian kernels. The density on a hyperplane provides a uniform upper bound on the value of the empirical density at points that belong to the hyperplane. This bound is tight and proportional to the density on the hyperplane. Therefore, the smallest upper bound on the value of the empirical density on a hyperplane is achieved by hyperplanes that minimise the density on a hyperplane criterion. An important feature of the proposed approach is that the density on a hyperplane can be evaluated exactly through a one-dimensional kernel density estimator, constructed from the projections of the data sample onto the vector normal to the hyperplane. This renders the computation of minimum density hyperplanes tractable even in high dimensional applications.

We establish a connection between the minimum density hyperplane and the maximum margin hyperplane in the finite sample setting. In particular, as the bandwidth of the kernel density estimator is reduced towards zero, the minimum density hyperplane converges to the maximum margin hyperplane. An intermediate result establishes that there exists a positive bandwidth such that the partition of the data sample induced by the minimum density hyperplane is identical to that of the maximum margin hyperplane.

The remaining paper is organised as follows: The formulation of the minimum density hyperplane problem as well as basic properties are presented in Section 2. Section 3

establishes the connection between minimum density hyperplanes and maximum margin hyperplanes. Section 4 discusses the estimation of minimum density hyperplanes and the computational complexity of the resulting algorithm. Experimental results are presented in Section 5, followed by concluding remarks and future research directions in Section 6.

2. Problem Formulation

We study the problem of estimating a hyperplane to partition a finite data set, $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n \subset \mathbb{R}^d$, without splitting any of the high-density clusters present. We assume that $\mathcal{X}$ is an i.i.d. sample of a random variable $\mathbf{X}$ on $\mathbb{R}^d$, with unknown probability density function $p : \mathbb{R}^d \to \mathbb{R}^+$. A hyperplane is defined as $H(\mathbf{v}, b) := \{\mathbf{x} \in \mathbb{R}^d \,|\, \mathbf{v} \cdot \mathbf{x} = b\}$, where without loss of generality we restrict attention to hyperplanes with unit normal vector, i.e., those parameterised by $(\mathbf{v}, b) \in \mathcal{S}^{d-1} \times \mathbb{R}$, where $\mathcal{S}^{d-1} = \{\mathbf{v} \in \mathbb{R}^d \,\big|\, \|\mathbf{v}\| = 1\}$. Following Ben-David et al. (2009) we define the *density on the hyperplane* $H(\mathbf{v}, b)$ as the integral of the probability density function along the hyperplane,

$$I(\mathbf{v}, b) := \int_{H(\mathbf{v},b)} p(\mathbf{x}) \mathrm{d}\mathbf{x}. \tag{1}$$

We approximate $p(\mathbf{x})$ through a kernel density estimator with isotropic Gaussian kernels,

$$\hat{p}(\mathbf{x}|\mathcal{X}, h^2 I) = \frac{1}{n(2\pi h^2)^{d/2}} \sum_{i=1}^{n} \exp\left\{-\frac{\|\mathbf{x} - \mathbf{x}_i\|^2}{2h^2}\right\}. \tag{2}$$

This class of kernel density estimators has the useful property that the integral in Equation (1) can be evaluated exactly by projecting $\mathcal{X}$ onto $\mathbf{v}$; constructing a one-dimensional density estimator with Gaussian kernels and bandwidth h; and evaluating the density at b,

$$\begin{aligned}\hat{I}(\mathbf{v}, b|\mathcal{X}, h^2 I) &:= \int_{H(\mathbf{v},b)} \hat{p}\left(\mathbf{x}|\mathcal{X}, h^2 I\right) \mathrm{d}\mathbf{x}, \\ &= \frac{1}{n\sqrt{2\pi h^2}} \sum_{i=1}^{n} \exp\left\{-\frac{(b - \mathbf{v} \cdot \mathbf{x}_i)^2}{2h^2}\right\} = \hat{p}\left(b \,|\, \{\mathbf{v} \cdot \mathbf{x}_i\}_{i=1}^n, h^2\right). \end{aligned} \tag{3}$$

The univariate kernel estimator $\hat{p}\left(\cdot \,|\, \{\mathbf{v} \cdot \mathbf{x}_i\}_{i=1}^n, h^2\right)$ approximates the *projected density on* $\mathbf{v}$, that is, the density function of the random variable, $X_{\mathbf{v}} = \mathbf{X} \cdot \mathbf{v}$. Henceforth we use $\hat{I}(\mathbf{v}, b)$ to approximate $I(\mathbf{v}, b)$. To simplify terminology we refer to $\hat{I}(\mathbf{v}, b)$ as the *density on* $H(\mathbf{v}, b)$, or the *density integral on* $H(\mathbf{v}, b)$, rather than the empirical density, or the empirical density integral, respectively. For notational convenience we write $\hat{I}(\mathbf{v}, b)$ for $\hat{I}(\mathbf{v}, b|\mathcal{X}, h^2 I)$, where $\mathcal{X}$ and h are apparent from context.

The following Lemma, adapted from (Tasoulis et al., 2010, Lemma 3), shows that $\hat{I}(\mathbf{v}, b)$ provides an upper bound for the maximum value of the empirical density at any point that belongs to the hyperplane.

Lemma 1 *Let $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n \subset \mathbb{R}^d$, and $\hat{p}(\mathbf{x}|\mathcal{X}, h^2 I)$ be a kernel density estimator with isotropic Gaussian kernels. Then, for any $(\mathbf{v}, b) \in \mathcal{S}^{d-1} \times \mathbb{R}$,*

$$\max_{\mathbf{x} \in H(\mathbf{v},b)} \hat{p}(\mathbf{x}|\mathcal{X}, h^2 I) \leqslant \left(2\pi h^2\right)^{\frac{1-d}{2}} \hat{I}(\mathbf{v}, b), \quad \textit{for all } \mathbf{x} \in H(\mathbf{v}, b). \tag{4}$$

This lemma shows that a hyperplane, $H(\mathbf{v}, b)$, cannot intersect level sets of the empirical density with level higher than $\left(2\pi h^2\right)^{\frac{1-d}{2}} \hat{I}(\mathbf{v}, b)$. The proof of the lemma relies on the fact that projection contracts distances, and follows from simple algebra. In Equation (4) equality holds if and only if there exists $\mathbf{x} \in H(\mathbf{v}, b)$ and $\mathbf{c} \in \mathbb{R}^n$ such that all $\mathbf{x}_i \in \mathcal{X}$, can be written as $\mathbf{x}_i = \mathbf{x} + c_i\mathbf{v}$. It is therefore not possible to obtain a uniform upper bound on the value of the empirical density at points that belong to $H(\mathbf{v}, b)$ that is lower than $\left(2\pi h^2\right)^{\frac{1-d}{2}} \hat{I}(\mathbf{v}, b)$ using only one-dimensional projections. Since the upper bound of Lemma 1 is tight and proportional to $\hat{I}(\mathbf{v}, b)$, minimising the density on the hyperplane leads to the lowest upper bound on the maximum value of the empirical density along the hyperplane separator.

To obtain hyperplane separators that are meaningful for clustering and semi-supervised classification, it is necessary to constrain the set of feasible solutions, because the density on a hyperplane can be made arbitrarily low by considering a hyperplane that intersects only the tail of the density. In other words, for any $\mathbf{v}$, $\hat{I}(\mathbf{v}, b)$ can be made arbitrarily low for sufficiently large $|b|$. In both problems the constraints restrict the feasible set to a subset of the hyperplanes that intersect the interior of the convex hull of $\mathcal{X}$. In detail, let $\operatorname{conv} \mathcal{X}$ denote the convex hull of $\mathcal{X}$, and assume $\operatorname{Int}(\operatorname{conv} \mathcal{X}) \neq \emptyset$. Define C to be the set of hyperplanes that intersect $\operatorname{Int}(\operatorname{conv} \mathcal{X})$,

$$C = \left\{ H(\mathbf{v}, b) \,\middle|\, (\mathbf{v}, b) \in \mathcal{S}^{d-1} \times \mathbb{R},\ \exists \mathbf{z} \in \operatorname{Int}(\operatorname{conv} \mathcal{X}) \text{ s.t. } \mathbf{v} \cdot \mathbf{z} = b \right\}. \tag{5}$$

Then denote by F the set of feasible hyperplanes, where $F \subset C$. We define the *minimum density hyperplane* (MDH), $H(\mathbf{v}^\star, b^\star) \in F$ to satisfy,

$$\hat{I}(\mathbf{v}^\star, b^\star) = \min_{(\mathbf{v},b)|H(\mathbf{v},b)\in F} \hat{I}(\mathbf{v}, b). \tag{6}$$

In the following subsections we discuss the specific formulations for clustering and semi-supervised classification in turn.

2.1 Clustering

Since high-density clusters are formed around the modes of $p(\mathbf{x})$, the convex hull of these modes would be a natural choice to define the set of feasible hyperplanes. Unfortunately, this convex hull is unknown and difficult to estimate. We instead propose to constrain the distance of hyperplanes to the origin, b. Such a constraint is inevitable as for any $\mathbf{v} \in \mathcal{S}^{d-1}$, $\hat{I}(\mathbf{v}, b)$ can become arbitrarily close to zero for sufficiently large $|b|$. Obviously, such hyperplanes are inappropriate for the purposes of bi-partitioning as they assign all the data to the same partition. Rather than fixing b to a constant, we constrain it in the interval,

$$F(\mathbf{v}) = \left[\mu_{\mathbf{v}} - \alpha\sigma_{\mathbf{v}}, \mu_{\mathbf{v}} + \alpha\sigma_{\mathbf{v}}\right], \tag{7}$$

where $\mu_{\mathbf{v}}$ and $\sigma_{\mathbf{v}}$ denote the mean and standard deviation, respectively, of the projections $\{\mathbf{v} \cdot \mathbf{x}_i\}_{i=1}^n$. The parameter $\alpha \geqslant 0$, controls the width of the interval, and has a probabilistic interpretation from Chebyshev's inequality. Smaller values of α favour more balanced partitions of the data at the risk of excluding low density hyperplanes that separate clusters more effectively. On the other hand, increasing α increases the risk of separating out only a

few outlying observations. We discuss in detail how to set this parameter in the experimental results section. If $\text{Int}(\text{conv}\,\mathcal{X}) \neq \emptyset$, then there exists $\alpha > 0$ such that the set of feasible hyperplanes for clustering, F_{CL}, satisfies,

$$F_{\text{CL}} = \left\{ H(\mathbf{v}, b) \,\middle|\, (\mathbf{v}, b) \in \mathcal{S}^{d-1} \times \mathbb{R},\, b \in F(\mathbf{v}) \right\} \subset C, \tag{8}$$

where C is the set of hyperplanes that intersect $\text{Int}(\text{conv}\,\mathcal{X})$, as defined in Equation (5).

The minimum density hyperplane for clustering is the solution to the following constrained optimisation problem,

$$\min_{(\mathbf{v},b)\in\mathcal{S}^{d-1}\times\mathbb{R}} \hat{I}(\mathbf{v}, b), \tag{9a}$$

$$\text{subject to:} \quad b - \mu_{\mathbf{v}} + \alpha\sigma_{\mathbf{v}} \geqslant 0, \tag{9b}$$

$$\mu_{\mathbf{v}} + \alpha\sigma_{\mathbf{v}} - b \geqslant 0. \tag{9c}$$

Since the objective function and the constraints are continuously differentiable, MDHs can be estimated through constrained optimisation methods like sequential quadratic programming (SQP). Unfortunately the problem of local minima due to the nonconvexity of the objective function seriously hinders the effectiveness of this approach.

To mitigate this we propose a parameterised optimisation formulation, which gives rise to a projection pursuit approach. Projection pursuit methods optimise a measure of "interestingness" of a linear projection of a data sample, known as the projection index. For our problem the natural choice of projection index for $\mathbf{v}$ is the minimum value of the projected density within the feasible region, $\min_{b\in F(\mathbf{v})} \hat{I}(\mathbf{v}, b)$. This index gives the minimum density integral of feasible hyperplanes with normal vector $\mathbf{v}$. To ensure the differentiability of the projection index we incorporate a penalty term into the objective function. We define the penalised density integral as,

$$f_{\text{CL}}(\mathbf{v}, b) = \hat{I}(\mathbf{v}, b) + \frac{L}{\eta^{\epsilon}} \max\left\{0, \mu_{\mathbf{v}} - \alpha\sigma_{\mathbf{v}} - b, b - \mu_{\mathbf{v}} - \alpha\sigma_{\mathbf{v}}\right\}^{1+\epsilon}, \tag{10}$$

where, $L = \left(e^{1/2}h^2\sqrt{2\pi}\right)^{-1} \geqslant \sup_{b\in\mathbb{R}} \left|\frac{\partial \hat{I}(\mathbf{v},b)}{\partial b}\right|$, $\epsilon \in (0, 1)$ is a constant term that ensures that the penalty function is everywhere continuously differentiable, and $\eta \in (0, 1)$. Other penalty functions are possible, but we only consider the above due to its simplicity, and the fact that its parameters offer a direct interpretation: L in terms of the derivative of the projected density on $\mathbf{v}$; and η in terms of the desired accuracy of the minimisers of $f_{\text{CL}}(\mathbf{v}, b)$ relative to the minimisers of Equation (9), as discussed in the following proposition.

