Multimodal Projection Pursuit using the Dip Statistic

Andreas Krause
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA, 15213, USA
krausea@cs.cmu.edu

Volkmar Liebscher
Institut für Mathematik und Informatik
Ernst-Moritz-Arndt-Universität Greifswald
D-17487 Greifswald, Germany
volkmar.liebscher@uni-greifswald.de

Abstract

Projection pursuit is the search for interesting low-dimensional projections of high-dimensional data. It optimizes projection indices, which increase with the interestingness of the projection image. Most classical approaches equate interestingness with non-gaussianity. However, in cluster analysis one should more be interested in departure from unimodality. The dip is an efficient nonparametric test measuring the distance of distributions from the class of unimodal distributions with respect to the maximum norm. In this paper, we demonstrate how the dip can be used in projection pursuit. We establish continuity and differentiability properties and develop efficient algorithms to search for projections maximizing the dip and extend them to find multiple interesting projections. Our algorithms are empirically evaluated on several surrogate and real-world data sets.

Keywords. Projection Index, Dip Test, Unimodality, Clustering, Bump Hunting, Nonparametric Tests

1 Introduction

Projection pursuit is the systematic search for interesting low-dimensional linear projections of high-dimensional data. Dimension reduction is an important problem in exploratory data analysis and visualization – the human ability for pattern recognition can only be exploited for very low dimensional data. Machine learning methods can also take advantage of this method since it provides a viable way to overcome the "curse of dimensionality" problem [Bellman, 1961]. Projection pursuit also enables non-parametric approaches towards regression and density estimation and provides a unifying framework for well-known techniques in multivariate analysis, such as principal component analysis (PCA) or discriminatory techniques such as linear discriminant analysis (LDA). An excellent introduction into theory and applications of projection pursuit can be found in [Huber, 1985].

Projection index	Mixture of Gaussians	Pareto distribution	Normal distribution	95 percent quantiles
Negative Entropy	-1.23	$\mathbf{-0.952}$	$-\log\sqrt{2\pi e} \approx -1.42$	-1.38
Friedman's Index	0.0511	$\mathbf{0.326}$	0	0.017
Dip	$\mathbf{0.0516}$	0.0084	0	0.016

Table 1: Projection indices for samples from distributions. The sample further from normality is emphasized in bold.

Projection pursuit seeks to optimize functionals – projection indices – over projections of the data distribution. These indices are designed to respond to *interesting* distributions. Of course, the definition of interestingness is strongly application-specific. In previous investigations of projection pursuit, interestingness has mostly been equated with non-normality [Huber, 1985, Friedman, 1987]. In the case of PCA, projections are sought which maximize the variance within their image, and in discriminant analysis, projections are selected which achieve an optimal separation of classified sample data. In his dissertation Nason [1992] discusses the design and analysis of projection indices.

In the context of cluster analysis, one is interested in finding low-dimensional projections which preserve the cluster structure in the projected data. For one-dimensional data, unimodal distributions per definition have no such structure, thus the farther a distribution is from unimodality, the more likely it is to bear clusters. Departure from unimodality implies departure from normality, but since the converse is not true in general, it is appropriate for cluster analysis to define interestingness by multimodality.

To motivate the necessity of a projection index focused on multimodality consider the following example: We generated two datasets of 500 samples each, one from a Pareto distribution ($a = 2$) and another from a mixture of two unit-variance Gaussians with equal weight, centered at -2 and +2. Both data sets were standardized to have unit variance and the results are depicted as violin plots in Figure 1. A commonly used test for normality is the differential entropy of a (standardized) distribution, since the normal distribution is the unique maximizer of this functional among all distributions of unit variance and zero mean. Interestingly, the entropy of the mixture of Gaussians is much closer to the entropy of a standard normal distribution than the entropy of the unimodal Pareto distribution. Friedman's index [Friedman, 1987], which was specifically designed to measure departure from normality in the body of the distribution as discussed in Section 2, is also much lower for the mixture of Gaussians than for the Pareto distribution. The *dip* test for unimodality [Hartigan & Hartigan, 1985] however significantly rejects the unimodality hypothesis for the mixture of Gaussians and correctly classifies the sample from the Pareto distribution as unimodal. Table 1 presents the values of the projection indices along with the empirically determined 95 percent quantiles for rejecting the non-normality or non-unimodality hypotheses.

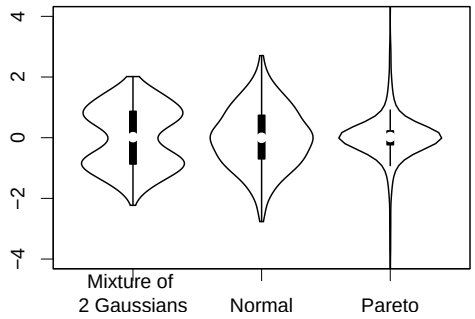


Figure 1: Unimodal distributions can be farther from normality than multimodal distributions.

In this paper, we propose to use the dip as a projection index in projection pursuit. We present an efficient algorithm to search for projections which maximize the dip. The algorithm is based on a gradient ascent scheme, for which we establish necessary differentiability properties of the dip in Section 3. We then demonstrate, how an efficient search algorithm can exploit these properties to find interesting projections (Section 4). In Section 5 we extend the procedure to find higher dimensional projections by discussing two approaches: The first method is based on an iterative orthogonal search, where one fixes $k-1$ orthogonal directions already found and optimizes the selection of the k-th orthogonal direction in the remaining subspace. The second method removes the interesting structure among each interesting direction, resulting in a recursive procedure. In case of the dip, this means making the distribution unimodal along these directions. To demonstrate the effectiveness of our method, we evaluate our algorithms on surrogate and real-world data. Our experiments indicate that the dip is a highly robust projection index, successfully identifying interesting directions, even in very high dimensional spaces, with a minimum of data preprocessing.

2 Projection Indices

This section presents a brief overview on common projection indices, and how the dip relates to them. We argue that projection indices can be classified into *parametric* indices measuring departure from a particular parametric family of distributions, and *nonparametric* indices targeted at detecting multimodality.

A *k-dimensional linear projection* from $\mathbb{R}^d$ to $\mathbb{R}^k$ is a linear mapping Ψ represented by a $k \times d$ matrix A of rank k. Ψ is an *orthogonal projection* if $A \cdot A^T = I_k$, where I_k is the $k \times k$ unit matrix. In projection pursuit, in most cases very low-dimensional projections are sought, where low usually means 1, 2 or 3. A *projection index* is a functional $\varphi : \Omega \to \mathbb{R}$ where $\Omega \subseteq \mathfrak{D}(\mathbb{R}^k)$ is a subset of the set of distribution functions on $\mathbb{R}^k$. For one dimensional projections, we will occasionally use the term *direction*. Furthermore when we speak about properties of projections, we will in general refer to the distribution of the projected data (e.g. “non-normal projections”).

Parametric statistical tests assume the data is sampled from a known family of distributions, of which the members can be selected by providing a finite set of parameters. On the contrary, non-parametric tests assume the data follows an arbitrary distribution. Although they are in general more difficult to handle analytically, they often appear to work in cases where the parametric approach fails (c.f. Cox et.al. [Cox & Hinkley, 1974]). Since projection indices can be considered as statistical tests for interestingness, these tests can be designed both as parametric and as non-parametric.

Parametric indices measuring departure from a distribution Two common tests for normality are *Shannon's negative entropy* and the *Fisher information*. Both indices are affine invariant, and have the normal distribution as unique minimizer, are differentiable with respect to the projection parameters and can be efficiently computed. Details can be found in Huber [Huber, 1985].

Since projection indices are repeatedly evaluated in iterative procedures, $\Theta(N)$ operations can be expensive for very large data sets. Using summary statistics, the computational effort can be reduced to a constant depending only on the dimension of the data. Jones et.al. [Jones & Sibson, 1987] propose a method based on the unbiased estimators of the third and fourth outer product moment tensors. They develop their index for one and two dimensions, and Nason [Nason, 1992] presents a generalization to three dimensions. The time complexity per iteration for these methods is $\Theta(d^4)$, hence for applications where d is low and N is high, this index might be computationally more efficient than the $\Theta(N)$ alternatives. Unfortunately, the cumulant approach is highly sensitive to outliers and heavy tails of the distributions [Friedman, 1987].

Friedman [Friedman, 1987] devised a projection index with the goal of computational efficiency and reduced sensitivity for heavy tails and outliers. To address robustness problems of moment methods, Friedman proposes to first transform the projected data by $R = 2\Phi(X) - 1$ where Φ is the distribution function of the standard normal, after which R will be distributed uniformly on $[-1, 1]$. To measure departure from normality, Friedman computes the L_2 distance of R to the uniform on $[-1, 1]$. He argues that by this transformation, central departure from normality is emphasized relatively to tail departure. To allow efficient computation, Friedman uses a moment approximation, expanding the index in terms of Legendre polynomials. For optimization, Friedman proposes a combination of global search and gradient ascent. Since Friedman requires spheredness of the data, he enforces the constraint $\|a\|_2 = 1$ by using Lagrange multipliers. A rotationally invariant two-dimensional version has been proposed by Morton [Morton, 1989].

