

Learning with Algebraic Invariances, and the Invariant Kernel Trick

Franz J. Király^{* 1}, Andreas Ziehe^{† 2}, and Klaus-Robert Müller^{‡ 2}

¹ Department of Statistical Science, University College London,
Gower Street, London WC1E 6BT, United Kingdom

² Machine Learning Group, Technische Universität Berlin,
Marchstrasse 23, 10587 Berlin, Germany

Abstract

When solving data analysis problems it is important to integrate prior knowledge and/or structural invariances. This paper contributes by a novel framework for incorporating algebraic invariance structure into kernels. In particular, we show that algebraic properties such as sign symmetries in data, phase independence, scaling etc. can be included easily by essentially performing the kernel trick twice. We demonstrate the usefulness of our theory in simulations on selected applications such as sign-invariant spectral clustering and underdetermined ICA.

1. Introduction

The construction of algorithms that encompass problem specific invariances has been an important line of research in pattern recognition and machine learning. Properly incorporating invariances into a model allows to constrain the underlying function class and thus to increase on one hand generalization and on the other hand resistance against outliers and robustness against nonstationarities that are caused by the respective invariance transformations. In principle, machine learning models could become invariant by learning from large corpora of data that contain the respective transformations with respect to which invariance needs to be achieved, e.g. ambient noise in speech recognition (Hinton et al., 2012) or robustness against translation, rotation or thickness transformations in handwritten digit recognition, (e.g. Jebara (2003), LeCun et al. (1998), Schölkopf et al. (1997), Simard et al. (1992)). Two general lines of research have enjoyed high popularity, kernel methods (Boser et al., 1992, Schölkopf et al., 1998, Vapnik, 1995), where kernels can be *engineered* to reflect complex invariances or prior knowledge, e.g. in Bioinformatics (Jiang and Ching, 2012, Zien et al., 2000), or (deep) neural networks where large amounts of data help to *learn* a representation (Bengio, 2009, Krizhevsky et al., 2012, LeCun et al., 1998).

We will contribute to kernel methods and address the fundamental question how prior knowledge on invariances can be *directly* incorporated into the kernel formalism. More specifically, the kernel trick extracts features, and it is important to ask how these kernel features

*f.kiraly@ucl.ac.uk

†andreas.ziehe@tu-berlin.de

‡klaus-robert.mueller@tu-berlin.de

can be made naturally invariant under data-specific symmetries, such as sign change, mirror symmetry, common complex phase factor, rotation, and so on.

In particular, we propose an algebraic method that we will call *the invariant kernel trick*. It allows (i) a modification of any (positive semi-definite) kernel into a suitable invariant kernel by (ii) applying the kernel trick twice. Namely, properties of two kernels are combined – one for the invariance, and one for the features. (iii) Invariant kernels come without an increase in the computational cost of kernel evaluation, and finally (iv) the derived kernel is canonical, and shares fundamental properties with the original. The kernel invariant trick can be readily applied to any kernel-based method and naturally and immediately implements the desired algebraic invariance structure.

To underline the versatility of our novel invariance inducing framework, we exemplarily show sign-invariant clustering simulations on toy data obtained by modifying the USPS data set, and sign- and scale invariant signal separation on the real world “flutes” data set. These experiments are primarily intended to illustrate ease of use, usefulness and the broad applicability of the kernel invariant trick. While we provide the theoretical details of many potential invariance inducing transformations, it is clearly unfeasible to show simulations for all of them in combination with all possible kernel algorithms, therefore we have focused on clustering, sign invariance, and sign-and-scale invariance.

In the next section we will lay out our theory on invariances, then discuss the spectral clustering algorithm variant used in section 3 and finally provide experimental results and discussion.

2. Learning with Algebraic Invariances

2.1. An invariant kernel example: clustering on the sphere

We first illustrate the main idea in a simple example. Suppose one is given data points in \mathbb{R}^n , coming from two different classes. Suppose these classes form clusters up to sign - that is, for each data point $x \in \mathbb{R}^n$ is considered equivalently to $-x$, e.g., if one is interested only in the span of the normal vector x . After flipping the sign for some of the points, they fall exactly in one of the clusters (see Figure 1).