Proposition 2 *For $\mathbf{v} \in \mathcal{S}^{d-1}$, define, the set of minimisers,*

$$B(\mathbf{v}) = \operatorname*{arg\,min}_{b\in F(\mathbf{v})} \hat{I}(\mathbf{v}, b), \tag{11}$$

$$B_C(\mathbf{v}) = \operatorname*{arg\,min}_{b\in\mathbb{R}} f_{\text{CL}}(\mathbf{v}, b) \tag{12}$$

For every $b^\star \in B(\mathbf{v})$ there exists $b^\star_C \in B_C(\mathbf{v})$ such that $|b^\star - b^\star_C| \leqslant \eta$. Moreover, there are no minimisers of $f_{\text{CL}}(\mathbf{v}, b)$ outside the interval $[\mu_v - \alpha\sigma_{\mathbf{v}} - \eta, \mu_v + \alpha\sigma_{\mathbf{v}} + \eta]$,

$$B_C(\mathbf{v}) \cap \mathbb{R}\backslash[\mu_v - \alpha\sigma_{\mathbf{v}} - \eta, \mu_v + \alpha\sigma_{\mathbf{v}} + \eta] = \emptyset.$$

Proof

Any minimiser in the interior of the feasible region, $b^\star \in B(\mathbf{v}) \cap \mathrm{Int}(F(\mathbf{v}))$, also minimises the penalised function, since $f_{\mathrm{CL}}(\mathbf{v}, b) = \hat{I}(\mathbf{v}, b)$ for all $b \in \mathrm{Int}(F(\mathbf{v}))$, hence $b^\star \in B_C(\mathbf{v})$.

Next we consider the case when either or both of the boundary points of $F(\mathbf{v})$, $b^- = \mu_v - \alpha\sigma_{\mathbf{v}}$ and $b^+ = \mu_v + \alpha\sigma_{\mathbf{v}}$, are contained in $B(\mathbf{v})$. It suffices to show that, $f_{\mathrm{CL}}(\mathbf{v}, b) > \hat{I}(\mathbf{v}, b^-)$ for all $b < b^- - \eta$, and $f_{\mathrm{CL}}(\mathbf{v}, b) > \hat{I}(\mathbf{v}, b^+)$ for all $b > b^+ + \eta$. We discuss only the case $b > b^+ + \eta$ as the treatment of $b < b^- - \eta$ is identical. Assume that $\hat{I}(\mathbf{v}, b) < \hat{I}(\mathbf{v}, b^+)$ (since in the opposite case the result follows immediately: $f_{\mathrm{CL}}(\mathbf{v}, b) > \hat{I}(\mathbf{v}, b) > \hat{I}(\mathbf{v}, b^+)$). From the mean value theorem there exists $\xi \in (b^+, b)$ such that,

$$\begin{aligned}
\hat{I}(\mathbf{v}, b^+) &= \hat{I}(\mathbf{v}, b) - (b - b^+) \left.\frac{\partial \hat{I}(\mathbf{v}, b)}{\partial b}\right|_{b=\xi} \\
&\leqslant \hat{I}(\mathbf{v}, b) + (b - b^+)L \\
&< \hat{I}(\mathbf{v}, b) + \frac{L(b - b^+)^{1+\epsilon}}{\eta^\epsilon} = f_{\mathrm{CL}}(\mathbf{v}, b).
\end{aligned}$$

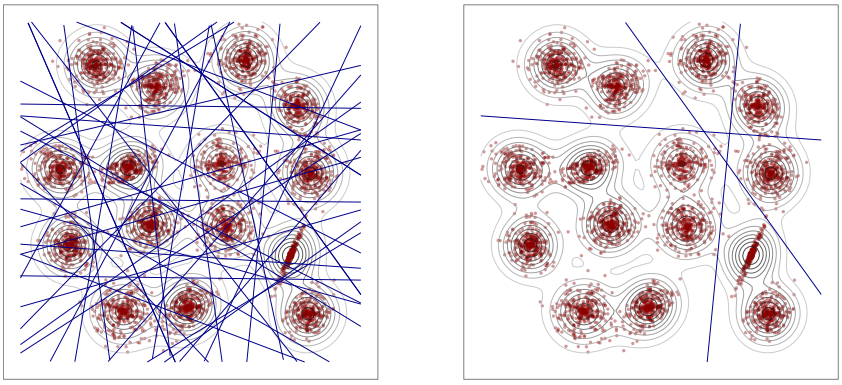
In the above we used the following facts: $\left.\frac{\partial \hat{I}(\mathbf{v},b)}{\partial b}\right|_{b=\xi} < 0$, $L \geqslant \sup_{b\in\mathbb{R}} \left|\frac{\partial \hat{I}(\mathbf{v},b)}{\partial b}\right|$, and $\frac{b-b^+}{\eta} > 1$. ■

We define the projection index for the clustering problem as the minimum of the penalised density integral,

$$\phi_{\mathrm{CL}}(\mathbf{v}) = \min_{b\in\mathbb{R}} f_{\mathrm{CL}}(\mathbf{v}, b). \tag{13}$$

Since the optimisation problem of Equation (13) is one-dimensional it is simple to compute the set of global minimisers $B_C(\mathbf{v})$. As we discuss in Section 4, this is necessary to compute directional derivatives of the projection index, as well as, to determine whether ϕ_{CL} is differentiable. We call the optimisation of ϕ_{CL}, *minimum density projection pursuit* (MDP^2). For each $\mathbf{v}$, MDP^2 considers only the optimal choice of b. This enables it to avoid local minima of $\hat{I}(\mathbf{v}, \cdot)$. Most importantly MDP^2 is able to accommodate a discontinuous change in the location of the global minimiser(s), $\arg\min_{b\in\mathbb{R}} f_{\mathrm{CL}}(\mathbf{v}, b)$, as $\mathbf{v}$ changes. Neither of the above can be achieved when the optimisation is jointly over $(\mathbf{v}, b)$ as in the original constrained optimisation problem, Equation (9). The projection index ϕ_{CL} is continuous, but it is not guaranteed to be everywhere differentiable when $B_C(\mathbf{v})$ is not a singleton. The resulting optimisation problem is therefore nonsmooth and nonconvex.

To illustrate the effectiveness of MDP^2 to estimate MDHs, we compare this approach with a direct optimisation of the constrained problem given in Equation (9) using SQP. To enable visualisation we consider the two-dimensional S1 data set (Fränti and Virmajoki, 2006), constructed by sampling from a Gaussian mixture distribution with fifteen components, where each component corresponds to a cluster. Figure 1 depicts the MDHs obtained over 100 random initialisations of SQP and MDP^2. It is evident that SQP frequently yields hyperplanes that intersect regions with high probability density thus splitting clusters. As SQP always converged in these experiments the poor performance is solely due to convergence to local minima. In contrast, MDP^2 converges to three different solutions over the 100 experiments, all of which induce high quality partitions, and none intersects a high-density



(a) SQP (b) MDP^2

Figure 1: Binary partitions induced by 100 MDHs estimated through SQP and MDP^2

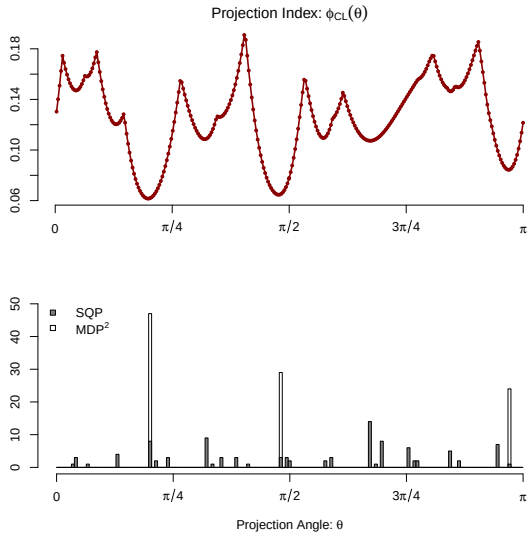


Figure 2: Projection index for S1 data set and solutions obtained through SQP and MDP^2

cluster. In polar coordinates any $\mathbf{v} \in \mathcal{S}^1$ can be parameterised through a single projection angle. Using this parameterisation, the upper plot of Figure 2 depicts the value of the projection index, $\phi_{\mathrm{CL}}(\mathbf{v}(\theta))$, for $\theta \in [0, \pi]$. The lower plot of the figure provides histograms of the distribution of the solutions (locally optimal projection angles) obtained over the 100 experiments with SQP (grey) and MDP^2 (white). The figure shows that $\phi_{\mathrm{CL}}(\mathbf{v})$ is continuous but not everywhere differentiable. The solution most frequently obtained through MDP^2 corresponds to the global optimum, while the only other two solutions identified are the local minimisers with the next two lowest function values. In contrast SQP converges to a much wider range of solutions. Note that this method is not guaranteed to identify the

optimal value of b for any $\mathbf{v}(\theta)$ and this indeed occurs in this example. Therefore the value of $\phi_{\text{CL}}(\mathbf{v})$ is a lower bound for the function values of the minimisers identified through SQP.

2.2 Semi-Supervised Classification

In semi-supervised classification labels are available for a subset of the data sample. The resulting classifier needs to predict as accurately as possible the labelled examples, while avoiding intersection with high-density regions of the empirical density. The MDH formulation can readily accommodate partially labelled data by incorporating the linear constraints associated with the labelled data into the clustering formulation. Without loss of generality assume that the first ℓ examples are labelled by $\mathbf{y} = (y_1, \ldots, y_\ell)^\top \in \{-1, 1\}^\ell$. The MDH for semi-supervised classification is the solution to the problem,

$$\min_{(\mathbf{v},b)\in\mathcal{S}^{d-1}\times\mathbb{R}} \quad \hat{I}(\mathbf{v}, b), \tag{14a}$$

$$\text{subject to:} \quad y_i(\mathbf{v}\cdot\mathbf{x}_i - b) \geqslant 0,\ \forall i = 1, \ldots, \ell, \tag{14b}$$

$$b - \mu_{\mathbf{v}} + \alpha\sigma_{\mathbf{v}} \geqslant 0, \tag{14c}$$

$$\mu_{\mathbf{v}} + \alpha\sigma_{\mathbf{v}} - b \geqslant 0, \tag{14d}$$

where $\hat{I}(\mathbf{v}, b), \mu_{\mathbf{v}}$, and $\sigma_{\mathbf{v}}$ are computed over the entire data set. If the labelled examples are linearly separable the constraints in Equation (14) define a nonempty feasible set of hyperplanes,

$$F_{\text{LB}} = \left\{ H(\mathbf{v}, b) \,|\, (\mathbf{v}, b) \in \mathcal{S}^{d-1} \times \mathbb{R}, b \in F(\mathbf{v}), y_i(\mathbf{v}\cdot\mathbf{x}_i - b) \geqslant 0,\ \forall i \in \{1, \ldots, l\} \right\} \subset C. \tag{15}$$

Equations (14c) and (14d) act as a *balancing constraint* which discourages MDHs that classify the vast majority of unlabelled data to a single class. Balancing constraints are included in the estimation of S^3VMs for the same reason (Joachims, 1999; Chapelle and Zien, 2005).

As in the case of clustering, the direct minimisation of Equation (14) frequently leads to locally optimal solutions. To mitigate this we again propose a projection pursuit formulation. We define the penalised density integral for semi-supervised classification as,

$$f_{\text{SSC}}(\mathbf{v}, b) = f_{\text{CL}}(\mathbf{v}, b) + \gamma \sum_{i=1}^{l} \max\left\{0, -y_i(\mathbf{v}\cdot\mathbf{x}_i - b)\right\}^{1+\epsilon} \tag{16}$$

where, $\gamma > 0$ is a user-defined constant, which controls the trade-off between reducing the density on the hyperplane, and misclassifying the labelled examples. The projection index is then defined as the minimum of the penalised density integral,

$$\phi_{\text{SSC}}(\mathbf{v}) = \min_{b\in\mathbb{R}} f_{\text{SSC}}(\mathbf{v}, b). \tag{17}$$

3. Connection to Maximum Margin Hyperplanes

In this section we discuss the connection between MDHs and maximum (hard) margin hyperplane separators. The margin of a hyperplane $H(\mathbf{v}, b)$ with respect to a data set $\mathcal{X}$ is

defined as the minimum Euclidean distance between the hyperplane and its nearest datum,

$$\operatorname{margin} H(\mathbf{v}, b) = \min_{\mathbf{x} \in \mathcal{X}} |\mathbf{v} \cdot \mathbf{x} - b|. \tag{18}$$

The points whose distance to the hyperplane $H(\mathbf{v}, b)$ is equal to the margin of the hyperplane, that is, $\arg\min_{\mathbf{x} \in \mathcal{X}} |\mathbf{v} \cdot \mathbf{x} - b|$, are called the *support points* of $H(\mathbf{v}, b)$. Let F denote the set of feasible hyperplanes; then the *maximum margin hyperplane* (MMH), $H(\mathbf{v}^m, b^m) \in F$ satisfies,

$$\operatorname{margin} H(\mathbf{v}^m, b^m) = \max_{(\mathbf{v}, b) | H(\mathbf{v}, b) \in F} \operatorname{margin} H(\mathbf{v}, b). \tag{19}$$

The main result of this section is Theorem 5, which states that as the bandwidth parameter, h, is reduced to zero the MDH converges to the MMH. An intermediate result, Lemma 4, shows that there exists a positive bandwidth, $h' > 0$ such that, for all $h \in (0, h')$, the partition of the data set induced by the MDH is identical to that of the MMH.

We first discuss some assumptions which allow us to present the theoretical results of this section. As before we assume a fixed and finite data set $\mathcal{X} \subset \mathbb{R}^d$, and approximate its (assumed) underlying probability density function via a kernel density estimator using Gaussian kernels with isotropic bandwidth matrix $h^2 I$. We assume that the interior of the convex hull of the data, $\operatorname{Int}(\operatorname{conv} \mathcal{X})$, is non-empty, and define C as the set of hyperplanes that intersect $\operatorname{Int}(\operatorname{conv} \mathcal{X})$, as in Equation (5). The set of feasible hyperplanes, F, for either clustering or the semi-supervised classification satisfies $F \subset C$. By construction every $H(\mathbf{v}, b) \in F$ defines a hyperplane which partitions $\mathcal{X}$ into two non-empty subsets. Observe that if for each $\mathbf{v} \in \mathcal{S}^{d-1}$ the set $\{b \in \mathbb{R} \,|\, H(\mathbf{v}, b) \in F\}$ is compact, then by the compactness of $\mathcal{S}^{d-1}$ a maximum margin hyperplane in F exists. For both the clustering and semi-supervised classification problems this compactness holds by construction.