Motivated by the observation that most non-normal projection indices tend to be sensitive to outliers and favor distributions with heavy tails, Nason [Nason, 1992] proposed a projection index measuring the departure from a Student-t distribution with n degrees of freedom.

Non-parametric indices for measuring multimodality For an e.c.d.f. F_n, the *depth* of F_n is defined as $\text{depth}(F_n) = \sup\left(\min\{F_n(x_6) - F_n(x_5), F_n(x_2) - F_n(x_1)\} - (F_n(x_4) - F_n(x_3))\right)$ where the $\sup$ is

taken over all points $x_1 \leq x_2 \leq x_3 \leq x_4 \leq x_5 \leq x_6$ such that $x_4 - x_3 \geq \max(x_2 - x_1, x_6 - x_5)$ (*c.f.* [Hartigan, 1977]). This functional identifies three intervals such that the middle interval is the largest and allocates relatively little mass compared to the outer intervals. Hartigan [Hartigan & Hartigan, 1985] argues that the depth, considered as a statistical test for unimodality, can in general be expected to have less power than the dip. Furthermore, it is of computational complexity $\Theta(n^2)$ to compute the depth for empirical distribution functions of n points.

Wegman [Wegman, 1970] defines the *likelihood ratio* $L(F_n)$ for an e.c.d.f. F_n as $L(F_n) = \left(\sup_{f \in \mathfrak{U}_{1,C}} \sum_{i=1}^{n} \log f(x_i)\right) / \left(\sup_{f \in \mathfrak{U}_{2,C}} \sum_{i=1}^{n} \log f(x_i)\right)$ where $\mathfrak{U}_{1,C}$ and $\mathfrak{U}_{2,C}$ respectively denote the classes of one-dimensional unimodal and bimodal densities, constrained to be bounded by C. The constraint on the densities is necessary, because $\sup_{f \in \mathfrak{U}} \sum_{k=1}^{n} \log f(x_i) = \infty$ where $\mathfrak{U}$ is the class of unimodal densities, independent from the selection of x_i. It is always possible to allocate the mass of the unimodals up to ε between two arbitrary data points. Hartigan [Hartigan & Hartigan, 1985] argues that the likelihood ratio is less robust than the dip, and critically depends on the choice of the constraint C on the densities.

Silverman [Silverman, 1981] constructs a test for multimodality, called *k-critical windows*, based on the following idea: A sample from a density with more than k modes will require more smoothing to exhibit k or less modes in the density estimate than a sample from a density with exactly k modes. Nason [Nason, 1992] argues that this test would not be an appropriate projection index since the analytical properties, such as continuity of the index with respect to the data are not clear, and furthermore it would be computationally expensive to compute. Additionally, it is not clear how k should be chosen, since the number of clusters in the data will in general not be known in advance.

In designing their statistical test for multimodality, Müller et.al. [Müller & Sawitzki, 1991] note that the analytical definition of a mode – being a local maximum – does not necessarily coincide with the statistical idea of a "high probability point". They define, study and estimate the *excess mass functional* of the data density. Nason [Nason, 1992] argues that the use of this test in the context of projection pursuit is again problematic due to the high computational complexity and the necessity for specifying an assumed number of modes k. The advantage of this test is however that it can be defined in an arbitrary number of dimensions.

Nason et.al. [Nason & Sibson, 1992] use a similar idea in constructing a projection index for measuring multimodality. Their design is based on the number of the connected components of the function graph, weighted by their respective volume. They argue that their index would work well in the context of projection pursuit, since it is rotation- and scale-invariant and responds weakly to outliers and strongly to clustering structure. However, the computation of this index is again computationally complex. The authors suggest to use a finite element approximation to the density and identify the number of connected components on the approximation grid. In addition to the computational complexity, their index again depends on

the selection of a kernel width.

The *dip* [Hartigan & Hartigan, 1985] measures the departure of arbitrary distribution functions from the class of unimodal distribution functions with respect to the supremum norm. It is efficiently computable, robust against outliers and location and scale-invariant. This index is the main focus of this paper and will be explored in detail in the subsequent sections.

3 The dip statistic

The dip is a statistical test for unimodality [Hartigan & Hartigan, 1985], defined as the distance of empirical distribution functions to the class of unimodal distribution functions with respect to the maximum norm. Hartigan proposed an algorithm to compute the dip in $\mathcal{O}(n)$ operations on sorted data. Since the gradient computations which we discuss in the following depend on quantities computed by the algorithm, it is presented as Algorithm 1 for reference in Section 3.1.

Definition 3.1. A distribution function F over $\mathbb{R}$ is called *unimodal* with *mode* m (which is not necessarily unique) if F is convex in $(-\infty, m]$ and concave in $[m, \infty)$.

A unimodal distribution function F has a density f everywhere but at most one point. f is monotonically non-decreasing in $(-\infty, m)$ and monotonically non-increasing in (m, ∞), where m is the mode of F. If F does not have a density in m', then m' must be the unique mode of F. If the mode is not unique, the set of modes forms a closed interval $[m_l, m_u]$, the *modal interval* of F. In this case, the density is constant within the modal interval and thus the distribution function restricted to $[m_l, m_u]$ is a line segment. Let $\mathfrak{D}(\mathbb{R}^k)$ be the class of distribution functions on $\mathbb{R}^k$, and we will use $\mathfrak{D}$ to represent $\mathfrak{D}(\mathbb{R})$. We will use $\mathfrak{B}(\mathbb{R}^k)$ to denote the class of bounded functions on $\mathbb{R}^k$, abbreviating $\mathfrak{B}(\mathbb{R})$ with $\mathfrak{B}$. $\mathfrak{D}$ is a complete metric space with the metric $\rho(F, G) := \|F - G\|_\infty := \sup_{x \in \mathbb{R}} |F(x) - G(x)|$ which is sometimes referred to as *Kolmogorov distance*.

Definition 3.2. Denote by $\mathfrak{U} \subset \mathfrak{D}$ the class of unimodal distribution functions over $\mathbb{R}$. The *dip* of a distribution function F is defined to be $\mathrm{D}(F) = \rho(F, \mathfrak{U})$, where for a subset $\mathcal{C} \subset \mathfrak{D}$ we define $\rho(F, \mathcal{C}) := \inf_{G \in \mathcal{C}} \rho(F, G)$.

The closedness of $\mathfrak{U}$ with respect to ρ immediately implies that the dip measures departure from unimodality, i.e. $\mathrm{D}(F) = 0 \Leftrightarrow F \in \mathfrak{U}$. Furthermore, the observation that $\mathrm{D}(F_1) \leq \mathrm{D}(F_2) + \rho(F_1, F_2)$ immediately proves that the mapping $\mathrm{D} : \mathfrak{D} \to [0, \infty)$ is continuous with respect to the uniform topology induced by ρ.

In fact, the dip is even Skorohod-continuous [Billingsley, 1968], a result which was established in [Krause, 2004]:

Theorem 3.3 ([Krause, 2004])**.** *The mapping $F \mapsto D(F)$ is continuous with respect to the Skorohod topology on $\mathfrak{S}$.*

In the context of projection pursuit, we only need continuity for empirical distribution functions. More specifically, in Corollary 3.18 we will show that the dip is continuous a.e. with respect to the locations of the masses. This result will be used later in order to show that the dip is also continuous a.e. with respect to linear projections of the data, an important property of a projection index.

3.1 Computing the dip

Since we are interested to find projections of high-dimensional data which maximize the dip, we need to be able to compute it at least for empirical distribution functions. In [Hartigan & Hartigan, 1985] a geometric characterization of the elements of best approximation is given, which directly leads to an efficient algorithm for computing the dip. The *greatest convex minorant* (g.c.m.) of $F \in \mathfrak{B}$ *in* $(-\infty, a]$ is $x \mapsto \sup\{G(x), G \in \mathfrak{B} \text{ convex in } (-\infty, a], G \leq F\}$. The *least concave majorant* (l.c.m.) of $F \in \mathfrak{B}$ *in* $[a, \infty)$ is $x \mapsto \inf\{L(x), L \in \mathfrak{B} \text{ concave in } [a, \infty), G \geq F\}$. The following theorem by Hartigan uses these concepts to characterize the elements of best-approximation:

Theorem 3.4 ([Hartigan & Hartigan, 1985])**.** *Let $F \in \mathfrak{D}$ be a distribution function. Then $D(F) = d$ only if there exists a nondecreasing function G such that, for some $x_L \leq x_U$, (i) G is the greatest convex minorant of $F + d$ in $(-\infty, x_L)$, (ii) G has constant maximum slope in (x_L, x_U), (iii) G is the least concave majorant of $F - d$ in $[x_U, \infty)$ and (iv) $d = \sup_{x \notin (x_L, x_U)} |F(x) - G(x)| \geq \sup_{x \in (x_L, x_U)} |F(x) - G(x)|$.*

As noted in [Hartigan & Hartigan, 1985], Theorem 3.4 suggests the following approach to computing the dip: Let F be an empirical distribution function for the sorted samples $x_1, \ldots, x_n$. There are $n(n-1)/2$ possible candidates for the modal interval. Compute for each candidate $[x_i, x_j], i \leq j$ the greatest convex minorant of F in $(-\infty, x_i]$ and the least concave majorant of F in $[x_j, \infty)$ and let d_{ij} be the maximum distance of F_n to these computed curves. Then $2\mathrm{D}(F_n)$ is the minimum value of d_{ij} for all candidate modal intervals, for which the line segment from $[x_i, F(x_i) + \frac{1}{2}d_{ij}]$ to $[x_j, F(x_j) - \frac{1}{2}d_{ij}]$ lies in the tube of width d_{ij} centered around the graph of F over the interval $[x_i, x_j]$.