The first step is as follows: observe that for any $x \in \mathbb{R}^n$, knowledge of the matrix $X = xx^\top \in \mathbb{R}^{n \times n}$ is equivalent to knowledge of x up to sign. Namely, x can be obtained from X by an eigenvector decomposition of X . One obtains $x = \pm v \cdot \|X\|$, where v is the eigenvector of X with biggest eigenvalue, and $\|X\|$ the Frobenius norm of X . Thus the matrix X can be interpreted as a collection of invariant features.

One could now replace every data point x by xx^\top , but through this maneuver the dimension of the data is roughly squared, thus the computational cost of learning in this representation increases considerably when compared to learning in the original ambient space. One may try to alleviate by computing random projections, e.g., clustering the vector of scalars $\text{Tr}(x_i x_i^* A_j)$ for $i = 1 \dots M$ and $j = 1 \dots m$ for some $M < n^2$ and random/generic matrices $A_j \in \mathbb{R}^{n \times n}$. But again, this is a heuristic step and can lead to inaccuracies and random fluctuations (even though it may work in some scenarios, and compressed-sensing-like guarantees can be derived for certain asymptotic settings).

We propose a different strategy which does not incur these problems. It combines the above invariant representation with the kernel trick and a kernel based learning method, for instance kernel k -means clustering. The typical kernel for clustering purposes is the Gaussian

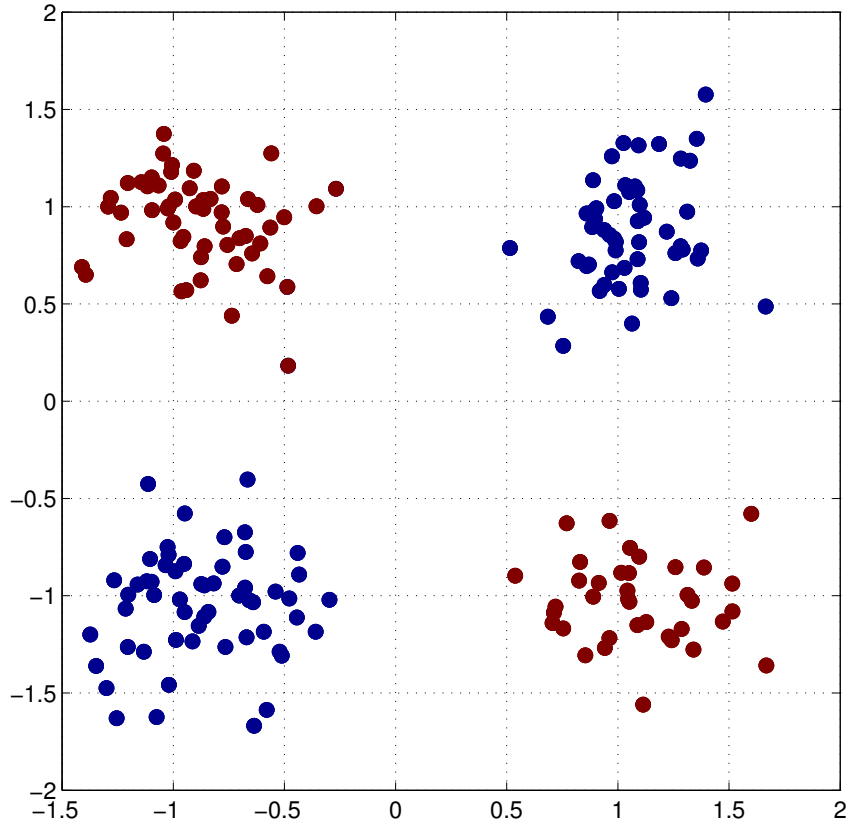


Figure 1: The XOR dataset as a sign-invariant clustering problem.

kernel

$$k(X, Y) = \exp\left(-\frac{\|X - Y\|^2}{2\sigma^2}\right).$$

The crucial second step is taking the kernel not with respect to the original data points x , but on the invariant features $X = xx^\top$, $Y = yy^\top$, obtaining a new kernel for x and y of the form

$$k_{\pm}(x, y) := k(X, Y) = \exp\left(-\frac{\|xx^\top - yy^\top\|^2}{2\sigma^2}\right),$$

where the norm is again the Frobenius norm on matrices. Applying the binomial expansion, one obtains

$$k_{\pm}(x, y) = \exp\left(-\frac{\|x\|^4 + \|y\|^4 - 2\langle x, y \rangle^2}{2\sigma^2}\right). \quad (1)$$

This is a sign-invariant variant of the Gaussian kernel which is indeed a (positive definite) kernel and none more expensive to evaluate than the original Gaussian kernel. It is further similar to the original by having an adaptable kernel width.