For any $h > 0$, let $(\mathbf{v}_h^\star, b_h^\star) \in \mathcal{S}^{d-1} \times \mathbb{R}$ parameterise a hyperplane which achieves the minimal density integral over all hyperplanes in F, for bandwidth matrix $h^2 I$. That is,

$$\hat{I}(\mathbf{v}_h^\star, b_h^\star) = \min_{(\mathbf{v}, b) | H(\mathbf{v}, b) \in F} \hat{I}(\mathbf{v}, b). \tag{20}$$

Following the approach of Tong and Koller (2000) we first show that as the bandwidth, h, is reduced towards zero, the density on a hyperplane is dominated by its nearest point. This is achieved by establishing that for all sufficiently small values of h, a hyperplane with non-zero margin has lower density integral than any other hyperplane with smaller margin.

Lemma 3 *Take $H(\mathbf{v}, b) \in F$ with non-zero margin and $0 < \delta < \operatorname{margin} H(\mathbf{v}, b) := M_{\mathbf{v},b}$. Then $\exists h' > 0$ such that $h \in (0, h')$ and $M_{\mathbf{w},c} := \operatorname{margin} H(\mathbf{w}, c) \leqslant M_{\mathbf{v},b} - \delta$ implies $\hat{I}(\mathbf{v}, b) < \hat{I}(\mathbf{w}, c)$.*

Proof

Using Equation (3) it is easy to see that,

$$\hat{I}(\mathbf{v}, b) \leqslant \frac{1}{h\sqrt{2\pi}} \exp\left\{ -\frac{M_{\mathbf{v},b}^2}{2h^2} \right\},$$

$$\inf\left\{ \hat{I}(\mathbf{w}, c) \,|\, M_{\mathbf{w},c} \leqslant M_{\mathbf{v},b} - \delta \right\} \geqslant \frac{1}{nh\sqrt{2\pi}} \exp\left\{ -\frac{(M_{\mathbf{v},b} - \delta)^2}{2h^2} \right\}.$$

Therefore,

$$0 \leqslant \lim_{h\to 0^+} \frac{\hat{I}(\mathbf{v},b)}{\inf\left\{\hat{I}(\mathbf{w},c) \,|\, M_{\mathbf{w},c} \leqslant M_{\mathbf{v},b} - \delta\right\}} \leqslant \lim_{h\to 0^+} \frac{n \exp\left\{-\frac{M^2_{\mathbf{v},b}}{2h^2}\right\}}{\exp\left\{-\frac{(M_{\mathbf{v},b}-\delta)^2}{2h^2}\right\}} = 0.$$

Therefore, $\exists h' > 0$ such that $h \in (0, h') \Rightarrow \frac{\hat{I}(\mathbf{v},b)}{\inf\left\{\hat{I}(\mathbf{w},c)\middle| M_{\mathbf{w},c} \leqslant M_{\mathbf{v},b}-\delta\right\}} < 1.$

■

An immediate corollary of Lemma 3 is that as h tends to zero the margin of the MDH tends to the maximum margin. However, this does not necessarily ensure the stronger result that the sequence of MDHs converges to the MMH. To establish this we require two technical results, which describe some algebraic properties of the MMH, and are provided as part of the proof of Theorem 5 which is given in Appendix A.

The next lemma uses the previous result to show that there exists a positive bandwidth, $h' > 0$, such that an MDH estimated using $h \in (0, h')$ induces the same partition of $\mathcal{X}$ as the MMH. The result assumes that the MMH is unique. Notice that if $\mathcal{X}$ is a sample of realisations of a continuous random variable then this uniqueness holds with probability 1.

Lemma 4 *Suppose there is a unique hyperplane in F with maximum margin, which can be parameterised by $(\mathbf{v}^m, b^m) \in \mathcal{S}^{d-1} \times \mathbb{R}$. Then $\exists h' > 0$ s.t. $h \in (0, h') \Rightarrow H(\mathbf{v}^\star_h, b^\star_h)$ induces the same partition of $\mathcal{X}$ as $H(\mathbf{v}^m, b^m)$.*

Proof

Let $M = \text{margin}\, H(\mathbf{v}^m, b^m)$, and let P be the collection of hyperplanes that induce the same partition of $\mathcal{X}$ as that induced by $H(\mathbf{v}^m, b^m)$. Since $\mathcal{X}$ is finite and $H(\mathbf{v}^m, b^m)$ is unique, $\exists \delta > 0$ s.t. $H(w, c) \notin P \Rightarrow \text{margin}\, H(w, c) \leqslant M - \delta$. By Lemma 3, $\exists h' > 0$ s.t.,

$$h \in (0, h') \Rightarrow H(\mathbf{v}^\star_h, b^\star_h) \notin \{H(\mathbf{w}, c) \,|\, \text{margin}\, H(\mathbf{w}, c) \leqslant M - \delta\},$$

therefore $H(\mathbf{v}^\star_h, b^\star_h) \in P$.

■

The next theorem is the main result of this section, and states that the MDH converges to the MMH as the bandwidth parameter is reduced to zero. Notice that by the non-unique representation of hyperplanes, the maximum margin hyperplane has two parameterisations in C, namely $(\mathbf{v}^m, b^m)$ and $(-\mathbf{v}^m, -b^m)$. Convergence to the maximum margin hyperplane is therefore equivalent to showing that,

$$\min\{\|(\mathbf{v}^\star_h, b^\star_h) - (\mathbf{v}^m, b^m)\|, \|(\mathbf{v}^\star_h, b^\star_h) + (\mathbf{v}^m, b^m)\|\} \to 0 \text{ as } h \to 0^+.$$

Theorem 5 *Suppose there is a unique hyperplane in F with maximum margin, which can be parameterised by $(\mathbf{v}^m, b^m) \in \mathcal{S}^{d-1} \times \mathbb{R}$. Then,*

$$\lim_{h\to 0^+} \min\{\|(\mathbf{v}^\star_h, b^\star_h) - (\mathbf{v}^m, b^m)\|, \|(\mathbf{v}^\star_h, b^\star_h) + (\mathbf{v}^m, b^m)\|\} = 0.$$

The set F used in Theorem 5 is generic so it can capture the constraints associated with both clustering and semi-supervised classification, Equations (9) and (14) respectively. In the case of semi-supervised classification we must also assume that the labelled data are linearly separable. Theorem 5 is not directly applicable to the MDP^2 formulations as in this case the function being minimised is not the density on a hyperplane. The next two subsections establish this result for the MDP^2 formulation of the clustering and semi-supervised classification problem.

3.1 MDP^2 for Clustering

We have shown that for the constrained optimisation formulation the MDH converges to the MMH within the feasible set, $F_{\text{CL}} \subset C$. In addition, for a fixed $\mathbf{v}$, Proposition 2 bounds the distance between minimisers of the penalised function f_{CL}, $\arg\min_{b\in\mathbb{R}} f_{\text{CL}}(\mathbf{v}, b)$, and the optimal b of the constrained problem, $\arg\min_{b\in F(\mathbf{v})} \hat{I}(\mathbf{v}, b)$. Combining these we can show that the optimal solution to the penalised MDP^2 formulation converges to the maximum margin hyperplane in F_{CL}, provided the parameters within the penalty term suitably depend on the bandwidth parameter, h. While the general case can be shown, for ease of exposition we make the simplifying assumption that the maximum margin hyperplane is strictly feasible, i.e., if $(\mathbf{v}^m, b^m)$ parameterises the maximum margin hyperplane then $b^m \in (\mu_{\mathbf{v}^m} - \alpha\sigma_{\mathbf{v}^m}, \mu_{\mathbf{v}^m} + \alpha\sigma_{\mathbf{v}^m})$.

For $h, \eta, L > 0$ define $(\mathbf{v}^\star_{h,\eta,L}, b^\star_{h,\eta,L})$ to be any global minimiser of f_{CL}, i.e.,

$$f_{\text{CL}}(\mathbf{v}^\star_{h,\eta,L}, b^\star_{h,\eta,L}) = \min_{(\mathbf{v},b)\in\mathcal{S}^{d-1}\times\mathbb{R}} f_{\text{CL}}(\mathbf{v}, b).$$

Lemma 6 *Suppose there is a unique hyperplane in F_{CL} with maximum margin, which can be parameterised by $(\mathbf{v}^m, b^m) \in \mathcal{S}^{d-1} \times \mathbb{R}$. Suppose further that $b^m \in (\mu_{\mathbf{v}^m} - \alpha\sigma_{\mathbf{v}^m}, \mu_{\mathbf{v}^m} + \alpha\sigma_{\mathbf{v}^m})$. For $h > 0$, let $L(h) = (e^{1/2}h^2\sqrt{2\pi})^{-1}$, and $0 < \eta(h) \leqslant h$. Then,*

$$\lim_{h\to 0^+} \min\{\|(\mathbf{v}^\star_{h,\eta(h),L(h)}, b^\star_{h,\eta(h),L(h)}) - (\mathbf{v}^m, b^m)\|, \|(\mathbf{v}^\star_{h,\eta(h),L(h)}, b^\star_{h,\eta(h),L(h)}) + (\mathbf{v}^m, b^m)\|\} = 0.$$

Proof

Let $M = \text{margin}H(\mathbf{v}^m, b^m)$ and as in the proof of Lemma 4, let $\delta > 0$ be such that any hyperplane inducing a different partition from $H(\mathbf{v}^m, b^m)$ has margin at most $M - \delta$. Consider the set $F^\delta_{\text{CL}} := \{(\mathbf{v}, b) \in \mathcal{S}^{d-1} \times \mathbb{R} | b \in \mathbb{B}_{\delta/2}(F(\mathbf{v}))\}$, where we used the notation $\mathbb{B}_{\delta/2}(F(\mathbf{v}))$ to denote the neighbourhood of $F(\mathbf{v})$ given by $\{r \in \mathbb{R} | d(r, F(\mathbf{v})) < \delta/2\}$. The set F^δ_{CL} increases the feasible set of hyperplanes by allowing b to range in $b \in \mathbb{B}_{\delta/2}(F(\mathbf{v}))$. For any fixed $\mathbf{v}$, the maximum margin of all hyperplanes with normal vector $\mathbf{v}$ can increase by at most $\delta/2$. Thus, any hyperplane inducing a different partition compared to $H(\mathbf{v}^m, b^m)$ has a margin at most $M - \delta/2$. Since $H(\mathbf{v}_m, b_m)$ is strictly feasible it therefore remains the unique maximum margin hyperplane in F^δ_{CL}. Observe now that for $0 < h < \delta/2$ we have $H(\mathbf{v}^\star_{h,\eta(h),L(h)}, b^\star_{h,\eta(h),L(h)}) \in F^\delta_{\text{CL}}$, by Proposition 2. In addition, by Theorem 5, we know that the minimisers of $\hat{I}(\mathbf{v}, b)$ over F^δ_{CL}, say $H(\mathbf{v}^\delta_h, b^\delta_h)$, satisfy

$$\lim_{h\to 0^+} \min\left\{\|(\mathbf{v}^\delta_h, b^\delta_h) - (\mathbf{v}^m, b^m)\|, \|(\mathbf{v}^\delta_h, b^\delta_h) + (\mathbf{v}^m, b^m)\|\right\} = 0.$$

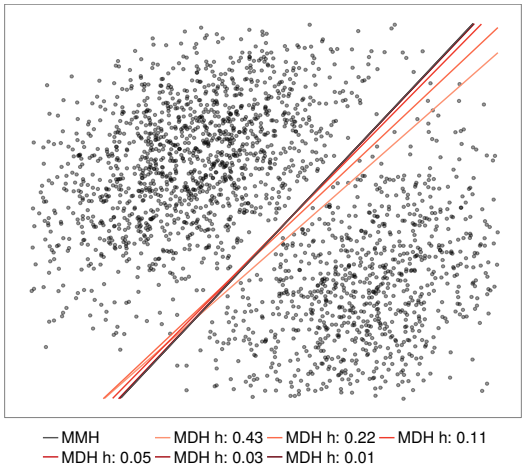


Figure 3: Convergence of the MDH to the maximum margin hyperplane for a decreasing sequence of bandwidth parameters, h.

Now, since $H(\mathbf{v}^m, b^m)$ is strictly feasible $\exists \epsilon' > 0$ s.t. $(\mathbf{v}, b) \in \mathbb{B}_{\epsilon'}(\{(\mathbf{v}^m, b^m), -(\mathbf{v}^m, b^m)\}) \Rightarrow H(\mathbf{v}, b) \in F_{\mathrm{CL}}$. Then for any $0 < \epsilon < \epsilon'$ there exists $h' > 0$ s.t. for $0 < h < h'$ both $(\mathbf{v}_h^\delta, b_h^\delta) \in \mathbb{B}_\epsilon(\{(\mathbf{v}^m, b^m), -(\mathbf{v}^m, b^m)\}) \Rightarrow H(\mathbf{v}_h^\delta, b_h^\delta) \in F_{\mathrm{CL}}$ *and* $H(\mathbf{v}^\star_{h,\eta(h),L(h)}, b^\star_{h,\eta(h),L(h)}) \in F_{\mathrm{CL}}^\delta$. Now for $H(\mathbf{v}, b) \in F_{\mathrm{CL}}^\delta \setminus F_{\mathrm{CL}}$ we know that $\hat{I}(\mathbf{v}, b) < f_{\mathrm{CL}}(\mathbf{v}, b)$, whereas for $H(\mathbf{v}, b) \in F_{\mathrm{CL}}, \hat{I}(\mathbf{v}, b) = f_{\mathrm{CL}}(\mathbf{v}, b)$ and therefore the minimiser of $f_{\mathrm{CL}}(\mathbf{v}, b)$ must lie in the neighbourhood $\mathbb{B}_\epsilon(\{(\mathbf{v}^m, b^m), -(\mathbf{v}^m, b^m)\})$, and the result follows.