Since the minorant and majorant computations can be done in advance in $\mathcal{O}(n)$, this algorithm is clearly of order $\mathcal{O}(n^2)$. Key to efficiency is to reduce the number of candidate modal intervals. For lower endpoints x_i, only those x_j need to be considered, for which the least concave majorant of F in $[x_i, \infty)$ touches F. Hartigan presents the following algorithm:

Consider a taut string stretched between the points $(x_1, F(x_1) + d)$ and $(x_n, F(x_n) - d)$. Assume the graphs of $F(x) + d$ and $F(x) - d$ are solid and bound a tube of width $2d$. As d decreases, the string bends to form a convex minorant from x_1 to x_L and a concave majorant from x_U to x_n. Both x_L and

x_U move towards each other with d decreasing. D(F) is the minimum value of d such that any further decrease forces the stretched string out of its unimodal shape. This idea is formalized in Algorithm 1. A concrete implementation of order $\mathcal{O}(n)$ is given in [Hartigan, 1985]. To see that $\mathcal{O}(n)$ is possible, note that the supremum computations in lines 1, 2 and 5 need to examine each index only once during all iterations of the **while**-loop.

Algorithm 1: Computation of the dip statistic

Input: $x_1, \ldots, x_n$ sorted
Output: Dip D and modal interval $[x_L, x_U]$
begin
  $x_L \leftarrow x_1; x_U \leftarrow x_n; d \leftarrow 0;$
  while *TRUE* **do**
    compute the g.c.m. G and l.c.m. L for F in $[x_L, x_U]$;
    compute the points of contact $g_1, \ldots, g_k$ of G and $l_1 \ldots l_m$ of L with F;
1   $d_G \leftarrow \sup_i |G(g_i) - L(g_i)|;$
2   $d_L \leftarrow \sup_i |G(l_i) - L(l_i)|;$
    if $d_G > d_L$ **then**
3   assume the sup occurs at $l_j \le g_i \le l_{j+1}$; $x_L^0 \leftarrow g_i; x_U^0 \leftarrow l_{j+1};$
    else
4   assume the sup occurs at $g_i \le l_j \le g_{i+1}$; $x_L^0 \leftarrow g_i; x_U^0 \leftarrow l_j;$
    end
    $d \leftarrow \max(d_G, d_L); x_U \leftarrow x_U^0; x_L \leftarrow x_L^0;$
    if $d \le D$ **then**
      $D \leftarrow \frac{D}{2}$; **break**;
    else
5     $D = \max(D, \sup_{x_L \le x \le x_L^0} |G(x) - F(x)|, \sup_{x_U^0 \le x \le x_U} |L(x) - F(x)|);$
    end
  end
end

3.2 Differentiability of the dip

We already mentioned that the dip is a continuous functional on the space of empirical distribution functions. This subsection will investigate the differentiability of the dip for this function class.

Definition 3.5. Let $\mathfrak{F}$ denote the space of empirical distribution functions, i.e. those which can be represented as a finite sum $F = \frac{1}{n} \sum_{i=1}^{n} \chi_{[x_i, \infty)}$ of characteristic functions, where $n > 0$, $x_i \le x_j$ for $i < j$, $i, j \in \{1, \ldots, n\}$.

Definition 3.6. Let $I = \{(x_1, \ldots, x_n), x_1 \leq \cdots \leq x_n, n \in \mathbb{N}_0\}$. The mapping $\mathrm{j} : I \to \mathfrak{F}, (x_1, \ldots, x_n) \mapsto \frac{1}{n} \sum_{i=1}^{n} \chi_{[x_i, \infty)}$ is a mapping from I into $\mathfrak{F}$ with the convention $\alpha_0 = 0$. If $F = \mathrm{j}(\xi)$, then ξ is called a *natural representation* of F. If $\xi = (x_1, \ldots, x_n)$, we write $|\xi| = n$.

It can be seen that the mapping j is surjective, i.e. any $F \in \mathfrak{F}$ can be naturally represented. In the following we will require the locations of the masses to be distinct. If this requirement is violated, we cannot even expect continuity of the dip:

Example 3.7. *Let $\xi = (0, 0)$, and for $\varepsilon > 0$ let $\delta = (-\varepsilon, \varepsilon)$. Then $D(\mathrm{j}(\xi)) = 0$ and $D(\mathrm{j}(\xi + \delta)) = \frac{1}{4}$, no matter how small ε has been chosen.* □

Definition 3.8. Let $\xi \in I$, $\xi = (x_1, \ldots, x_n)$. The difference quotient $\mathrm{Q}_{\xi,i}$ of the dip at ξ with respect to x_i is $\mathrm{Q}_{\xi,i}(h) = \left(D(\mathrm{j}(x_1, \ldots, x_i + h, \ldots, x_n)) - D(\mathrm{j}(x_1, \ldots, x_n))\right)/h$. The dip is *partially differentiable with respect to x_i at ξ*, if the limit $\lim_{h \to 0} \mathrm{Q}_{\xi,i}(h)$ exists. If the limit is d, we write $\frac{\partial D}{\partial x_i}(F) = d$. If only the one-sided limits exist, we write $\frac{\partial D}{\partial_+ x_i}$ if the limit exists for $h \to 0+$ and $\frac{\partial D}{\partial_- x_i}$ if the limit exists for $h \to 0-$.

One approach towards proving local differentiability results is to analyze how Algorithm 1 computes the dip. The only place where the estimate of the dip is updated is line 5. From the fact that the distance between F and the g.c.m. or l.c.m. can be maximized only on the locations of the masses, and that the g.c.m. and l.c.m. have piecewise linear structure, it is clear, that the dip is computed as vertical distance between some x_i and a line segment between some x_j and x_k.

Definition 3.9. Let $F = \mathrm{j}(x_1, \ldots, x_n)$. The index i is called *active* if for every $\varepsilon > 0$ there exists a δ_x, $|\delta_x| < \varepsilon$ such that for $\hat{F} = \mathrm{j}(x_1, \ldots, x_i + \delta_x, \ldots, x_n, \alpha_n)$ it holds that $\mathrm{D}(F) \neq \mathrm{D}(\hat{F})$. The index i is called *strongly active* if there exists an $\varepsilon > 0$ such that for all δ_x, $|\delta_x| < \varepsilon$ it holds that for $\hat{F} = \mathrm{j}(x_1, \ldots, x_i + \delta_x, \ldots, x_n)$, $\mathrm{D}(F) \neq \mathrm{D}(\hat{F})$. The index i is called *inactive* if there exists an $\varepsilon > 0$ such that for all δ_x, $|\delta_x| < \varepsilon$ it holds that for $\hat{F} = \mathrm{j}(x_1, \ldots, x_i + \delta_x, \ldots, x_n)$, $\mathrm{D}(F) = \mathrm{D}(\hat{F})$.

It can be seen that the dip cannot be differentiable with respect to an index which is neither inactive nor strongly active. In the following, we will only consider natural representations of empirical distribution functions containing at least three indices. The definition of inactive indices immediately implies the following lemma:

Lemma 3.10. *Let $F = j(\xi)$, and i an inactive index of ξ. Then D is partially differentiable at F with respect to x_i and $\frac{\partial D}{\partial x_i}(F) = 0$.*

Definition 3.11. Let $F = \mathrm{j}(\xi)$. A *touching triangle* of ξ is a triple (i_1, i_2, i_3) of indices of ξ such that $i_1 \leq i_2 \leq i_3$, $i_1 < i_3$ and x_{i_1} and x_{i_3} are touching points of F and its greatest convex minorant or least concave majorant, on $(-\infty, x_{i_3}]$ or $[x_{i_1}, \infty)$ respectively, depending on whether $(x_{i_2}, \frac{i_2}{n})$ lies above (computed in Line 3 of Algorithm 1) or below (Line 4) the line segment between $(x_{i_1}, \frac{i_1}{n})$ and $(x_{i_3}, \frac{i_3}{n})$.