We will show and argue that these two steps are an instance of a very general and mathematically natural strategy for arbitrary kernels, which allows to construct modified invariant kernels for a large class of invariances. We will use the above example as a running illustration of the general strategy.

2.2. Invariant kernels

We start by introducing an abstract and general **setting** for invariances, which is as follows: we start with data points $x_1, \dots, x_N \in \mathbb{K}^n$, with $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$, and an invariant group action (finite or continuous)

$$G \times \mathbb{K}^n \rightarrow \mathbb{K}^n, \quad (g, x) \mapsto g.x.$$

That is, an data point $x \in W$ is considered invariance-equivalent to the elements of the so-called orbit

$$G.x = \{g.x : g \in G\}.$$

We will assume (for technical reasons: ensuring existence of a well-defined quotient) that the group action is algebraic, that is, all maps $x \mapsto g.x$ are algebraic maps. In the example presented in section 2.1, the group is the “sign group” $G = (\{+1, -1\}, \cdot)$, with group operation \cdot given by $+1 = +1 \cdot +1 = -1 \cdot -1$ and $-1 = +1 \cdot -1 = -1 \cdot +1$, and group action given by $+1.x = x$ and $-1.x = -x$ (this group G is, as a group, isomorphic to the group $\mathbb{Z}/2$).

Our main goal is to start from an arbitrary class of kernels $k : \mathbb{K}^m \times \mathbb{K}^m \rightarrow \mathbb{R}$ (where the kernel is defined for any integer m not necessary equal to n , such as for Gaussian or polynomial kernels) and make them invariant under the G -action. Mathematically, there are two meaningful but different ways for k to be invariant:

Definition 2.1. *Let $W \subseteq \mathbb{K}^n$. Consider a group G acting on W , and a function $k : W \times W \rightarrow \mathbb{R}$. The function k is called:*

- (i) *G -invariant, if $k(x, y) = k(g.x, h.y)$ for all $g, h \in G$ and all $x, y \in W$.*
- (ii) *diagonally G -invariant, if $k(x, y) = k(g.x, g.y)$ for all $g \in G$ and all $x, y \in W$.*

If k is a kernel, we similarly call it a G -invariant or diagonally G -invariant kernel.

Note that G -invariant kernels are also diagonally G -invariant, but the converse is in general not true.

2.3. Invariant kernels: characterization, existence and uniqueness

Before constructing invariant kernels, we will characterize them abstractly. Similarly to the Moore-Aronszajn theorem, which asserts the existence of a unique feature space, a classical theorem of invariant theory asserts existence of a unique invariant space. We start by stating (technically convenient variants of) both theorems:

Theorem 1 (Moore-Aronszajn). *Let $W \subseteq \mathbb{K}^n$ compact, let $k : W \times W \rightarrow \mathbb{R}$ be a symmetric, positive definite kernel. Then there exist a \mathbb{K} -Hilbert space \mathcal{F} and a continuous map $\phi : W \rightarrow \mathcal{F}$, unique up to isomorphism, such that $k(x, y) = \langle x, y \rangle_{\mathcal{F}}$ for all $x, y \in W$.*

The Hilbert space \mathcal{F} is called *reproducing kernel Hilbert space* (RKHS) or *feature space* associated to k . The map ϕ is called the *feature map*.

Theorem 2 (Universal property of group quotient). *Let $W \subseteq \mathbb{K}^n$ compact, and G a group acting algebraically on W . Then, there exist a \mathbb{K} -Hilbert space of invariants \mathcal{J} and an algebraic map $q : W \rightarrow \mathcal{J}$, unique up to isomorphism, such that any G -invariant (piecewise) continuous map $\phi : W \rightarrow \mathcal{F}$ to a Hilbert space \mathcal{F} admits a unique factorization $\phi = \phi' \circ q$ with $\phi' : W/G \rightarrow \mathcal{F}$ continuous, where $W/G = q(W)$. If ϕ is algebraic, then so is ϕ' .*

Here, W/G is called the *quotient space* of W w.r.t. G , and is also called orbifold if G is finite and W is a differentiable manifold. The map q is called the *quotient map* or *invariant map*. It encodes the invariants of W with respect to the group action G , and Theorem 2 asserts that it does so in a unique and canonical way.