■

To illustrate the convergence of the MDH to the MMH we use the two-dimensional data set shown in Figure 3. The data is sampled from a mixture of two Gaussian distributions with equal covariance matrix. The MDH with respect to the true underlying density is $H((1, -1), 0)$. A large margin separator is artificially introduced by removing a few observations in a narrow margin around a hyperplane different from $H((1, -1), 0)$. The margin is intentionally small to ensure that identifying the MMH is non-trivial. Figure 3 illustrates the MDH solutions arising from the MDP^2 method for a decreasing sequence of bandwidths, h. Initially the MDH approximately coincides with the optimal MDH with respect to the true density of the Gaussian mixture. As h decreases, the MDH approaches the MMH and for the smallest values of h the two are indistinguishable.

3.2 MDP^2 for Semi-Supervised Classification

Denote the set of hyperplanes which correctly classify the labelled data by F_{LB}. Under the assumption that $\exists H(\mathbf{v}, b) \in F_{\mathrm{LB}} \cap F_{\mathrm{CL}}$ with non-zero margin, we can show that, provided the parameter γ does not shrink too quickly with h, the hyperplane that minimises f_{SSC} converges to the MMH contained in $F_{\mathrm{LB}} \cap F_{\mathrm{CL}}$, where as before we assume that such an MMH is strictly feasible. To establish this result it is sufficient to show that there exists $h' > 0$ such

that for all $h \in (0, h')$, the optimal hyperplane $H(\mathbf{v}^\star_{h,\eta,L,\gamma}, b^\star_{h,\eta,L,\gamma})$ correctly classifies all the labelled examples. If this holds, then $f_{\mathrm{SSC}}(\mathbf{v}^\star_{h,\eta,L,\gamma}, b^\star_{h,\eta,L,\gamma}) = f_{\mathrm{CL}}(\mathbf{v}^\star_{h,\eta,L,\gamma}, b^\star_{h,\eta,L,\gamma})$ for all sufficiently small h, and hence Lemma 6 can be applied to establish the result. The proof relies on the fact that the penalty terms associated with the known labels in Equation (16) are polynomials in b. Provided that γ is bounded below by a polynomial in h, the value of the penalty terms for hyperplanes that do not correctly classify the labelled data dominate the value of the density integral as h approaches zero. Therefore the optimal hyperplane must correctly classify the labelled data for small values of h.

Lemma 7 *Define* $F_{\mathrm{LB}} = \{H(\mathbf{v}, b) \big| y_i(\mathbf{v}\cdot x_i - b) > 0, \forall i = 1, \ldots, \ell\}$ *and* $F_{\mathrm{CL}} = \{H(\mathbf{v}, b) \big| \mu_{\mathbf{v}} - \alpha\sigma_{\mathbf{v}} \leqslant b \leqslant \mu_{\mathbf{v}} + \alpha\sigma_{\mathbf{v}}\}$ *and assume that* $F_{\mathrm{SSC}} = F_{\mathrm{LB}} \cap F_{\mathrm{CL}} \neq \emptyset$ *and that* $\exists H(\mathbf{v}, b) \in F_{\mathrm{SSC}}$ *with non-zero margin. For* $h > 0$, *let* $L(h) = (e^{1/2}h^2\sqrt{2\pi})^{-1}$, $0 < \eta(h) \leqslant h$ *and* $\gamma(h) \geqslant h^r$ *for some* $r > 0$. *Then* $\exists h' > 0$ *s.t.* $h \in (0, h') \Rightarrow H(\mathbf{v}^\star_{h,\eta(h),L(h),\gamma(h)}, b^\star_{h,\eta(h),L(h),\gamma(h)}) \in F_{\mathrm{LB}}$.

Proof

Consider $H(\mathbf{v}, b) \notin F_{\mathrm{LB}}$. Then,

$$f_{\mathrm{SSC}}(\mathbf{v}, b) \geqslant \frac{1}{n\sqrt{2\pi}h} \exp(-\nu_\star^2/2h^2) + \gamma(h)\nu_\star^{1+\epsilon} > \gamma(h)\nu_\star^{1+\epsilon},$$

where $\nu_\star > 0$ minimises $\frac{1}{n\sqrt{2\pi}h} \exp(-\nu^2/2h^2) + \gamma(h)\nu^{1+\epsilon}$. Therefore, $\nu_\star$ is the unique positive number satisfying,

$$\begin{aligned}
\frac{1}{n\sqrt{2\pi}h} \exp\left(-\frac{\nu_\star^2}{2h^2}\right)\left(-\frac{\nu_\star}{h^2}\right) + (1+\epsilon)\gamma(h)\nu_\star^{\epsilon} &= 0 \\
\Rightarrow \nu_\star^{1-\epsilon} &= (1+\epsilon)\gamma(h)n\sqrt{2\pi}h^3 \exp\left(\frac{\nu_\star^2}{2h^2}\right) \\
\Rightarrow \nu_\star &\geqslant \left((1+\epsilon)\gamma(h)n\sqrt{2\pi}h^3\right)^{1/1-\epsilon}.
\end{aligned}$$

We therefore have,

$$\begin{aligned}
f_{\mathrm{SSC}}(\mathbf{v}, b) &> \gamma(h)\left((1+\epsilon)\gamma(h)n\sqrt{2\pi}h^3\right)^{\frac{1+\epsilon}{1-\epsilon}} \\
&= K\gamma(h)^{\frac{2}{1-\epsilon}} h^{\frac{3(1+\epsilon)}{1-\epsilon}} \\
&\geqslant Kh^{\frac{2r+3(1+\epsilon)}{1-\epsilon}},
\end{aligned}$$

where K is a constant which can be chosen independent of $(\mathbf{v}, b)$. Finally, for any $H(\mathbf{v}', b') \in F_{\mathrm{SSC}}$ with non-zero margin, $\exists h' > 0$ s.t.

$$h \in (0, h') \Rightarrow f_{\mathrm{SSC}}(\mathbf{v}', b') = \hat{I}(\mathbf{v}', b') < Kh^{\frac{2r+3(1+\epsilon)}{1-\epsilon}} < f_{\mathrm{SSC}}(\mathbf{v}, b).$$

Since K is independent of $(\mathbf{v}, b)$, the result follows. The final set of inequalities holds since the hyperplane $H(\mathbf{v}', b')$ is assumed to have non-zero margin, say $M_{\mathbf{v}',b'} > 0$, and hence $\hat{I}(\mathbf{v}', b') \leqslant \frac{1}{h\sqrt{2\pi}} \exp\{-M_{\mathbf{v}',b'}/2h^2\}$, which tends to zero faster than any polynomial in h. ■

4. Estimation of Minimum Density Hyperplanes

In this section we discuss the computation of MDHs. We first investigate the continuity and differentiability properties required to optimise the projection indices $\phi_{\mathrm{CL}}(\mathbf{v})$ and $\phi_{\mathrm{SSC}}(\mathbf{v})$.

Since the domain of both projection indices, $\phi_{\mathrm{CL}}(\mathbf{v})$ and $\phi_{\mathrm{SSC}}(\mathbf{v})$, is the boundary of the unit-sphere in $\mathbb{R}^d$ it is more convenient to express $\mathbf{v}$ in terms of spherical coordinates,

$$v_i(\theta) = \begin{cases} \cos(\theta_i)\prod_{j=1}^{i-1}\sin(\theta_j), & i = 1,\ldots,d-1 \\ \prod_{j=1}^{d-1}\sin(\theta_j), & i = d, \end{cases} \tag{21}$$

where $\theta \in \Theta = [0,\pi]^{d-2} \times [0, 2\pi]$ is called the *projection angle*. Using spherical coordinates renders the domain, Θ, convex and compact, and reduces dimensionality by one.

As the following discussion applies to both $\phi_{\mathrm{CL}}(\mathbf{v})$ and $\phi_{\mathrm{SSC}}(\mathbf{v})$ we denote a generic projection index $\phi : \Theta \to \mathbb{R}$, and the associated set of minimisers, as,

$$\phi(\theta) = \min_{b\in A} f(\mathbf{v}(\theta), b), \tag{22}$$

$$B(\theta) = \left\{ b \in A \,\middle|\, f(\mathbf{v}(\theta), b) = \phi(\theta) \right\}, \tag{23}$$

where $f(\mathbf{v}(\theta), b)$ is continuously differentiable, $A \subset \mathbb{R}$ is compact and convex, and the correspondence $B(\theta)$ gives the set of global minimisers of $f(\mathbf{v}(\theta), b)$ for each θ. The definition of A is not critical in our formulation. Setting,

$$A \supset \left[\min_{\mathbf{v}\in\mathcal{S}^{d-1}} \{\mu_{\mathbf{v}}\} - \alpha\sigma_{\mathrm{pc}_1} - \eta,\ \max_{\mathbf{v}\in\mathcal{S}^{d-1}} \{\mu_{\mathbf{v}}\} + \alpha\sigma_{\mathrm{pc}_1} + \eta \right], \tag{24}$$

where $\sigma^2_{\mathrm{pc}_1}$ is the variance of the projections along the first principal component, ensures that the set of hyperplanes that satisfy the constraint of Equation (7) will be a subset of A for all $\mathbf{v}$.

Berge's maximum theorem (Berge, 1963; Polak, 1987), establishes the continuity of $\phi(\theta)$ and the upper-semicontinuity (u.s.c.) of the correspondence $B(\theta)$. Theorem 3.1 in (Polak, 1987) enables us to establish that $\phi(\theta)$ is locally Lipschitz continuous. Using Theorem 4.13 of Bonnans and Shapiro (2000) we can further show that $\phi(\theta)$ is directionally differentiable everywhere. The directional derivative at θ in the direction ν is given by,

$$d\phi(\theta;\nu) = \min_{b\in B(\theta)} D_\theta f(\mathbf{v}(\theta), b)\cdot\nu, \tag{25}$$

where D_θ denotes the derivative with respect to θ. It is clear from Equation (25) that $\phi(\theta)$ is differentiable if $D_\theta f(\mathbf{v}(\theta), b)$ is the same for all $b \in B(\theta)$. If $B(\theta)$ is a singleton then this condition is trivially satisfied and $\phi(\theta)$ is continuously differentiable at θ.

It is possible to construct examples in which $B(\theta)$ is not a singleton. However, with the exception of contrived examples, our experience with real and simulated data sets indicates that when h is set through standard bandwidth selection rules $B(\theta)$ is almost always a singleton over the optimisation path.

Proposition 8 *Suppose $B(\theta)$ is a singleton for almost all $\theta \in \Theta$. Then $\phi(\theta)$ is continuously differentiable almost everywhere.*

Proof The result follows immediately from the fact that if $B(\theta) = \{b\}$ is a singleton, then the derivative $D\phi(\theta) = D_\theta f(\mathbf{v}(\theta), b)$, which is continuous. ■

Wolfe (1972) has provided early examples of how standard gradient-based methods can fail to converge to a local optimum when used to minimise nonsmooth functions. In the last decade a new class of nonsmooth optimisation algorithms has been developed based on gradient sampling (Burke et al., 2006). Gradient sampling methods use generalised gradient descent to find local minima. At each iteration points are randomly sampled in a radius ε of the current candidate solution, and the gradient at each point is computed. The convex hull of these gradients serves as an approximation of the ε-Clarke generalised gradient (Burke et al., 2002). The minimum element in the convex hull of these gradients is a descent direction. The gradient sampling algorithm progressively reduces the sampling radius so that the convex hull approximates the Clarke generalised gradient. When the origin is contained in the Clarke generalised gradient there is no direction of descent, and hence the current candidate solution is a local minimum. Gradient sampling achieves almost sure global convergence for functions that are locally Lipschitz continuous and almost everywhere continuously differentiable. It is also well documented that it is an effective optimisation method for functions that are only locally Lipschitz continuous.

4.1 Computational Complexity

In this subsection we analyse the computational complexity of MDP^2. At each iteration the algorithm projects the data sample onto $\mathbf{v}(\theta)$ which involves $\mathcal{O}(nd)$ operations. To compute the projection index, $\phi(\theta)$, we need to minimise the penalised density integral, $f(\mathbf{v}(\theta), b)$. This can be achieved by first evaluating $f(\mathbf{v}(\theta), b)$ on a grid of m points, to bracket the location of the minimiser, and then applying bisection to compute the minimiser(s) within the desired accuracy. The main computational cost of this procedure is due to the first step which involves m evaluations of a kernel density estimator with n kernels. Using the improved fast Gauss transform (Morariu et al., 2008) this can be performed in $\mathcal{O}(m + n)$ operations, instead of $\mathcal{O}(mn)$. Bisection requires $\mathcal{O}(-\log_2 \varepsilon)$ iterations to locate the minimiser with accuracy ε.

If the minimiser of the penalised density integral $b^\star = \arg\min_{b \in A} f(\mathbf{v}(\theta), b)$, is unique the projection index is continuously differentiable at θ. To obtain the derivative of the projection index it is convenient to define the projection function, $P(\mathbf{v}) = (\mathbf{x}_1 \cdot \mathbf{v}, \ldots, \mathbf{x}_n \cdot \mathbf{v})^\top$. An application of the chain rule yields,

$$d_\theta \phi = D_\theta f(\mathbf{v}(\theta), b^\star) = D_P f(\mathbf{v}(\theta), b^\star) D_\mathbf{v} P D_\theta \mathbf{v} \tag{26}$$

where the derivative of the projections of the data sample with respect to $\mathbf{v}$ is equal to the data matrix, $D_\mathbf{v} P = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$; and $D_\theta \mathbf{v}$ is the derivative of $\mathbf{v}$ with respect to the projection angle, which yields a $d \times (d-1)$ matrix. The computation of the derivative therefore requires $\mathcal{O}(d(n+d))$ operations.