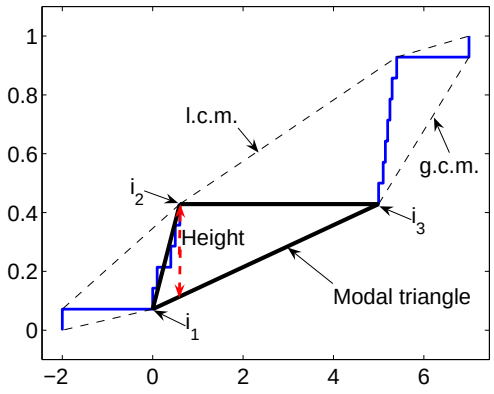


Figure 2: Height of the modal triangle

The case $i_1 = i_2$ or $i_2 = i_3$ is degenerate and might void some uniqueness statements valid for the non-degenerate case. This does not affect the following proofs in any way and will not be considered in the subsequent subsections for the sake of clarity.

Definition 3.12. Let $F = \mathrm{j}(\xi)$ and $\Delta = (i_1, i_2, i_3)$ be a touching triangle of ξ. Then the *height* of Δ is

$$h_\xi(\Delta) = \frac{1}{N}\left| i_2 - i_1 - \frac{x_{i_2} - x_{i_1}}{x_{i_3} - x_{i_1}}(i_3 - i_1)\right| + \frac{1}{N}. \tag{1}$$

If additionally $h_\xi(\Delta) = 2\mathrm{D}(F)$, then Δ is called *modal triangle*.

Figure 2 visualizes the concept of the height of a touching triangle.The next lemma establishes the existence of modal triangles for all empirical distribution functions with at least three indices. Its proof is based on the computation of the dip in Algorithm 1.

Lemma 3.13. *Let $F = j(\xi)$. Then i is an active index of ξ only if it is part of a modal triangle (i_1, i_2, i_3). Furthermore, ξ has at least one modal triangle.*

The following lemma establishes the relationship between touching and modal triangles. It essentially states that the modal triangle is the second highest touching triangle.

Lemma 3.14. *Let $F = j(\xi)$, and Θ denote the set of touching triangles of F. Then $D(F) = \frac{1}{2}\min_{\Delta\in\Theta}\left(\max_{\hat{\Delta}\in\Theta\setminus\{\Delta\}} h_\xi(\hat{\Delta})\right)$ with the convention that $\max(\emptyset) = 0$.*

The next lemma presents sufficient conditions for partial differentiability of the dip with respect to the location of the masses:

Lemma 3.15. *Let $F = j(\xi)$, $\Delta = (i_1, i_2, i_3)$ be the only modal triangle of ξ, and assume that the corner i_2 does not lie exactly on the line segment between i_1 and i_3. Then the dip is partially differentiable with*

respect to x_{i_2}. If in addition i_1 or i_3 are strongly active, then the dip is additionally differentiable with respect to x_{i_1} or x_{i_3} respectively.

Lemma 3.15 implies the following corollary about sufficient conditions for total differentiability:

Corollary 3.16. *Let $F = j(\xi)$, $\Delta = (i_1, i_2, i_3)$ be the only modal triangle of ξ, and assume that the corner i_2 does not lie exactly on the line segment between i_1 and i_3. Furthermore assume that i_1, i_2 and i_3 are strongly active. Then the dip is totally differentiable with respect to x_{i_1}, x_{i_2} and x_{i_3}.*

Even if the conditions of Lemma 3.15 are violated, we are guaranteed one-sided differentiability, which is the main result of our analysis:

Theorem 3.17. *The dip has one-sided partial derivatives, i.e. for $F = j(\xi)$ there exists an $M_\xi > 0$ such that for any index i of ξ it holds that $\left|\frac{\partial D}{\partial_\pm x_i}(F)\right| \leq M_\xi$. If $M : \xi \to [0, \infty)$ is the function assigning to ξ the infimum possible bound M_ξ, then M is locally bounded on I.*

Theorem 3.17 allows to prove the following Corollary on the continuity of the dip as a function of the locations of masses.

Corollary 3.18. *Let $\xi \in I$ and $\varepsilon > 0$. Then there exists a δ, $0 < \delta < \min\{x_{i+1} - x_i, i \in \{1, \ldots, n-1\}\}$ such that for every $\delta_\xi = (\delta_{x_1}, \ldots, \delta_{x_n})$ with $\|\delta_\xi\|_\infty < \delta$ it holds that $|D(j(\xi)) - D(j(\xi + \delta_\xi))| < \varepsilon$.*

Note that Corollary 3.18 only applies to empirical distribution functions where masses are not collocated. It is straight-forward to generalize the results to the case of varying weights, e.g. for reasons of robustness, but this is not done here for the sake of clarity. Details are presented in [Krause, 2004].

4 Multimodal Projection Pursuit

This section explains how the dip can be used to guide the search for multimodal projections. According to the terminology established in Section 2, we investigate the use of the dip as a projection index. In Section 4.1, an overview over the chosen approach is given and it is sketched how a local search can be performed to find local maxima of the dip. Differentiability results required for this approach are proved in Section 4.2. Details on algorithms are presented in Section 4.3, and Section 4.4 explains how multi-dimensional projections can be chosen based on the (one-dimensional) dip.

4.1 Approach

In Section 3.2 it was shown that for empirical distribution functions, the one-sided partial derivatives with respect to the natural representation exist. These results suggest that it is possible to increase the dip by appropriately moving the masses of the empirical data. For one-dimensional projections of multi-dimensional

data, the choice of the projection allows certain degrees of freedom regarding the positioning of the masses. The key notion in Section 3.2 was the concept of *modal triangles* as proposed in Definition 3.12. A modal triangle is a triple of sample indices directly influencing the dip statistic, whose value is determined by the height of the modal triangle. The height is a continuously differentiable function, although modifying its parameters might cause the triangle's modality to cease. In the context of linear projection pursuit, the height can also be considered a piecewise differentiable function of the *projection parameters*, keeping the empirical data fixed. This idea leads to a gradient ascent method explained in detail in Section 4.2.

4.2 The gradient of the dip

One-dimensional projections of d-dimensional data $\{x_1, \ldots, x_N\} \subset \mathbb{R}^d$ can be specified by a single vector $a \in \mathbb{R}^d$ and the standard scalar product $y_i = \langle a, x_i \rangle = a^T x_i$ for $1 \leq i \leq N$. For fixed data $X \in \mathbb{R}^{N \times d}$, its projection $a \mapsto Xa$ can be considered as a linear function in a. From Section 3.2 we know that for empirical distribution functions $F = \mathrm{j}(\xi)$ it holds that $2\mathrm{D}(F) = h_\xi(\Delta)$ where Δ is a modal triangle of ξ which is guaranteed to exist. The analysis in Section 3.2 required that the locations of the masses were distinct, otherwise the one-sided partial derivatives do not necessarily exist (*c.f.* Example 3.7).

Example 4.1. *Consider the data set X consisting of two points $x_1 = (1,0)^T$ and $x_2 = (2,0)^T$. Any projection onto a unit vector $a = (a_1, a_2)$ will map x_1 to a_1 and x_2 to $2a_1$. Thus if $a = (0,1)$, both points will be mapped onto the same image. Hence $D_X(a) = 0$ if a is orthogonal to $(1,0)^T$ and $D_X(a) = \frac{1}{4}$ otherwise.*

Thus, even if all data points in X are distinct, two data points can be projected onto the same image. In this case we cannot guarantee differentiability or even continuity of the dip with respect to the projection. Despite this negative result, the search technique presented below appears to perform well both on surrogate and empirical data.

Definition 4.2. Let $X \in \mathbb{R}^{N \times d}$ and denote by $x_1, \ldots, x_N$ the rows of X. A vector $a \in \mathbb{R}^d$ is called X-singular, if there exist $1 \leq i < j \leq N$ such that $x_i \neq x_j$ and $\langle a, x_i \rangle = \langle a, x_j \rangle$.

It is clear that 0 is X-singular for any non-trivial X. In the following we will always assume that the data points of X are all distinct.

Definition 4.3. Let $X \in \mathbb{R}^{N \times d}$ and $a \in \mathbb{R}^d$. Let $(y_1, \ldots, y_N)$ be the ordered sequence of the elements of $\psi_X(a)$ and let $\xi_X(a) := (y_1, \ldots, y_N)$ be the empirical distribution function of the projected data $\psi_X(a)$. Then denote by $\mathrm{D}_X(a) := \mathrm{D}(\mathrm{j}(\xi_X(a)))$ the dip of the projection of X with respect to a.

Utilizing the continuity established in Corollary 3.18, we arrive at the following continuity result:

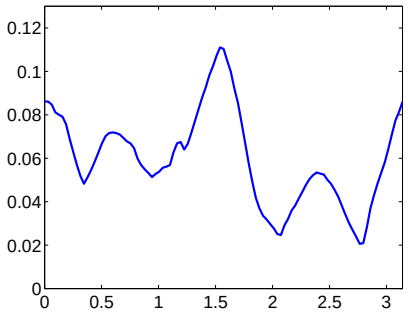


Figure 3: Dip elevation over a half-circle on the plane spanned by the two optimum dip axes from the sphered mixture of 20 Gaussians sample.