If G is finite, it can be shown that \mathcal{J} is as well. If G is infinite, \mathcal{J} needs not to be of finite dimension, but it can be shown to be countable always, and finite for all G in this manuscript. See the appendix for more details. Both theorems also admit more technical variants where W can be more general (e.g. Moore-Aronszajn for non-compact W , group quotient for non-compact W), which we omit for sake of exposition and readability.

The two theorems can immediately be combined into a corollary which characterizes invariant kernels:

Theorem 3. *Let $W \subseteq \mathbb{R}^n$ compact. Consider a group G acting on W , and a let $k : W \times W \rightarrow \mathbb{R}$ be a (piecewise) continuous symmetric positive definite kernel.*

- (i) *If k is G -invariant, then there is a symmetric positive definite kernel $k' : W/G \times W/G \rightarrow \mathbb{R}$, such that $k(x, y) = k'(q(x), q(y))$, where $q : W \rightarrow W/G$ is the canonical quotient map. k' is unique (up to isomorphism of W/G), the feature spaces associated to k and k' coincide, and for the respective feature maps ϕ, ϕ' , it holds that $\phi = \phi' \circ q$, thus $k(x, y) = \langle \phi'(q(x)), \phi'(q(y)) \rangle_{\mathcal{F}}$.*
- (ii) *If k is diagonally G -invariant, then there is a function $f : (W \times W)/G \rightarrow \mathbb{R}$, such that $k(x, y) = f(Q(x, y))$, where $Q : W \times W \rightarrow (W \times W)/G$ is the canonical (diagonal) quotient map.*

It should be specifically noted that Theorem 3 asserts a unique (up to isomorphism) decomposition of the kernel in an invariant part independent of k , given by q resp. Q , and a “structural” part, given by k' resp. f .

Some previous results on diagonally G -invariant kernels, including some statements on the corresponding feature spaces, can be already found in section 2.3 of Haasdonk and Burkhardt (2007) and chapters 4.4 of Kondor (2008). We would like to note that they do not explicitly describe - therefore do not allow to explicitly construct - the invariant space in terms of the quotient map and its canonical invariants.

Intuitively, a G -invariant kernel is always a kernel *on* the invariant features $q(x), q(y)$, while a diagonally G -invariant kernel is always a function in certain bivariate invariant functions $Q(x, y)$ (where in general x and y can not be separated).

Conversely, if one wants to construct kernels with certain invariance properties, the factorization Theorem 3 (ii) for diagonal G -invariances immediately implies that any positive definite function of the diagonal invariants $Q(x, y)$ will be diagonally invariant. Theorem 3 (i) for G -invariances implies that any class of kernels can be made into a G -invariant version by applying it to the features $q(x), q(y)$ - by the theorem, this procedure is canonical (up to isomorphism of the quotient W/G). The example in section 2.1 constructs a $\mathbb{Z}/2$ -invariant variant of the Gaussian kernel by observing that the quotient/invariant map is the map $q : x \mapsto xx^\top$.

Notation 2.2. *In general, for a kernel $k : W \times W \rightarrow \mathbb{R}$ and an invariant group action G , we will denote the G -invariant kernel canonically obtained by the procedure outlined above by k_G . If k has a name, e.g. Gaussian kernel, we call k_G the G -invariant kernel of that name, e.g. the G -invariant Gaussian kernel (keeping in mind that uniqueness holds only up to isomorphism of W/G).*

2.4. The invariant kernel trick

While the existence of q or Q is guaranteed by Theorem 3, the maps and associated invariant features can in principle be difficult or intractable to compute in practice - just as most of the common kernel features are difficult to obtain explicitly. The set of ideas outlined above is therefore only practically appealing when q or Q (or relevant parts thereof) are obtainable in (low-order) polynomial time, or in the following situation:

(a) the original kernel $k : W \times W \rightarrow \mathbb{R}$ is a function of Euclidean scalar products $k(x, y) = f(\langle x, x \rangle, \langle x, y \rangle, \langle y, y \rangle)$.