The original GS algorithm requires $\mathcal{O}(d)$ gradient evaluations at each iteration which is costly. Curtis and Que (2013) have developed an adaptive gradient sampling algorithm that requires $\mathcal{O}(1)$ gradient evaluations in each iteration. More recently, Lewis and Overton (2013) have strongly advocated that for the minimisation of nonsmooth, nonconvex, locally

	n	d	c
banknote[a]	1372	4	2
br. cancer[a]	699	9	2
forest[a]	523	27	4
ionosphere[a]	351	33	2
optidigits[a]	5618	64	10
pendigits[a]	10992	16	10
seeds[a]	210	7	3
smartphone [a]	10929	561	12
image seg.[a]	2309	18	7
satellite[a]	6435	36	6
synth[a]	600	60	6
voting[a]	435	16	2
wine[a]	178	13	3
yeast[b]	698	72	5

a. UCI machine learning repository `https://archive.ics.uci.edu/ml/datasets.html`

b. Stanford Yeast Cell Cycle Analysis Project `http://genome-www.stanford.edu/cellcycle/`

Table 1: Details of benchmark data sets: size (n), dimensionality (d), number of clusters (c).

Lipschitz functions, a simple BFGS method using inexact line searches is much more efficient in practice than gradient sampling, although no convergence guarantees have been established for this method. BFGS requires a single gradient evaluation at each iteration and a matrix vector operation to update the Hessian matrix approximation. In our experiments we use the BFGS algorithm.

5. Experimental Results

In this section we assess the empirical performance of MDHs for clustering and semi-supervised classification. We compare performance with existing state-of-the-art methods for both problems on the following 14 benchmark data sets: Banknote authentication (banknote), Breast Cancer Wisconsin original (br. cancer), Forest type mapping (forest), Ionosphere, Optical recognition of handwritten digits (optidigits), Pen-based recognition of hand-written digits (pendigits), Seeds, Smartphone-Based Recognition of Human Activities and Postural Transitions (smartphone), Statlog Image Segmentation (image seg.), Statlog Landsat Satellite (satellite), Synthetic control chart time series (synth control), Congressional voting records (voting), Wine, and Yeast cell cycle analysis (yeast). Details of these data sets, in terms of their size, n, dimensionality, d and number of clusters, c, can be seen in Table 1.

5.1 Clustering

Since an MDH yields a bi-partition of a data set rather than a complete clustering, we propose two measures to assess the quality of a binary partition of a data set containing an

arbitrary number of clusters. Both take values in $[0, 1]$ with larger values indicating a better partition. These measures are motivated by the fact that a good binary partition should (a) avoid dividing clusters between elements of the partition, and (b) be able to discriminate at least one cluster from the rest of the data. To capture this we modify the cluster labels of the data by assigning each cluster to the element of the binary partition which contains the majority of its members. In the case of a tie the cluster is assigned to the smaller of the two partitions. We thus merge the true clusters into two aggregate clusters, C_1 and C_2.

The first measure we use is the binary V-measure which is simply the V-measure (Rosenberg and Hirschberg, 2007) computed on C_1, C_2 with respect to the binary partition, which we denote Π_1, Π_2. The V-measure is the harmonic mean of homogeneity and completeness. For a data set containing clusters $C_1, \ldots, C_c$, partitioned as $\Pi_1, \ldots, \Pi_k$, homogeneity is defined as the conditional entropy of the cluster distribution within each partition, Π_i. Completeness is symmetric to homogeneity and measures the conditional entropy of each partition within each cluster, C_j. An important characteristic of the V-measure for evaluating binary partitions is that if the distribution of clusters within each partition is equal to the overall cluster distribution in the data set then the V-measure is equal to zero (Rosenberg and Hirschberg, 2007). This means that if an algorithm fails to distinguish the majority of any of the clusters from the remainder of the data, the binary V-measure returns zero performance. Other evaluation metrics for clustering, such as purity and the Rand index, can assign a high value to such partitions.

To define the second performance measure we first determine the number of correctly and incorrectly classified samples. The error of a binary partition, $\mathrm{E}(\Pi_1, \Pi_2)$, given in Equation (27), is defined as the number of elements of each aggregate cluster which are not in the same partition as the majority of their original clusters. In contrast, the success of a partition, $\mathrm{S}(\Pi_1, \Pi_2)$, Equation (28), measures the number of samples which are in the same partition as the majority of their original clusters. The Success Ratio, $\mathrm{SR}(\Pi_1, \Pi_2)$, Equation (29), captures the extent to which the majority of at least one cluster is well-distinguished from the rest of the data.

$$\mathrm{E}(\Pi_1, \Pi_2) = \min\left\{|\Pi_1 \cap C_1| + |\Pi_2 \cap C_2|, |\Pi_1 \cap C_2| + |\Pi_2 \cap C_1|\right\}, \tag{27}$$

$$\mathrm{S}(\Pi_1, \Pi_2) = \min\left\{\max\left\{|\Pi_1 \cap C_1|, |\Pi_1 \cap C_2|\right\}, \max\left\{|\Pi_2 \cap C_1|, |\Pi_2 \cap C_2|\right\}\right\}, \tag{28}$$

$$\mathrm{SR}(\Pi_1, \Pi_2) = \frac{\mathrm{S}(\Pi_1, \Pi_2)}{\mathrm{S}(\Pi_1, \Pi_2) + \mathrm{E}(\Pi_1, \Pi_2)}. \tag{29}$$

The Success Ratio takes the value zero if an algorithm fails to distinguish the majority of any cluster from the remainder of the data.

5.1.1 Parameter Settings for MDP^2

The two most important settings for the performance of the proposed approach are the initial projection direction, and the choice of α, which controls the width of the interval $F(\mathbf{v})$ within which the optimal hyperplane falls. Despite the ability of the MDP^2 formulation to mitigate the effect of local minima of the projected density, the problem remains non-convex and local minima in the projection index can still lead to suboptimal performance. We have found that this effect is amplified in general when either or both the number of dimensions, and the number of high density clusters in the data set is large. To better handle the effect

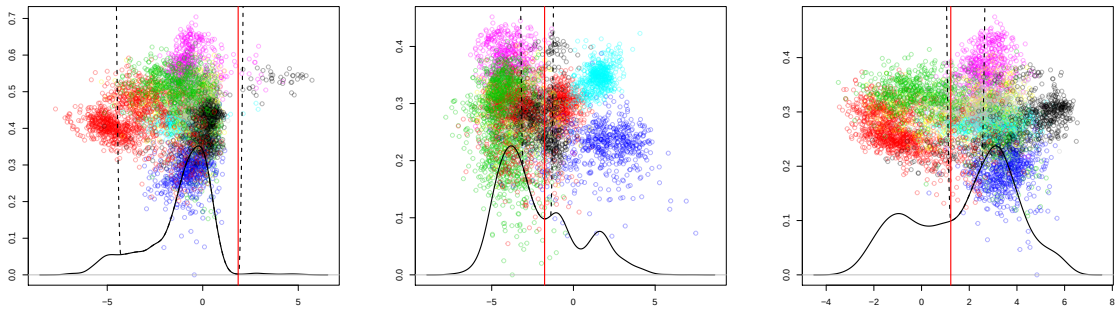
of local optima, we use multiple initialisations and select the MDH that maximises the *relative depth* criterion, defined in Equation (30). The relative depth of an MDH, $H(\mathbf{v}, b)$, is defined as the smaller of the relative differences in the density on the MDH and its two adjacent modes in the projected density,

$$\text{RelativeDepth}(\mathbf{v}, b) = \frac{\min\left\{\hat{I}(\mathbf{v}, m_l), \hat{I}(\mathbf{v}, m_r)\right\} - \hat{I}(\mathbf{v}, b)}{\hat{I}(\mathbf{v}, b)} \tag{30}$$

where m_l and m_r are the two adjacent modes in the projected density on $\mathbf{v}$. If an MDH does not separate the modes of the projected density, then its relative depth is set to zero, signalling a failure of MDP^2 to identify a meaningful bi-partition. The relative depth is appealing because it captures the fact that a high quality separating hyperplane should have a low density integral, and separate well the modes of the projected density. Note also that the relative depth is equivalent to the inverse of a measure used to define cluster overlap in the context of Gaussian mixtures (Aitnouri et al., 2000). In all the reported experiments we initialise MDP^2 to the first and second principal component and select the MDH with the largest relative depth. For the data sets listed above it was never the case that both initialisations led to MDHs with zero relative depth.

The choice of α determines the trade-off between a balanced bi-partition and the ability to discover lower density hyperplanes. The difficulties associated with choosing this parameter are illustrated in Figure 4. In each sub-figure the horizontal axis is the candidate projection vector, $\mathbf{v}$, while the right vertical axis is the direction of maximum variability orthogonal to $\mathbf{v}$. Points correspond to projections of the data sample onto this two-dimensional space, while colour indicates cluster membership. The solid line depicts the projected density on $\mathbf{v}$, while the dotted line depicts the penalised function, $f_{\text{CL}}(\mathbf{v}, \cdot)$. The scale of both functions is depicted on the left vertical axis. The solid vertical line indicates the MDH along $\mathbf{v}$. Setting α to a large value can cause MDP^2 to focus on hyperplanes that have low density because they partition only a small subset of the data set as shown in Figure 4(a). In contrast smaller values of α may cause the algorithm to disregard valid lower density hyperplane separators (see Figure 4(b)), or for the separating hyperplane to not be a local minimiser of the projected density (see Figure 4(c)).

Rather than selecting a single value for α we recommend solving MDP^2 repeatedly for an increasing sequence of values in the range $\{\alpha_{\min}, \alpha_{\max}\}$, where each implementation beyond the first is initialised using the solution to the previous. Setting $\alpha_{\min}$ close to zero forces MDP^2 to seek low density hyperplanes that induce a balanced data partition. This tends to find projections which display strong multimodal structure, yet prevents convergence to hyperplanes that have low density because they partition a few observations, as in the case shown in Figure 4(a). Increasing α progressively fine-tunes the location of the MDH. To avoid sensitivity to the value of $\alpha_{\max}$ (set to 0.9) the output of the algorithm is the last hyperplane that corresponds to a minimiser of the projected density. Figure 5 illustrates this approach using the optical recognition of handwritten digits data set from the UCI machine learning repository (Lichman, 2013). Figure 5(a) depicts the projected density on the initial projection direction, which in this case is the second principal component. As shown, the density is unimodal and the clusters are not well separated along this vector. Although not shown, if a large value of α is used from the outset, MDP^2 will identify a vector



(a) MDH separating few observations

(b) Lower density hyperplane beyond feasible region

(c) MDH not a minimiser of the projected density

Figure 4: Impact of choice of α on minimum density hyperplane.

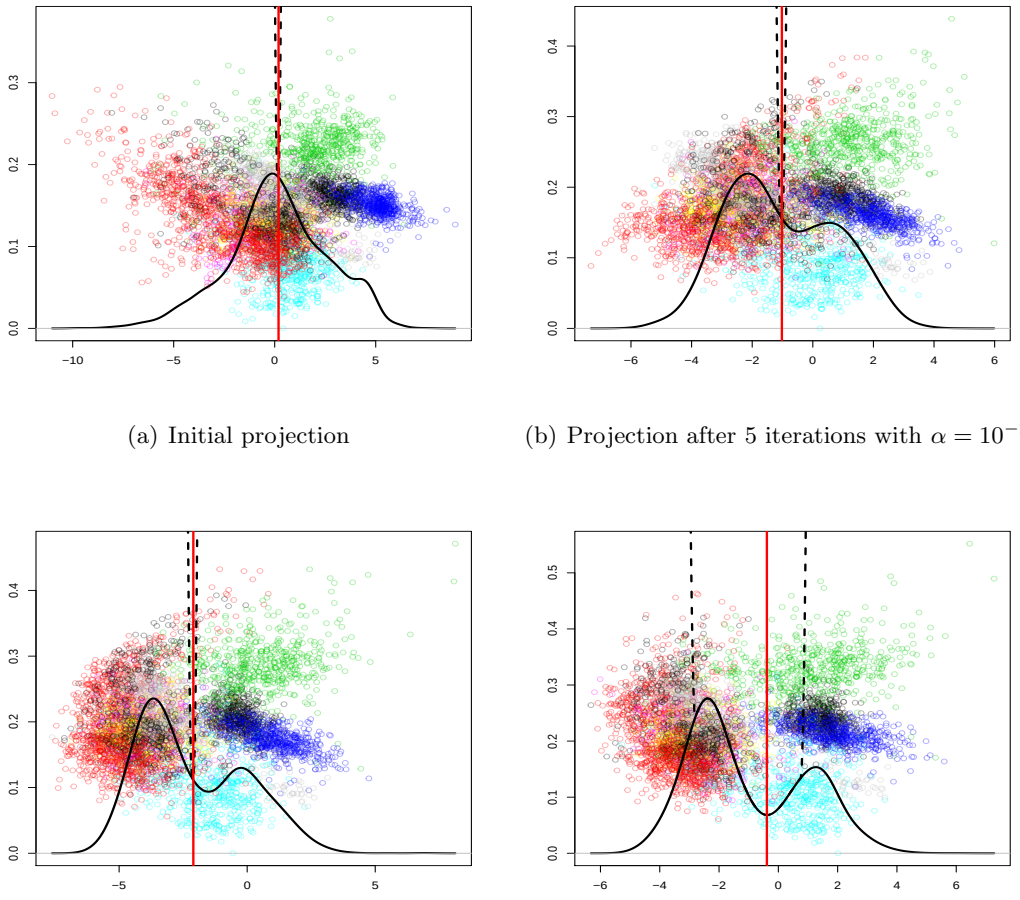
along which the projected density is unimodal and skewed. Figure 5(b) shows that after five iterations with $\alpha = 10^{-2}$ MDP^2 has identified a projection vector with bimodal density. In subsequent iterations the two modes become more clearly separated, Figure 5(c), while increasing α enables MDP^2 to locate an MDH that corresponds to a minimiser of $\hat{I}(\mathbf{v}, b)$, as illustrated in Figure 5(d).