Theorem 4.4. *Let $X \in \mathbb{R}^{N\times d}$. Then the mapping $a \mapsto D_X(a)$ is continuous at all points of $\mathbb{R}^d$ which are not X-singular.* □

Theorem 4.4 immediately implies that the dip is continuous almost everywhere, if the projections are chosen randomly from any density on the unit sphere. Further continuity properties are described in [Krause, 2004].

To implement a gradient ascent method, we need to compute the partial derivatives of the dip with respect to the projection parameters. Let $\Delta = (i_1, i_2, i_3)$ be the modal triangle of $\xi_X(a)$. From Definition 3.12 we have that

$$h_{\xi_X(a)}(\Delta) = \frac{1}{N}\left| i_2 - i_1 - \frac{a^T\beta}{a^T\gamma}(i_3 - i_1)\right| + \frac{1}{N} \tag{2}$$

using the abbreviations $\beta = (\beta_1, \ldots, \beta_d) := x_{i_2} - x_{i_1}$ and $\gamma = (\gamma_1, \ldots, \gamma_d) = x_{i_3} - x_{i_1}$. After computation of the partial derivatives we arrive at the following result:

Theorem 4.5. *Let $X \in \mathbb{R}^{N\times d}$ and $a \in \mathbb{R}^d$ not X-singular such that $\xi_X(a)$ has exactly one modal triangle $\Delta = (i_1, i_2, i_3)$ such that i_1, i_2 and i_3 are strongly active. Then D_X is continuously differentiable in a with the partial derivatives*

$$\frac{\partial D_X}{\partial a_i}(a) = \begin{cases} -\frac{i_3 - i_1}{N} \cdot \frac{a_i a^T(\beta_i\gamma - \gamma_i\beta)}{(a^T\gamma)^2} & \text{if } \eta > 0 \\ \frac{i_3 - i_1}{N} \cdot \frac{a_i a^T(\beta_i\gamma - \gamma_i\beta)}{(a^T\gamma)^2} & \text{if } \eta < 0 \end{cases}$$

where $\eta = i_2 - i_1 - (i_3 - i_1)(a^T\beta)/(a^T\gamma)$ and β, γ as above. □

Note that the case $a^T\gamma = 0$ can only occur if a is X-singular. Figure 3 presents the change of the dip statistic for the projection axes rotating on the plane spanned by the first two coordinate axes of the Mixture of 20 Gaussians data set discussed in Section 6. This indicates that even for a highly clustered data set with $N = 985$ samples, the dip changes very smoothly with changing projection direction.

4.3 Search strategies

The differentiability considerations from Section 4.2 allow us to implement a gradient ascent algorithm. Such an approach will most likely not find global but rather local maxima, which is a common problem in multivariate analysis. However, there are various ways to avoid getting trapped in local maxima. The simplest possibility is to do restarts with random initial values. More sophisticated methods include simulated annealing and initialization with directions estimated from existing clustering results.

For the local search with random restarts, our algorithm picks the starting direction a uniformly at random from the sphere. It then computes the gradient ∇ of the dip, and performs a line search: It tries out all step sizes 1 through $2 \cdot 10^{-16}$, decreasing the step size by a factor γ of 2 per try, computing $a_\gamma := a + \gamma\nabla$, normalizing and choosing the step size a_γ maximizing the dip.

4.4 A note on efficient sorting

In the context of projection pursuit for large samples, the requirement of Algorithm 1 to receive ordered input in order to guarantee $\mathcal{O}(n)$ performance becomes a problem. It is a well known fact that comparison based sorting strategies have an $\Omega(n \log n)$ lower bound time complexity [Heun, 2000], which is asymptotically slower than the linear time dip computation. To overcome this problem and guarantee linear time computation of the dip, there are several possibilities. First it has to be noted that apparently the data only has to be sorted anew if the change in the projection parameters is so large, that the order of the projected data changes. Since linear projection is continuous, there is an upper bound for the parameter change below which sorting anew is not necessary.

Lemma 4.6. *Let $a, \delta \in \mathbb{R}^d$ and $x_i \in \mathbb{R}^d \backslash \{0\}$, $\varepsilon_i = \min\{|a^T x_j - a^T x_i|, j \in \{1, \ldots, N\} \backslash \{i\}\}$ for $1 \leq i \leq N$. If $\|\delta\|_2 < \frac{\varepsilon_i}{2\|x_i\|_2}$ for $1 \leq i \leq N$ and $(a^T x_1, \ldots, a^T x_N)$ is sorted, then $((a+\delta)^T x_1, \ldots, (a+\delta)^T x_N)$ is also sorted.* □

Lemma 4.6 provides a bound δ for which smaller changes do not disturb the sort order on the projected data. It is clear that δ can be computed in linear time once the data is initially sorted, since the ε_i are just the minimum absolute difference to the immediate neighbors of the $a^T x_i$, and since the norms $\|x_i\|_2$ can be pre-computed in $\mathcal{O}(d)$ time once and for all at the beginning of the projection pursuit.

Although this result can be used for optimizing the local search performance, apparently the data will still have to be sorted anew from time to time. A way to keep the linear time performance of the dip computation is to resort on using a non-comparison based sorting method. Commonly used algorithms for sorting numbers in real time are forward and backward radix sort as well as radix sort with groups [Heun, 2000]. These scale linearly with the number of samples and with the number of digits in the input data, which can normally be considered as constant.

5 Generalization to multiple dimensions

The literature on how to find several interesting one-dimensional projections [Huber, 1985] considers basically two approaches

1. *Iterative* methods fix the $k-1$ directions already found and then optimize among projections onto the k-dimensional space spanned by the fixed $k-1$ directions plus one additional direction. These methods only give a nested sequence of subspaces instead of an ordered list of directions. This technique is explained in Section 5.1.

2. *Recursive* methods find the most interesting direction, remove the interesting structure along this direction and iterate. Details on this technique are given in Section 5.2.

In this section, we will specialize these two approaches to the dip projection index.

5.1 Orthogonal directions

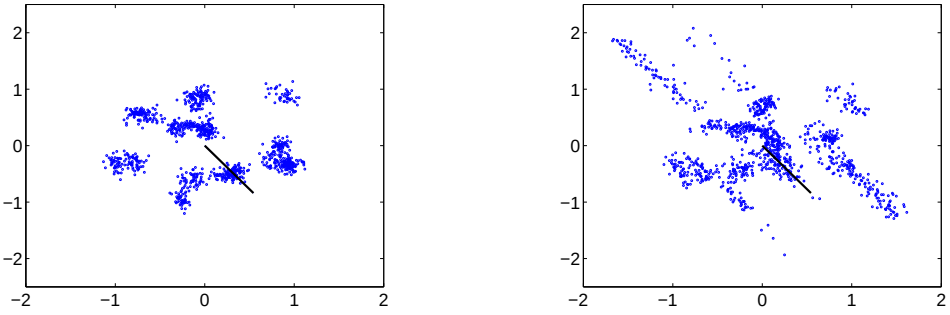
This section discusses the iterative approach on finding multi-dimensional orthogonal projections of high dip. Our local search algorithm is used to compute an initial direction $a_1 \in \mathbb{R}^D$. The local search algorithm then finds another interesting projection a_2 for the data projected onto the orthogonal complement of a_1. This procedure is continued until the desired number of directions $(a_1, \ldots, a_d)$, $d \leq D$, have been computed. $P = [a_1, \ldots, a_d]$ is the desired multidimensional projection.

5.2 Unimodalization of the data

In [Friedman, 1987] a method for structure removal for non-normal projection pursuit is proposed. The method renders the projection along the directions found normal. In the case of non-unimodal projection pursuit, we substitute the *normalization* of the data by *unimodalization*. It is however desirable to modify the data as little as possible. This suggests modifying the data such that the projection along the directions found after the modification is as close as possible to the best fitting unimodal of the original projection. The best fitting unimodal can be directly read off from the data structures used in Algorithm 1.

Unimodalization along direction a is done by computing the appropriate quantiles of the best-fitting unimodal distribution G of $\mathrm{j}(\xi_X(a))$. Then, the necessary displacement $\delta \in \mathbb{R}^N$ with respect to the original one dimensional projection is computed. This effectively determines a monotone mapping $\tau : \mathbb{R} \to \mathbb{R}$ such that $F \circ \tau = G$. The unimodalized data X' is then computed from X by $X' \leftarrow X + \delta \cdot a^T$.

The computation of the quantiles by evaluating G^{-1} is very simple, since the best fitting unimodal G is piecewise linear, and the appropriate piece can be found by simple index computations. Empirical evidence



(a) Before unimodalization (b) After unimodalization

Figure 4: Effect of unimodalization

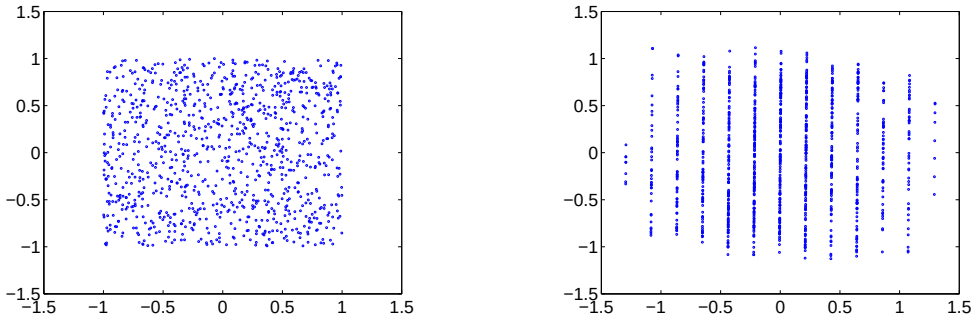
indicates that usually almost no additional multimodality is induced in directions which are non-orthogonal to the direction of unimodalization.