In all experiments we set the bandwidth parameter to $h = 0.9\hat{\sigma}_{\text{pc}_1} n^{-1/5}$, where $\hat{\sigma}_{\text{pc}_1}$ is the estimated standard deviation of the data projected onto the first principal component. This bandwidth selection rule is recommended when the density being approximated is assumed to be multimodal (Silverman, 1986). The parameter η controls the distance between the minimisers of $\arg\min_{b\in\mathbb{R}} f_{\text{CL}}(\mathbf{v}, b)$ and $\arg\min_{b\in F(\mathbf{v})} \hat{I}(\mathbf{v}, b)$, while larger values of ϵ increase the smoothness of the penalised function f_{CL}. Values of η close to zero affect the numerical stability of the one-dimensional optimisation problem, due to the term $\frac{L}{\eta^\epsilon}$ in f_{CL} becoming very large. We used $\eta = 10^{-2}$ and $\epsilon = 1 - 10^{-6}$ to avoid numerical instability. Beyond these numerical problems the values of η and ϵ do not affect the solutions obtained through MDP^2.

5.1.2 Performance Evaluation

We compare the performance of MDP^2 for clustering with the following methods:

1. k-means++ (Arthur and Vassilvitskii, 2007), a version of k-means that is guaranteed to be $\mathcal{O}(\log k)$-competitive to the optimal k-means clustering.
2. The adaptive linear discriminant analysis guided k-means (LDA-km) (Ding and Li, 2007). LDA-km attempts to discover the most discriminative linear subspace for clustering by iteratively using k-means, to assign labels to observations, and LDA to identify the most discriminative subspace.
3. The principal direction divisive partitioning (PDDP) (Boley, 1998), and the density-enhanced PDDP (dePDDP) (Tasoulis et al., 2010). Both methods project the data onto the first principal component. PDDP splits at the mean of the projections, while dePDDP splits at the lowest local minimum of the one-dimensional density estimator.



(a) Initial projection

(b) Projection after 5 iterations with $\alpha = 10^{-2}$

(c) Projection after 25 iterations with $\alpha = 10^{-2}$

(d) Final projection with $\alpha = \alpha_{\max}$

Figure 5: Evolution of the minimum density hyperplane through consecutive iterations.

4. The iterative support vector regression algorithm for MMC (Zhang et al., 2009) using the inner product and Gaussian kernel, iSVR-L and iSVR-G respectively. Both are initialised with the output of 2-means++.
5. Normalised cut spectral clustering (SCn) (Ng et al., 2002) using the Gaussian affinity function, and the automatic bandwidth selection method of Zelnik-Manor and Perona (2004). This choice of kernel and bandwidth produced substantially better performance than alternative choices considered. For data sets that are too large for the eigen decomposition of the Gram matrix to be feasible we employed the Nyström method (Fowlkes et al., 2004).

We also considered the density-based clustering algorithm PdfCluster (Menardi and Azzalini, 2014), but this algorithm could not be executed on the larger data sets and so its performance is not reported in this paper. With the exception of SCn and iSVR-G, the methods considered bi-partition the data through a hyperplane in the original feature

space. For the 2-means and LDA-2m algorithm the hyperplane separator bisects the line segment joining the two centroids. iSVR-L directly seeks the maximum margin hyperplane in the original space, while iSVR-G seeks the maximum margin hyperplane in the feature space defined by the Gaussian kernel. PDDP and dePDDP use a hyperplane whose normal vector is the first principal component. PDDP uses a fixed split point while dePDDP uses the hyperplane with minimum density along the fixed projection direction.

Table 2 reports the performance of the considered methods with respect to the success ratio (SR) and the binary V-measure (V-m) on the fourteen data sets. In addition Figures 6(a) and 6(b) provide summaries of the overall performance on all data sets using boxplots of the raw performance measures as well as the associated *regret*. The regret of an algorithm on a given data set is defined as the difference between the best performance attained on this data set and the performance of this algorithm. By comparing against the best performing clustering algorithm regret accommodates for differences in difficulty between clustering problems, while also making use of the magnitude of performance differences between algorithms. The distribution of performance with respect to both SR and V-m is negatively skewed for most methods, and as a result the median is higher than the mean (indicated with a red dot).

It is clear from Table 2 that no single method is consistently superior to all others, although MDP^2 achieves the highest or tied highest performance on seven data sets (more than any other method). More importantly MDP^2 is among the best performing methods in almost all cases. This fact is better captured by the regret distributions in Figure 6(b). Here we see that the average, median, and maximum regret of MDP^2 is substantially lower than any of the competing methods. In addition MDP^2 achieves the highest mean and median performance with respect to both SR and V-m, while also having much lower variability in performance when compared with most other methods.

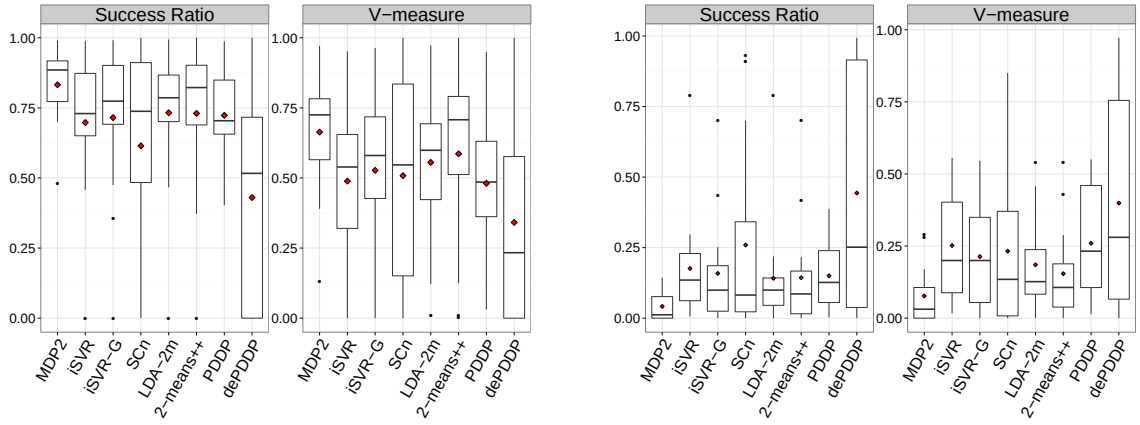
Pairwise comparisons between MDP^2 and other methods reveal some less obvious facts. SCn achieves higher performance than MDP^2 in more examples (six) than any other competing method, however it is much less consistent in its performance, obtaining very poor performance on five of the data sets. The iSVR maximum margin clustering approach is arguably the closest competitor to MDP^2. iSVR-L and iSVR-G achieve the second and third highest average performance with respect to V-m and SR respectively. The PDDP algorithm is the second best performing method on average with respect to SR, but performs poorly with respect to V-m. The density enhanced variant, dePDDP, performs on average much worse than MDP^2. This approach is similarly motivated by obtaining hyperplanes with low density integral, and its low average performance indicates the usefulness of searching for high quality projections as opposed to always using the first principal component. Finally, neither of the k-means variants appears to be competitive with MDP^2 in general.

5.2 Semi-Supervised Classification

In this section we evaluate MDHs for semi-supervised classification. We compare MDHs against three state-of-the-art semi-supervised classification methods: Laplacian Regularised Support Vector Machines (LapSVM) (Belkin et al., 2006), Simple Semi-Supervised Learning (SSSL) (Ji et al., 2012), and Correlated Nyström Views (XNV) (McWilliams et al., 2013). For all methods the inner product kernel was used to render the resulting classifiers linear,

	MDP^2		iSVR-L		iSVR-G		SCn		LDA-2m		2-means++		PDDP		dePDDP	
Data set	SR	V-m	SR	V-m	SR	V-m	SR	V-m	SR	V-m	SR	V-m	SR	V-m	SR	V-m
banknote	**0.79**	**0.55**	0	0	0.35	0	0.46	0.10	0	0.01	0.37	0.01	0.40	0.03	0	0.03
br. cancer	**0.91**	**0.79**	0.73	0.56	0.73	0.56	0	0.13	0.87	0.71	0.87	0.72	0.91	0.78	0.90	0.77
forest	0.78	0.67	0.90	0.72	**0.91**	**0.74**	0.56	0.41	0.76	0.63	0.72	0.58	0.64	0.36	0	0
image seg.	0.89	0.72	0.82	0.59	0.88	0.71	0.92	0.87	0.78	0.58	0.78	0.71	0.87	0.67	**1**	**1**
ionosphere	0.48	0.13	0.47	0.13	0.47	0.13	**0.55**	**0.22**	0.47	0.12	0.47	0.12	0.47	0.12	0.42	0.09
optidigits	**0.93**	**0.85**	0.63	0.29	0.82	0.60	0	0	0.81	0.62	0.92	0.82	0.68	0.30	0	0
pendigits	0.74	0.39	0.79	0.55	**0.88**	**0.68**	0.80	0.68	0.79	0.55	0.78	0.57	0.79	0.54	0.61	0.42
satellite	0.89	0.75	0.73	0.40	0.73	0.40	**0.92**	**0.86**	0.73	0.40	0.87	0.81	0.71	0.37	0	0
seeds	0.88	0.73	0.71	0.53	0.71	0.53	0.89	0.76	**0.96**	**0.90**	0.86	0.70	0.75	0.59	0.73	0.60
smartphone	**0.99**	**0.97**	0.99	0.95	0.99	0.96	0.99	0.94	0.99	0.97	0.99	0.94	0.99	0.95	0	0
synth	0.98	0.94	0.94	0.83	0.94	0.83	**1**	**1**	0.88	0.76	**1**	**1**	0.69	0.51	**1**	**1**
voting	**0.70**	**0.43**	0.46	0.09	0	0	0	0.05	0.69	0.41	0	0	0.70	0.40	0.68	0.38
wine	**0.77**	**0.61**	0.70	0.52	0.69	0.50	0.67	0.48	0.66	0.48	0.68	0.49	0.65	0.46	0.68	0.49
yeast	**0.92**	**0.76**	0.89	0.68	0.91	0.72	0.84	0.61	0.86	0.63	0.91	0.73	0.87	0.65	0	0
Average Improvement	0.13	0.18	0.12	0.14	0.22	0.16	0.10	0.11	0.10	0.08	0.11	0.18	0.40	0.32		

Table 2: Performance on the task of binary partitioning. (Ties in best performance were resolved by considering more decimal places)



(a) Raw Performance Measure

(b) Regret

Figure 6: Performance and Regret Distributions for all Methods Considered

and thereby comparable to our method. As the MDH is asymptotically equivalent to a linear S^{3}VM we also considered the continuous formulation for the estimation of a S^{3}VM proposed by Chapelle and Zien (2005). These results are omitted as this method was not competitive on any of the considered data sets.

5.2.1 Parameter Settings for MDP^2

The existence of a few labelled examples enables an informed initialisation of MDP^2. We consider the first and second principal components as well as the weight vector of a linear SVM trained on the labelled examples only, and initialise MDP^2 with the vector that minimises the value of the projection index, ϕ_{SSC}. The penalty parameter γ is first set to 0.1 and with this setting α is progressively increased in the same way as for clustering. After this, α is kept at $\alpha_{\max}$ and γ is increased to 1 and then 10. Thus the emphasis is initially on finding a low density hyperplane with respect to the marginal density $\hat{p}(\mathbf{x})$. As the algorithm progresses the emphasis on correctly classifying the labelled examples increases, so as to obtain a hyperplane with low training error within the region of low density already determined.

5.2.2 Performance Evaluation

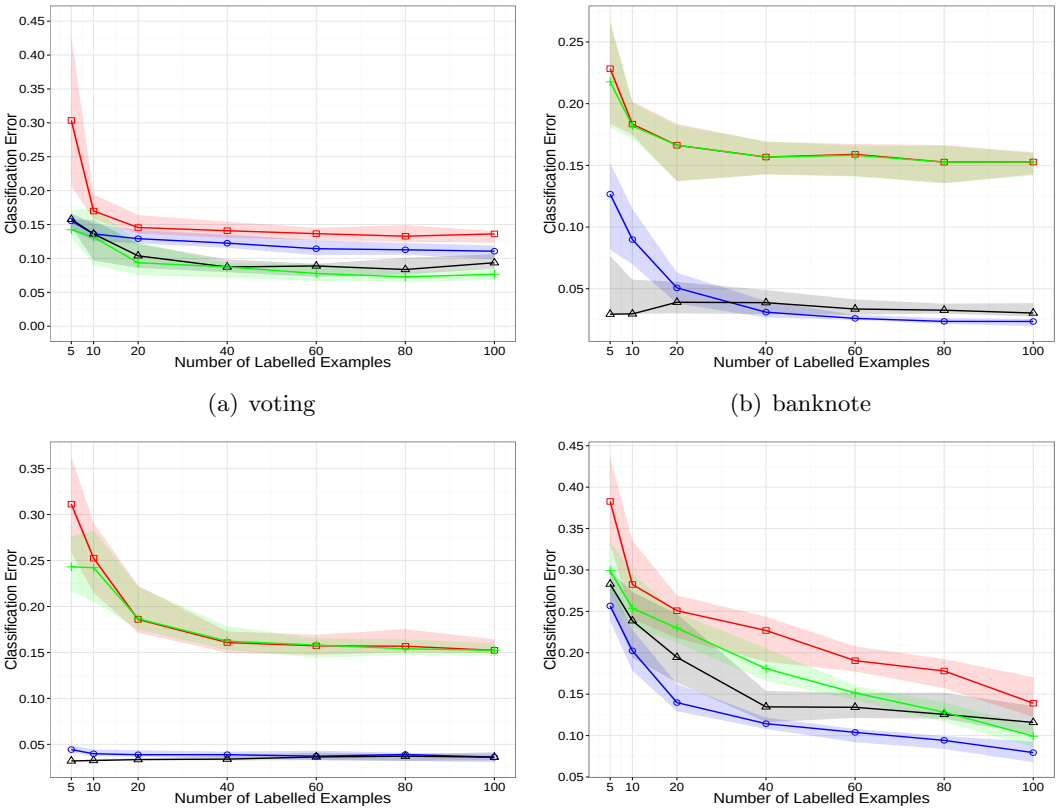
To assess the effect on performance of the number of labelled examples, ℓ, we consider a range of values. We compare the methods using the subset of data sets used in the previous section in which the size of the smallest class exceeds 100. In total eight data sets are used. For each value of ℓ, 30 random partitions into labelled and unlabelled data are considered. As classes are balanced in the data sets considered, performance is measured only in terms of classification error on the unlabelled data. For data sets with more than two classes all pairwise combinations of classes are considered and aggregate performance is reported.