5.3 Problems with a multidimensional generalization of the dip

In general it can be expected that higher-dimensional projections bear more information than those of lower-dimension. These have the advantage that structure can be identified which lower-dimensional indices are oblivious about (e.g. “holes” in the data, cf. [Huber, 1985]). There is however no agreement in the literature on how unimodality should be generalized to multiple dimensions [Wells, 1978]. Hartigan [Hartigan & Hartigan, 1985] suggests several ways to generalize the dip two multiple dimensions. One possibility is to define the multidimensional dip as the maximum dip over all one-dimensional directions. This idea however does not help for the *search* for interesting multi-dimensional projections: If the dip is defined this way, then no matter which subspace is chosen, the projection onto it is assigned the same dip as long as it contains the direction of maximum dip with respect to the original data. Hence this criterion cannot act as a guide for selecting interesting multi-dimensional projections. The two other linearizations proposed are based on a minimum spanning tree on the data, but it is not clear if the continuity and differentiability properties are preserved in this approach.

6 Experimental Results

This section provides empirical evidence about the performance of the proposed algorithms on several surrogate and real-world data sets. These data sets are described and analyzed in detail in Sections 6.1 and 6.2.



(a) RANDU triples projected on the first two dimensions

(b) RANDU triples projected on the first two orthogonal directions optimizing the dip

Figure 5: RANDU planes sample data

6.1 Results on surrogate data

Surrogate data is an important tool in empirically evaluating projection pursuit techniques. Instead of using noisy real-world data, data is sampled from distributions which are "hand-designed" and thus well-understood. We first evaluate the proposed algorithms on data generated by a well-known, poorly designed random number generator. In another experiment we point out, how high variability within one variable of the data can make it impossible to find clustering structure with the commonly used technique of Principal Component Analysis. In our last surrogate data set experiment, we study a mixture of 20 Gaussian distribution in a 10 dimensional space. This experiment was designed to analyze how various projection indices achieve cluster separation.

RANDU. RANDU is an infamous example of a poorly designed random number generator. It is based on the recurrence relation $x_{i+1} = ax_i \mod m$ for $a = 2^{16} + 3 = 65539$ and $m = 2^{31}$, and an arbitrary integer seed $x_0 > 0$. It was employed in this form in the IBM SYSTEM/360 machines. Marsaglia [Marsaglia, 1968] observed that for $i \in \mathbb{N}$ it holds that $9x_i - 6x_{i+1} + x_{i+2} \equiv 0 \mod 2^{31}$. Hence the triples lie on planes, 15 of which have nonempty intersection with the cube $[1..2^{31}]^3$ containing the triples. Figure 5 (a) presents a scaled scatter plot obtained from 1000 RANDU samples with seed $x_0 = 10000$. Although this plot appears to depict a sample of the uniform on the square, Figure 5 (b) shows the projection obtained from the application of our local search algorithm.

A PCA trap. Occasionally, the projections of the data which explain most of the sample variance bear little information. Figure 6 presents a scatter plot for the 200 point sample obtained from the normal mixture

$$F_T = \frac{1}{2}\mathcal{N}\left([-\frac{1}{2}, 0]^T, \begin{bmatrix} 0.2^2 & 0 \\ 0 & 3^2 \end{bmatrix}\right) + \frac{1}{2}\mathcal{N}\left([\frac{1}{2}, 0]^T, \begin{bmatrix} 0.2^2 & 0 \\ 0 & 3^2 \end{bmatrix}\right)$$

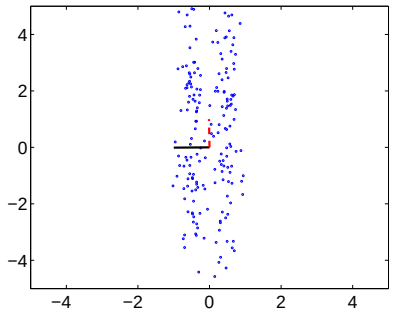


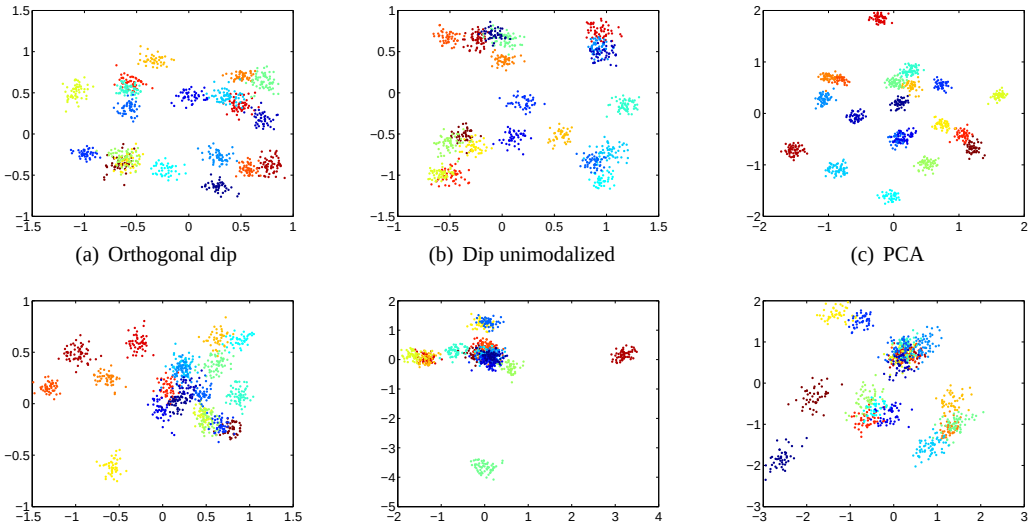
Figure 6: Comparison of the direction which maximizes the dip and the first principal component

It can be seen that the projection along the dip axis provides a clear separation of the two clusters, whereas the projection along the first principal component will conceal this structure. Nason [1992] refers to this phenomenon as *switching point* – for moment methods such as PCA, a single outlier can solely determine the maximum of the projection index. Since the dip works on distributions instead of densities, it appears to be more robust with respect to these issues.

Mixture of 20 Gaussians in $[-1,1]^{10}$. To empirically analyze how various projection indices achieve cluster separation for a highly clustered, relatively low dimensional data set, a surrogate data set was created as described in the following. The experimental data was sampled from a mixture of Gaussian distributions, where the means $\mu_1, \ldots, \mu_8$ where sampled uniformly from the cube $[-1,1]^{10}$ in $\mathbb{R}^{10}$, and the square roots of the diagonal entries $\sigma_{i,j}, 1 \leq i \leq 20, 1 \leq j \leq 10$ of the diagonal covariance matrices where sampled uniformly from the interval $[\frac{1}{10}, \frac{1}{6}]$ – all parameters were sampled in advance. The number of samples of each component varied uniformly between 40 and 60.

Here and in the following, in addition to the iterative and recursive dip algorithms and PCA, three other projection indices were used: Negative Entropy, Friedman's index [Friedman, 1987] and Posse's index, a method based on the χ^2 test [Posse, 1990]. Figure 7 visualizes the results. For each projection index, the top scoring two-dimensional projection was used (or the two top-scoring one-dimensional projections were combined).

For this data set, the non-robust sphering procedure lead to distortions, resulting in inferior results for all projection indices based in on the sphered data Figure 7 (d), (e), (f). The very low within-class variance lead to very good results of the simple PCA procedure (c). Both dip maximization procedures (a), (b) resulted in highly structured projections – the typical segmentation of the data in the directions of the coordinate axes can be perceived.



(a) Orthogonal dip (b) Dip unimodalized (c) PCA

(d) Entropy (e) Friedman's Index (f) Posse's Index

Figure 7: Mixture of 20 Gaussians in $[0,1]^{10}$

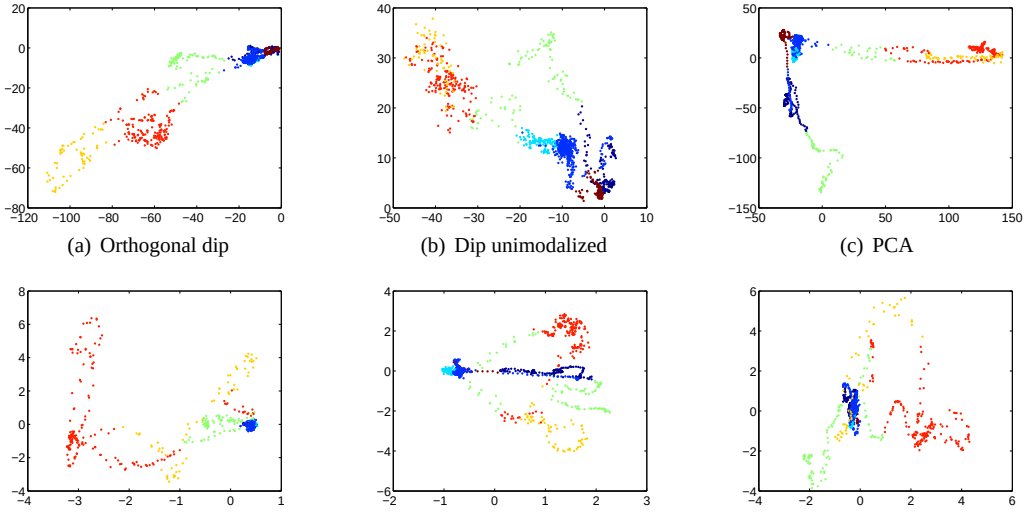
6.2 Results on real-world data

In this section we present exploratory data analyses of three real-world data sets. In our analyses, we compare the dip statistic with several other projection indices.

Human movement data. In [Krause *et al.*, 2003], a method for unsupervised and dynamic identification of physiological and activity context was proposed. Their unsupervised machine learning algorithms require dimension reduction techniques such as PCA. Although PCA worked reasonably well in their approach, it is interesting to compare their results with the application of the dip maximization procedure. In their experiments, they used the SenseWear armband from BodyMedia Inc. to measure activity and physiological sensory information.

In their motion classification experiment, several distinct kinds of movements were performed over a period of about three minutes (1632 samples in total), in the following order: Walking, running, walking, sitting, knee-bends, walking, waving his arms, walking, climbing up and down stairs, one stair a step. For this experiment, only the raw sensor values of the armband's accelerometers were used. The sampling rate was set to 8 Hz. On the sphered data, a non-windowed 64 sample Fast Fourier Transform was computed, individually for each axis. This 128 dimensional data – of which only half the components are unique due to aliasing effects – was then given as input the clustering algorithms described in [Krause *et al.*, 2003], and to the iterative and recursive versions of the dip-maximizer.

Figure 8 presents the results of several dip projection pursuit algorithms. Again, the dip maximization methods are less influenced by the variance in the sample data and appear to give robust clustering



(a) Orthogonal dip (b) Dip unimodalized (c) PCA

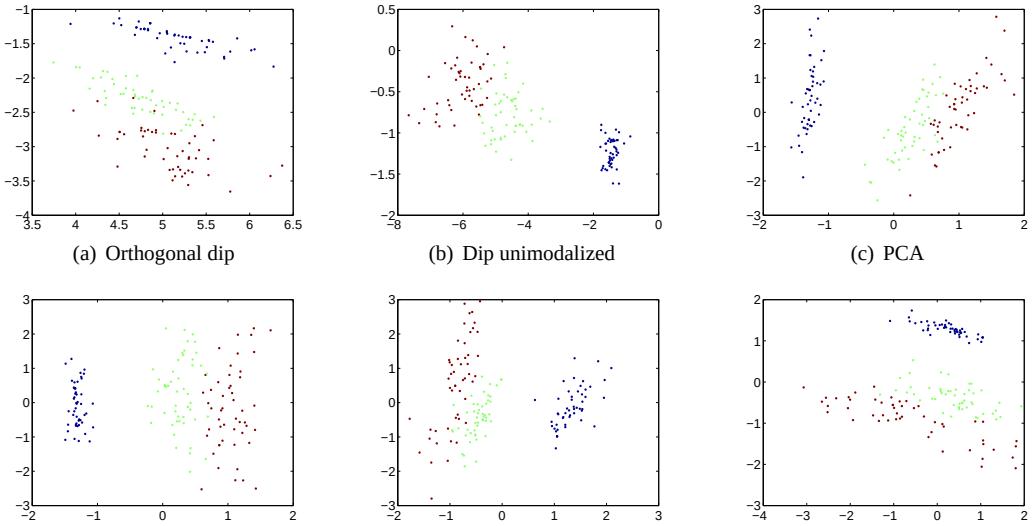
(d) Negative Entropy (e) Friedman's index (f) Posse's index

Figure 8: SenseWear accelerometer sample data

results. For the indices requiring sphered data, only the first ten principal components were used, which explained more than 95% of the variance. In our experiments without preselecting these components, the projection pursuit results could not be distinguished from noise. The projections obtained from the negative entropy, Friedman's and Posse's index (Figure 8 (d), (e), (f)) appear more structured than the results obtained from PCA (c), but seem to be attracted more to heavy tailed marginals than the dip maximization procedures. This experiment also shows that the dip maximization algorithms can operate well on very high-dimensional data.

Iris data. We also analyzed the popular Iris benchmark data set from the UCI machine learning repository (*c.f.* Figure 9). For this four-dimensional data set, all projection indices achieved a clear separation of the three classes Iris Setosa, Iris Versicolor and Iris Virginica. The most compact and well-separated clusters were produced by the unimodalized dip – the only method not necessarily producing orthogonal projections. This experiment shows that the relaxation of orthogonality for the selected projections can result in more compactly clustered projections.

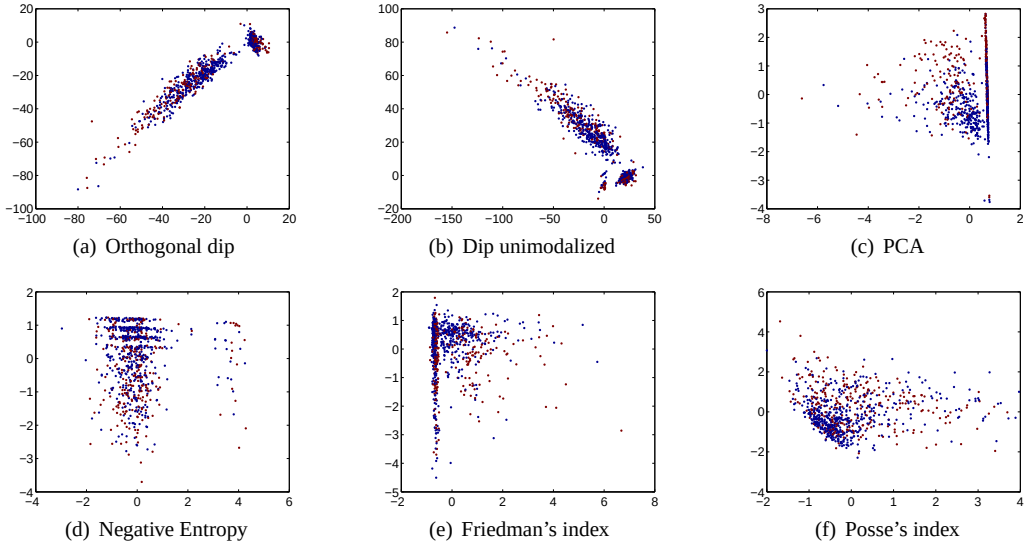
Pima Indians data. Additionally, we visualized the Pima Indians Diabetes benchmark data set from the UCI machine learning repository (*c.f.* Figure 10). This data set is eight-dimensional, and contains data for two classes of subjects – partitioned into whether the subjects tested positive for diabetes. The dip maximization procedures detect two clusters in the orthogonal case and three clusters in the unimodalized case. Again, the other projection indices prefer heavy tailed marginals to compactly clustered projections. The projections identified using Posse's index do not exhibit any clustering structure.



(a) Orthogonal dip (b) Dip unimodalized (c) PCA

(d) Negative Entropy (e) Friedman's index (f) Posse's index

Figure 9: Iris data set



(a) Orthogonal dip (b) Dip unimodalized (c) PCA

(d) Negative Entropy (e) Friedman's index (f) Posse's index

Figure 10: Pima Indians data set

7 Conclusions

In this paper, we investigated the dip, a statistical test for unimodality, as a projection index in the context of projection pursuit. We established continuity and differentiability properties of the dip as a functional on the space of empirical distribution functions. These results were exploited for the design of a gradient ascent based local search algorithm. To enable the search for multi-dimensional projections, two methods – orthogonalization and unimodalization – were proposed. We furthermore presented extensive empirical evidence that the dip is a viable projection index both for surrogate and real-world data. The dip does not require sphered data, eliminating a highly non-robust preprocessing step. It is also insensitive to the occurrence of outliers and prefers compactly clustered projections to heavy tailed distributions. Compared to classical indices such as PCA, negative entropy, Friedman's and Posse's index, the dip appears to provide comparable and in several cases superior performance. Combining their empirical performance with attractive analytical and computational properties, we hope that our proposed methods provide a valuable contribution to the field of exploratory data analysis.

Acknowledgements. Volkmar Liebscher and Andreas Krause were employed at GSF – National Research Center for Environment and Health during part of this work.