Figure 7 provides plots of the median and interquartile range of the classification error for values of ℓ between 5 and 100 for the four data sets with two classes. Overall MDP^2 appears to be most competitive when the number of labelled examples is small. In addition, MDP^2 is comparable with the best performing method in almost every case. The only exception is the ionosphere data set where LapSVM outperforms MDP^2 for all values of ℓ. Figure 8 provides plots of the median and interquartile range of the aggregate classification error on data sets containing more than two classes. As these data sets are larger we consider up to 300 labelled examples. Note that the interquartile range for XNV is not depicted for the satellite data set. The variability of performance of XNV on this data set was so high that including the interquartile range would obscure all other information in the figure. MDP^2 exhibits the best performance overall, and obtains the lowest median classification error, or tied lowest, for all data sets and values of ℓ.

5.3 Summary of Experimental Results

We evaluated the performance of the MDP^2 formulation for finding MDHs for both clustering and semi-supervised classification, on a large collection of benchmark data sets, and in comparison with state-of-the-art methods for both problems.

For clustering, we found that no single method was consistently superior to all others. This is a result of the vastly differing nature of the data sets in terms of size, dimensionality, number and shape of clusters, etc. MDP^2 achieved the best performance on more data sets than any of the competing methods, and importantly was competitive with the best performing method in almost every data set considered. All other methods performed poorly in at least as many examples. Boxplots of both the raw performance and performance regret, which measures the difference between each method and the best performing method on each

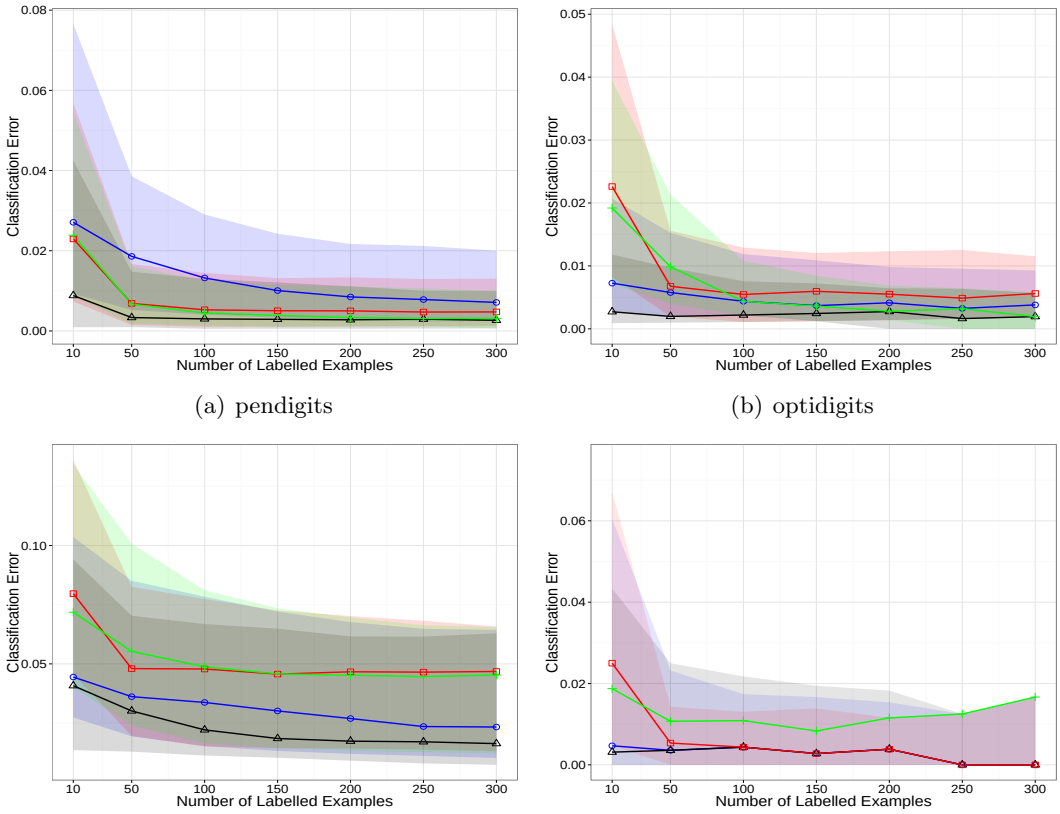


(a) voting (b) banknote

(c) breast cancer (d) ionosphere

MDP^2 median (—△—), LapSVM median (—○—), SSSL median (—□—), XNV median (—+—), with corresponding interquartile ranges given by shaded regions.

Figure 7: Classification error for different number of labelled examples for data sets with two clusters.



(a) pendigits

(b) optidigits

(c) satellite

(d) image segmentation

MDP^2 median (—△—), LapSVM median (—○—), SSSL median (—□—), XNV median (—+—), with corresponding interquartile ranges given by shaded regions.

Figure 8: Classification error for different numbers of labelled examples over all pairwise combinations of classes.

data set, allowed us to summarise the comparative performance of the different methods across data sets. The mean and median raw performance of MDP^2 is substantially higher than the next best performing method, and the regret is also substantially lower.

In the case of semi-supervised classification it was apparent that MDP^2 is extremely competitive when the number of labelled examples is (very) small, but that in some cases its performance does not improve as much as that of the other methods considered, when the labelled examples become more abundant. Our experiments suggest that overall MDP^2 is very competitive with the state-of-the-art for semi-supervised classification problems.

6. Conclusions

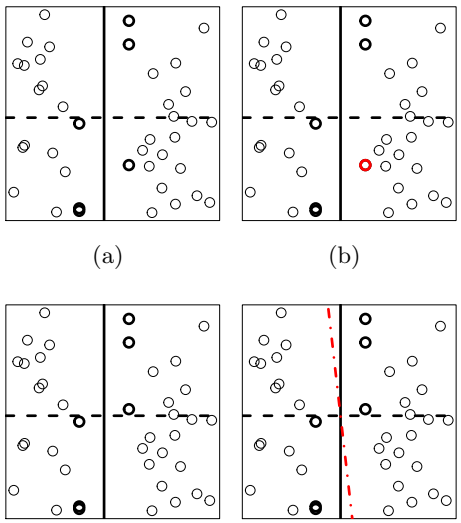
We proposed a new hyperplane classifier for clustering and semi-supervised classification. The proposed approach is motivated by determining low density linear separators of the high-density clusters within a data set. This is achieved by minimising the integral of the empirical density along the hyperplane, which is computed through kernel density estimation. To the best of our knowledge this is the first direct implementation of the low density separation assumption that underlies high-density clustering and numerous influential semi-supervised classification methods. We show that the minimum density hyperplane is asymptotically connected with maximum margin hyperplane, thereby establishing an important link between the proposed approach, maximum margin clustering, and semi-supervised support vector machines.

The proposed formulation allows us to evaluate the integral of the density on a hyperplane by projecting the data onto the vector normal to the hyperplane, and estimating a univariate kernel density estimator. This enables us to apply our method effectively and efficiently on data sets of much higher dimensionality than is generally possible for density based clustering methods. To mitigate the problem of convergence to locally optimal solutions we proposed a projection pursuit formulation.

We evaluated the minimum density hyperplane approach on a large collection of benchmark data sets. The experimental results obtained indicate that the method is competitive with state-of-the-art methods for clustering and semi-supervised classification. Importantly the performance of the proposed approach displays low variability across a variety of data sets, and is robust to differences in data size, dimensionality, and number of clusters. In the context of semi-supervised classification, the proposed approach shows especially good performance when the number of labelled data is small.

Acknowledgments

We would like to thank the reviewers for their insightful comments which substantially improved this paper. We also thank Prof. David Leslie, and Dr. Teemu Roos for valuable comments and suggestions on this work. Nicos Pavlidis would like to thank the Isaac Newton Institute for Mathematical Sciences, Cambridge, for support and hospitality during the programme 'Inference for Change-Point and Related Processes', where part of the work on this paper was undertaken. David Hofmeyr gratefully acknowledges the support of the EPSRC funded EP/H023151/1 STOR-i centre for doctoral training, as well as the Oppenheimer



Proposed MMH ——, Orthogonal hyperplane - - -, Hyperplane with larger margin - · - · -, Regular points ◯, Support points ◯, Differently assigned support point ◯

Figure 9: Two dimensional illustration of Lemma 9

Memorial Trust. The underlying code and data are openly available from Lancaster University data repository at `http://dx.doi.org/10.17635/lancaster/researchdata/97`.

Appendix A. Proof of Theorem 5

Before proving Theorem 5 we require the following two technical lemmata which establish some algebraic properties of the maximum margin hyperplane. The following lemma shows that any hyperplane orthogonal to the maximum margin hyperplane results in a different partition of the support points of the maximum margin hyperplane. The proof relies on the fact that if this statement does not hold then a hyperplane with larger margin exists which is a contradiction. Figure 9 provides an illustration of why this result holds. (a) Any hyperplane orthogonal to MMH generates a different partition of the support points of MMH, e.g., the point highlighted in red in (b) is grouped with the lower three by the dotted line but with the upper two by the solid line, the MMH. If an orthogonal hyperplane *can* generate the same partition (c), then a larger margin hyperplane than the proposed MMH exists (d).

Lemma 9 *Suppose there is a unique hyperplane in F with maximum margin, which can be parameterised by $(\mathbf{v}^m, b^m) \in \mathcal{S}^{d-1} \times \mathbb{R}$. Let $M = \operatorname{margin} H(\mathbf{v}^m, b^m)$, $C^+ = \{\mathbf{x} \in \mathcal{X} \,|\, \mathbf{v}^m \cdot \mathbf{x} - b^m = M\}$ and $C^- = \{\mathbf{x} \in \mathcal{X} \,|\, b^m - \mathbf{v}^m \cdot \mathbf{x} = M\}$. Then, $\forall \mathbf{w} \in \operatorname{Null}(\mathbf{v}^m)$, $c \in \mathbb{R}$ either $\min\{\mathbf{w} \cdot \mathbf{x} - c \,|\, \mathbf{x} \in C^+\} \leqslant 0$, or $\max\{\mathbf{w} \cdot \mathbf{x} - c \,|\, \mathbf{x} \in C^-\} \geqslant 0$.*

Proof

Suppose the result does not hold, then $\exists(\mathbf{w}, c)$ with $\|\mathbf{w}\| = 1, \mathbf{w} \cdot \mathbf{v}^m = 0$ and $\min\{\mathbf{w} \cdot \mathbf{x} - c | \mathbf{x} \in C^+\} > 0$ and $\max\{\mathbf{w} \cdot \mathbf{x} - c | \mathbf{x} \in C^-\} < 0$. Let $m = \min\{|\mathbf{w} \cdot \mathbf{x} - c| \,\big|\, \mathbf{x} \in C^+ \cup C^-\}$. Define $\lambda = \frac{m}{2M} < 1$. Define $\mathbf{u} = \frac{1}{\sqrt{\lambda^2 + (1-\lambda)^2}}(\lambda \mathbf{w} + (1-\lambda)\mathbf{v}^m)$ and $d = \frac{\lambda c + (1-\lambda) b^m}{\sqrt{\lambda^2 + (1-\lambda)^2}}$. By construction $\|\mathbf{u}\| = 1$. For any $\mathbf{x}_+ \in C^+$ we have,

$$\begin{aligned}
\mathbf{u} \cdot \mathbf{x}_+ - d &= \frac{\lambda(\mathbf{w} \cdot \mathbf{x}_+ - c) + (1-\lambda)(\mathbf{v}^m \cdot \mathbf{x}_+ - b^m)}{\sqrt{\lambda^2 + (1-\lambda)^2}} \\
&\geqslant \frac{\lambda m + (1-\lambda) M}{\sqrt{\lambda^2 + (1-\lambda)^2}} \\
&= \frac{m^2 + 2M^2 - Mm}{\sqrt{m^2 + (2M - m)^2}} \\
&> M.
\end{aligned}$$

Similarly one can show that $d - \mathbf{u} \cdot \mathbf{x}_- > M$ for any $\mathbf{x}_- \in C^-$, meaning that $(\mathbf{u}, d)$ achieves a larger margin on C^+ and C^- than $(\mathbf{v}^m, b^m)$, a contradiction.

■

The next lemma uses the above result to provide an upper bound on the distance between pairs of support points projected onto any vector, in terms of the angle between that vector and $\mathbf{v}^m$.

Lemma 10 *Suppose there is a unique hyperplane in F with maximum margin, which can be parameterised by $(\mathbf{v}^m, b^m) \in \mathcal{S}^{d-1} \times \mathbb{R}$. Define $M = \operatorname{margin} H(\mathbf{v}^m, b^m)$, $C^+ = \{\mathbf{x} \in \mathcal{X} | \mathbf{v}^m \cdot \mathbf{x} - b^m = M\}$, and $C^- = \{\mathbf{x} \in \mathcal{X} | b^m - \mathbf{v}^m \cdot \mathbf{x} = M\}$. There is no vector $\mathbf{w} \in \mathbb{R}^d$ for which $\mathbf{w} \cdot \mathbf{x}_+ - \mathbf{w} \cdot \mathbf{x}_- > 2M \mathbf{v}^m \cdot \mathbf{w}$ for all pairs $\mathbf{x}_+ \in C^+, \mathbf{x}_- \in C^-$.*

Proof

Suppose such a vector exists. Define $\mathbf{w}' = \mathbf{w} - (\mathbf{v}^m \cdot \mathbf{w})\mathbf{v}^m$. By construction $\mathbf{w}' \in \text{Null}(\mathbf{v}^m)$. For any pair $\mathbf{x}_+ \in C^+, \mathbf{x}_- \in C^-$ we have

$$\begin{aligned}
\mathbf{w}' \cdot x_+ - \mathbf{w}' \cdot x_- &= \mathbf{w} \cdot x_+ - (\mathbf{v}^m \cdot \mathbf{w})\mathbf{v}^m \cdot x_+ - \mathbf{w} \cdot x_- + (\mathbf{v}^m \cdot \mathbf{w})\mathbf{v}^m \cdot x_- \\
&> \mathbf{v}^m \cdot \mathbf{w}(2M - \mathbf{v}^m \cdot x_+ + b^m - b^m + \mathbf{v}^m \cdot x_-) \\
&= 0.
\end{aligned}$$

Define $c := \frac{1}{2}(\min\{\mathbf{w}' \cdot \mathbf{x}_+ \big| \mathbf{x}_+ \in C^+\} + \max\{\mathbf{w}' \cdot \mathbf{x}_- \big| \mathbf{x}_- \in C^-\})$. Then $\min\{\mathbf{w}' \cdot \mathbf{x}_+ - c | \mathbf{x}_+ \in C^+\} > 0$ and $\max\{\mathbf{w}' \cdot \mathbf{x}_- - c | \mathbf{x}_- \in C^-\} < 0$, a contradiction.