A Proofs

Proof of Lemma 3.10. The definition of inactivity guarantees the existence of an $\varepsilon > 0$ such that $\mathrm{Q}_{\xi,i}(h_1, h_2) = 0$ for $|h_1| \leq \varepsilon$ and $|h_2| \leq \varepsilon$. □

Proof of Lemma 3.14. This can be seen from the computation of the dip in Algorithm 1 – the active triangle with maximum height determines the modal interval returned by the algorithm, the active triangle with second largest height determines the dip, and thus must be modal. □

Proof of Lemma 3.13. Since i can only be active if there exists a local deformation of i which changes the dip, and since the dip can only change if the height of a modal triangle changes, i must participate in a modal triangle. The existence of at least one triangle Δ such that $h_\xi(\Delta) = 2\mathrm{D}(\mathrm{j}(\xi))$ can be directly seen from the computation of the dip in Algorithm 1 in line 5. □

Proof of Lemma 3.15. If Δ is the only modal triangle, then, by continuity, for x_{i_2} there exists an open neighborhood O_2 such that deformations within this environment render this situation unchanged, i.e. Δ remains the only modal triangle. Consider the mapping $\gamma : \mathbb{R}^3 \to \mathbb{R}$, $(\delta_{x_{i_1}}, \ldots, \delta_{x_{i_3}}) \mapsto h_{\xi+(0,\ldots,\delta_{x_1},\ldots,\delta_{x_3},\ldots,0)}(\Delta)$. From the construction it is clear that γ is differentiable with respect to i_2 in

$O_2 - x_2 := \{x : x + x_2 \in O_2\}$. If additionally i_1 or i_3 are strongly active, then there exist also open neighborhoods O_1 of x_{i_1} or O_3 of x_{i_3} such that for perturbations within these environments, Δ remains the only modal triangle of ξ and γ is differentiable with respect to i_1 or i_3 within $O_1 - x_1$ or $O_3 - x_3$ respectively. The observation that $\mathrm{D}(\mathrm{j}(\xi + (\dots, \delta_{x_{i_1}}, \dots))) = \frac{1}{2}\gamma(\delta_{x_{i_1}}, \dots)$ within the respective environments proves the claim. □

Lemma A.1. *Let $F = j(\xi)$ and let i be an active index of ξ. Then there exists an $\varepsilon > 0$ such that for $0 < \delta_{x_i} < \varepsilon$, $\delta_x = (0, \dots, 0, \delta_{x_i}, 0, \dots)$ it holds that the dip is partially differentiable at $j(\xi + \delta_x)$ with respect to index i. An analogous statement holds for $-\varepsilon < \delta_{x_i} < 0$.*

Proof. From Lemma 3.13 we know that i is part of a modal triangle. Without loss of generality we only consider the case $\delta_{x_i} > 0$. If x_i is perturbed a little, i can become either inactive or remain active. In the first case, Lemma 3.10 guarantees differentiability of the dip. The key observation in the second case is that there exists an $\varepsilon > 0$ such that as soon as x_i is moved by no more than ε, i remains part of only a single modal triangle which can be seen by a basic geometry consideration. Furthermore, i becomes strongly active. Thus Lemma 3.15 proves the claim. □

Proof of Theorem 3.17. Again, the statement is trivial for inactive indices. Consider the case that i is active and hence is part of a modal triangle. By adding a small positive or negative offset to x_i, the membership of i in modal triangles can change. Without loss of generality we consider the case of increasing x_i. Again there exists a $\delta > 0$ such that for $0 < \delta_{x_i} < \delta$, i is either part of a single modal triangle, or inactive. In the first case, i becomes strongly active. Lemma 3.15 can be used to compute the one-sided partial derivative $\frac{\partial \mathrm{D}}{\partial_+ x_i}$ and to show that it is locally bounded on I. Thus all one-sided partial derivatives with respect to the locations of the masses exist, are locally bounded, and M_ξ can be chosen as $M_\xi = \max\left\{\left|\frac{\partial \mathrm{D}}{\partial_\pm x_i}(F)\right|, i \in \{1, \dots, n\}\right\}$. By construction, $M(\xi) = M_\xi$ holds and since the maximum is taken over a finite number of locally bounded functions, it is clear that M is locally bounded. □

Proof of Corollary 3.18. From Theorem 3.17 it follows that for arbitrary ξ, the dip is continuous with respect to relocation of a single index within an open neighborhood around ξ. Let δ_1 be such that for $|\delta_{x_1}| \leq \delta_1$ and $x_1 + \delta_{x_1} < x_2$ it holds that $|\mathrm{D}(\mathrm{j}(\xi + (\delta_{x_1}, 0, \dots, 0))) - \mathrm{D}(\mathrm{j}(\xi))| < \frac{\varepsilon}{2^1}$. Now let $0 < \delta_2 \leq \delta_1$ be such that for $|\delta_{x_2}| < \delta_2$ it holds that $|\mathrm{D}(\mathrm{j}(\xi + (\delta_{x_1}, \delta_{x_2}, 0 \dots, 0))) - \mathrm{D}(\mathrm{j}(\xi + (\delta_{x_1}, 0, \dots, 0)))| < \frac{\varepsilon}{2^2}$ regardless of the choice of δ_{x_1} which is possible due to a compactness argument using the local boundedness of $M(\xi)$ from Theorem 3.17. Proceeding in the same manner, we arrive at a monotonically decreasing sequence of δ_k such that $\frac{\delta_{2n}}{2}$ is the constant δ promised in the claim. The sequence of changes in dip converges, since it is bounded by the geometric series $\varepsilon \sum_{k=1}^{\infty} \frac{1}{2^k} = \varepsilon$. □

Proof of Theorem 4.4. This is an immediate consequence of the observation that Corollary 3.18 applies to all locations of masses generated by a projection which is not X-singular. □

Proof of Lemma 4.6. It is clear by the Cauchy-Schwarz inequality that $|(a+\delta)^T x_i - a^T x_i| = |\delta^T x_i| \leq \|\delta\|_2 \|x_i\|_2$. Let $i < j$. Then $(a+\delta)^T x_j - (a+\delta)^T x_i \geq a^T x_j - a^T x_i - \frac{\varepsilon_j + \varepsilon_i}{2} \geq 0$ since $a^T x_j \geq a^T x_i$ and $\max(\varepsilon_i, \varepsilon_j) \leq (a^T x_j - a^T x_i)$ per definition. □

References

[Bellman, 1961] Bellman, R. (1961). *Adaptive Control Processes*. Princeton University Press.

[Billingsley, 1968] Billingsley, P. (1968). *Convergence of Probability Measures*. John Wiley & Sons, Inc.

[Cox & Hinkley, 1974] Cox, D. R. & Hinkley, D. V. (1974). *Theoretical Statistics*. Chapman and Hall.

[Friedman, 1987] Friedman, J. H. (1987). *J American Statistical Association,* **82** (397), 249–266.

[Hartigan, 1977] Hartigan, J. A. (1977). *Classification and Clustering*. Academic.

[Hartigan & Hartigan, 1985] Hartigan, J. A. & Hartigan, P. M. (1985). *Annals of Statistics,* **13** (1), 70–84.

[Hartigan, 1985] Hartigan, P. M. (1985). *Applied Statistics,* **34** (3), 320–325.

[Heun, 2000] Heun, V. (2000). *Grundlegende Algorithmen.* vieweg.

[Huber, 1985] Huber, P. J. (1985). *The Annals of Statistics,* **13** (2), 435–475.

[Jones & Sibson, 1987] Jones, M. C. & Sibson, R. (1987). *J Royal Statistical Association A,* **150**, 1–36.

[Krause, 2004] Krause, A. (2004). Diplomarbeit Technische Universität München Munich.

[Krause *et al.*, 2003] Krause, A., Siewiorek, D., Smailagic, A., & Farringdon, J. (2003). In: *The Seventh International Symposium on Wearable Computers* , IEEE White Plains, NY, USA:.

[Marsaglia, 1968] Marsaglia, G. (1968). *Proceedings of the National Academy of Sciences,* **61**, 25–28.

[Morton, 1989] Morton, S. C. (1989). Technical Report 106 Department of Statistics, Stanford University.

[Müller & Sawitzki, 1991] Müller, D. W. & Sawitzki, G. (1991). *Journal of the American Statistical Association,* **86** (415), 738–746.

[Nason, 1992] Nason, G. P. (1992). *Design and choice of projection indices*. PhD thesis Univ. of Bath.

[Nason & Sibson, 1992] Nason, G. P. & Sibson, R. (1992). *Statistics and Computing,* .

[Posse, 1990] Posse, C. (1990). *Comm. Statist. Simul. Comput.* **19** (4), 1143–1164.

[Silverman, 1981] Silverman, B. W. (1981). *Journal of the Royal Statistical Society B,* **43**, 97–99.

[Wegman, 1970] Wegman, E. J. (1970). *The Annals of Mathematical Statistics,* **41** (6), 2169–2174.

[Wells, 1978] Wells, D. R. (1978). *The Annals of Statistics,* **6** (4), 926–931.