■

We are now in a position to provide the main proof of this appendix. The theorem states that if the maximum margin hyperplane is unique, and can be parameterised by $(\mathbf{v}^m, b^m) \in \mathcal{S}^{d-1} \times \mathbb{R}$, then

$$\lim_{h \to 0^+} \min\{\|(\mathbf{v}_h^\star, b_h^\star) - (\mathbf{v}^m, b^m)\|, \|(\mathbf{v}_h^\star, b_h^\star) + (\mathbf{v}^m, b^m)\|\} = 0,$$

where $\{H(v_h^\star, b_h^\star)\}_h$ is any collection of minimum density hyperplanes indexed by their bandwidth $h > 0$.

Proof of Theorem 5

Define $M = \operatorname{margin} H(\mathbf{v}^m, b^m)$, $C^+ = \{\mathbf{x} \in \mathcal{X} \,|\, \mathbf{v}^m \cdot \mathbf{x} - b^m = M\}$ and $C^- = \{\mathbf{x} \in \mathcal{X} \,|\, b^m - \mathbf{v}^m \cdot \mathbf{x} = M\}$. Let $B = \max\{\|\mathbf{x}\| \,\big|\, \mathbf{x} \in \mathcal{X}\}$. Take any $\epsilon > 0$ and set $0 < \delta$ to satisfy $\frac{2\delta}{M}(1 + B^2) + 2B\delta^{3/2}\sqrt{\frac{2}{M}} + \delta^2 = \epsilon^2$. Now, suppose $(\mathbf{w}, c) \in \mathcal{S}^{d-1} \times \mathbb{R}$ satisfies,

$$\mathbf{w} \cdot \mathbf{x}_+ - c > M - \delta, \ \forall \mathbf{x}_+ \in C^+ \text{ and } c - \mathbf{w} \cdot \mathbf{x}_- > M - \delta, \ \forall \mathbf{x}_- \in C^-.$$

By Lemma 10 we know that $\exists \mathbf{x}_+ \in C^+, \mathbf{x}_- \in C^-$ s.t. $\mathbf{w} \cdot \mathbf{x}_+ - \mathbf{w} \cdot \mathbf{x}_- \leqslant 2M\mathbf{v}^m \cdot \mathbf{w}$. Thus

$$\begin{aligned}
\mathbf{v}^m \cdot \mathbf{w} &\geqslant \frac{\mathbf{w} \cdot \mathbf{x}_+ - \mathbf{w} \cdot \mathbf{x}_-}{2M} \\
&= \frac{\mathbf{w} \cdot \mathbf{x}_+ - c + c - \mathbf{w} \cdot \mathbf{x}_-}{2M} \\
&> \frac{2M - 2\delta}{2M} = 1 - \frac{\delta}{M}.
\end{aligned}$$

Thus $\|\mathbf{v}^m - \mathbf{w}\|^2 < \frac{2\delta}{M}$. Now, for each $\mathbf{x}_+ \in C^+, \mathbf{v}^m \cdot \mathbf{x}_+ - b = M$ and for each $\mathbf{x}_- \in C^-, b - \mathbf{v}^m \cdot \mathbf{x}_- = M$. Thus for any such $\mathbf{x}_+, \mathbf{x}_-$ we have,

$$\begin{gathered}
M - \delta + \mathbf{w} \cdot \mathbf{x}_- < c < \mathbf{w} \cdot \mathbf{x}_+ - M + \delta, \\
b^m - \mathbf{v}^m \cdot \mathbf{x}_- - \delta + \mathbf{w} \cdot \mathbf{x}_- < c < \mathbf{w} \cdot \mathbf{x}_+ - \mathbf{v}^m \cdot \mathbf{x}_+ + b^m + \delta, \\
b^m - \delta - (\mathbf{v}^m - \mathbf{w}) \cdot \mathbf{x}_- < c < b^m + \delta + (\mathbf{w} - \mathbf{v}^m) \cdot \mathbf{x}_+, \\
b^m - \delta - B\|\mathbf{v}^m - \mathbf{w}\| < c < b^m + \delta + B\|\mathbf{w} - \mathbf{v}^m\|, \\
|c - b^m| < |\delta + B\|\mathbf{w} - \mathbf{v}^m\|| \,.
\end{gathered}$$

We can now bound the distance between $(\mathbf{w}, c)$ and $(\mathbf{v}^m, b^m)$,

$$\begin{aligned}
\|(\mathbf{v}^m, b^m) - (\mathbf{w}, c)\|^2 &= \|\mathbf{v}^m - \mathbf{w}\|^2 + |b^m - c|^2 \\
&< \|\mathbf{v}^m - \mathbf{w}\|^2(1 + B^2) + 2B\delta\|\mathbf{v}^m - \mathbf{w}\| + \delta^2 \\
&< \frac{2\delta}{M}(1 + B^2) + 2B\delta\sqrt{\frac{2\delta}{M}} + \delta^2 \\
&= \epsilon^2.
\end{aligned}$$

We have shown that for any hyperplane $H(\mathbf{w}, c)$ that achieves a margin larger than $M - \delta$ on the support points of the maximum margin hyperplane, $\mathbf{x} \in C^+ \cup C^-$, the distance between $(\mathbf{w}, c)$ and $(\mathbf{v}^m, b^m)$ is less than ϵ. Equivalently, any hyperplane $H(\mathbf{w}, c)$ such that $\|(\mathbf{w}, c) - (\mathbf{v}^m, b^m)\| > \epsilon$ has a margin less than $M - \delta$, as $\min\left\{|\mathbf{w} \cdot \mathbf{x} - c| \,\big|\, \mathbf{x} \in C^+ \cup C_-\right\} < M - \delta$. By symmetry, the same holds for any $(\mathbf{w}, c)$ within distance ϵ of $(-\mathbf{v}^m, -b^m)$.

By Lemma 4 $\exists h_1 > 0$ such that for all $h \in (0, h_1)$, the minimum density hyperplane for h, $H(\mathbf{v}_h^\star, b_h^\star)$, induces the same partition of $\mathcal{X}$ as the maximum margin hyperplane, $H(\mathbf{v}^m, b^m)$. By Lemma 3 $\exists h_2 > 0$ such that for all $h \in (0, h_2)$, $\operatorname{margin} H(\mathbf{v}_h^\star, b_h^\star) > M - \delta$. Therefore for $h \in (0, \min\{h_1, h_2\})$, $\min\left\{\|(\mathbf{v}_h^\star, b_h^\star) - (\mathbf{v}^m, b^m)\|, \|(\mathbf{v}_h^\star, b_h^\star) + (\mathbf{v}^m, b^m)\|\right\} < \epsilon$. Since $\epsilon > 0$ was arbitrarily chosen, this gives the result.

■

References

E. Aitnouri, S. Wang, and D. Ziou. On comparison of clustering techniques for histogram pdf estimation. *Pattern Recognition and Image Analysis*, 10(2):206–217, 2000.

D. Arthur and S. Vassilvitskii. k-means++: The advantages of careful seeding. In *ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1027–1035, 2007.

A. Azzalini and N. Torelli. Clustering via nonparametric density estimation. *Statistics and Computing*, 17(1):71–80, 2007.

M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labelled and unlabeled examples. *Journal of Machine Learning Research*, 7:2399–2434, 2006.

S. Ben-David, T. Lu, D. Pál, and M. Sotáková. Learning low-density separators. In D. van Dyk and M. Welling, editors, *International Conference on Artificial Intelligence and Statistics (AISTATS)*, JMLR Workshop and Conference Proceedings, pages 25–32, 2009.

C. Berge. *Topological Spaces*. Macmillan, New York, 1963.

D. Boley. Principal direction divisive partitioning. *Data Mining and Knowledge Discovery*, 2(4):325–344, 1998.

J. F. Bonnans and A. Shapiro. *Perturbation Analysis of Optimization Problems*. Springer Series in Operations Research. Springer, 2000.

J. V. Burke, A. S. Lewis, and M. L. Overton. Approximating subdifferentials by random sampling of gradients. *Mathematics of Operations Research*, 27(3):567–584, 2002.

J. V. Burke, A. S. Lewis, and M. L. Overton. A robust gradient sampling algorithm for nonsmooth, nonconvex optimization. *SIAM Journal on Optimization*, 15(3):751–779, 2006.

V. Castelli and T. M. Cover. On the exponential value of labeled samples. *Pattern Recognition Letters*, 16(1):105–111, 1995.

V. Castelli and T. M. Cover. The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter. *IEEE Transactions on Information Theory*, 42(6):2102–2117, 1996.

O. Chapelle and A. Zien. Semi-supervised classification by low density separation. In R. G. Cowell and Z. Ghahramani, editors, *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 57–64, 2005.

A. Cuevas, M. Febrero, and R. Fraiman. Estimating the number of clusters. *Canadian Journal of Statistics*, 28(2):367–382, 2000.

A. Cuevas, M. Febrero, and R. Fraiman. Cluster analysis: a further approach based on density estimation. *Computational Statistics and Data Analysis*, 36(4):441–459, 2001.

F. E. Curtis and X. Que. An adaptive gradient sampling algorithm for nonsmooth optimization. *Optimization Methods and Software*, 28(6):1302–1324, 2013.

C. Ding and T. Li. Adaptive dimension reduction using discriminant analysis and k-means clustering. In *International Conference on Machine Learning (ICML)*, pages 521–528, 2007.

C. Fowlkes, S. Belongie, F. Chung, and J. Malik. Spectral grouping using the Nyström method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):214–225, 2004.

P. Fränti and O. Virmajoki. Iterative shrinking method for clustering problems. *Pattern Recognition*, 39(5):761–775, 2006.

J. A. Hartigan. *Clustering Algorithms*. Wiley Series in Probability and Mathematical Statistics. Wiley, New York, 1975.

M. Ji, T. Yang, B. Lin, R. Jin, and J. Han. A simple algorithm for semi-supervised learning with improved generalization error bound. In J. Langford and J. Pineau, editors, *International Conference on Machine Learning (ICML)*, pages 1223–1230, 2012.

T. Joachims. Transductive inference for text classification using support vector machines. In *International Conference on Machine Learning (ICML)*, volume 99, pages 200–209, 1999.

A. Lewis and M. Overton. Nonsmooth optimization via quasi-Newton methods. *Mathematical Programming*, 141:135–163, 2013.

M. Lichman. UCI machine learning repository, 2013. URL `http://archive.ics.uci.edu/ml`.

B. McWilliams, D. Balduzzi, and J. M. Buhmann. Correlated random features for fast semi-supervised learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 440–448, 2013.

G. Menardi and A. Azzalini. An advancement in clustering via nonparametric density estimation. *Statistics and Computing*, 24(5):753–767, 2014.

V. I. Morariu, B. V. Srinivasan, V. C. Raykar, R. Duraiswami, and L. S. Davis. Automatic online tuning for fast Gaussian summation. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1113–1120, 2008.

A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: analysis and an algorithm. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 14, pages 849–856, 2002.

E. Polak. On the mathematical foundations of nondifferentiable optimization in engineering design. *SIAM Review*, 29(1):21–89, 1987.

P. Rigollet. Generalization error bounds in semi-supervised classification under the cluster assumption. *Journal of Machine Learning Research*, 8:1369–1392, 2007.

A. Rinaldo and L. Wasserman. Generalized density clustering. *The Annals of Statistics*, 38 (5):2678–2722, 2010.

A. Rosenberg and J. Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In *Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, volume 7, pages 410–420, 2007.

B. W. Silverman. *Density estimation for statistics and data analysis*, volume 26. CRC press, 1986.

A. Singh, R. D. Nowak, and X. Zhu. Unlabeled data: Now it helps, now it doesn't. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems (NIPS)*, pages 1513–1520, 2009.

W. Stuetzle and R. Nugent. A generalized single linkage method for estimating the cluster tree of a density. *Journal of Computational and Graphical Statistics*, 19(2):397–418, 2010.

S. K. Tasoulis, D. K. Tasoulis, and V. P. Plagianakos. Enhancing principal direction divisive clustering. *Pattern Recognition*, 43(10):3391–3411, 2010.

S. Tong and D. Koller. Restricted Bayes optimal classifiers. In *National Conference on Artificial Intelligence (AAAI)*, pages 658–664, 2000.

V. Vapnik and A. Sterin. On structural risk minimization or overall risk in a problem of pattern recognition. *Automation and Remote Control*, 10(3):1495–1503, 1977.

G. Walther. Granulometric smoothing. *The Annals of Statistics*, 25(6):2273–2299, 1997.

P. Wolfe. On the convergence of gradient methods under constraint. *IBM Journal of Research and Development*, pages 407–411, 1972.

L. Xu, J. Neufeld, B. Larson, and D. Schuurmans. Maximum margin clustering. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 17, pages 1537–1544, 2004.

L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1601–1608, 2004.

K. Zhang, I. W. Tsang, and J. T. Kwok. Maximum margin clustering made practical. *IEEE Transactions on Neural Networks*, 20(4):583–596, 2009.