(b) the quotient/invariant map is the feature map of another efficiently computable kernel $\iota(x, y)$, that is, $\iota(x, y) = \langle q(x), q(y) \rangle$ where $q : W \rightarrow W/G$ is the quotient map.

\Rightarrow in this case, the G -invariant kernel may be obtained as $k_G(x, y) = f(\iota(x, x), \iota(x, y), \iota(y, y))$.

We call this the “*invariant kernel trick*”, as the double application of the kernel trick allows us to avoid an explicit and potentially tedious computation of $q(x)$ and $q(y)$. In case of existence, we call ι the *invariant kernel*. Note that in this case, $\mathcal{J} = \text{span } W/G$ is the feature space associated to ι .

Conditions (a) and (b) are fulfilled in section 2.1, but they may seem very special and a weak generalization of that example. This is, however, not true, as we argue in the following.

Condition (a) is true for the majority of the more common kernels on \mathbb{K}^n in practical use, such as: Gaussian kernel, Laplace kernel and all other RBF kernels, homogenous and inhomogenous polynomial kernels, sigmoid kernel, ANOVA kernel. The only exceptions are kernels where a maximum/minimum is taken, such as spline and histogram kernels, and the more combinatorial kernels, e.g. on strings and graphs, which in many cases do not take arguments in \mathbb{K}^n . Moreover, the following result, which is an application of Theorem 3 (ii), implies that (a) is fulfilled for any kernel invariant under orthogonal/unitary, or Euclidean isometries; absence of such an invariance would imply an unparsimonious imbalancedness on the input representation.

Lemma 2.3. *Let $k : W \times W \rightarrow \mathbb{R}, W \subseteq \mathbb{K}^n$ be a (piecewise) continuous function which is diagonally invariant under orthogonal/unitary transform, that is, w.r.t. the canonical group action of the orthogonal/unitary matrices $O(n)$ resp. $U(n)$ on \mathbb{K}^n . Then, there is a function $f : \mathbb{K}^3 \rightarrow \mathbb{R}$ such that $k(x, y) = f(\langle x, x \rangle, \langle x, y \rangle, \langle y, y \rangle)$.*

Lemma 2.4. *Let $k : W \times W \rightarrow \mathbb{R}, W \subseteq \mathbb{K}^n$ be a (piecewise) continuous function which is diagonally invariant under the Euclidean isometries $Euc(n)$ on \mathbb{K}^n . Then, there is a function $f : \mathbb{R} \rightarrow \mathbb{R}$ such that $k(x, y) = f(\|x - y\|)$.*

The converses of these Lemmas are easily seen to hold. Proofs of Lemmas 2.3 and 2.4 follow from Lemmas 5.6 and 5.8 in the appendix, together with Theorem 3 (ii).

Lemma 2.4 states that isometry-invariant kernels can always be expressed as a radial basis function kernels $k(x, y) = f(\|x - y\|)$. Note that the isometry-invariant kernels are also unitary invariant, thus both fulfill condition (a) by Lemma 2.3; more explicitly, note that $\|x - y\|^2 = \langle x, x \rangle - \langle x, y \rangle - \langle y, x \rangle + \langle y, y \rangle$, thus RBF kernels fulfill condition (a).

Condition (b), on the other hand, is fulfilled for a variety of simple invariances, which we non-exhaustively list in the following section.

2.5. Invariant maps and invariant kernels

We proceed by deriving invariant kernels for a list of common invariances. For exposition, we will explicitly write down the invariant versions of the Euclidean scalar product $k^E(x, y) = \langle x, y \rangle$, the inhomogenous polynomial kernel $k^\pi(x, y) = (\langle x, y \rangle + 1)^d$, and the Gaussian kernel $k^\gamma(x, y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$.

2.5.1. Finite rotation invariance This is a slight generalization of the running example from section 2.1: the group is $G = \mathbb{Z}/m$ acting on \mathbb{K}^n , with the action being $\ell \cdot x = \zeta_m^\ell x$, where ζ_m is a complex m -th root of unity. As $\zeta_2 = -1$, the case $m = 2$ is sign invariance. For general m , the action rotates each coordinate of x in the complex plane by an angle of $2\pi/m$. One can show: the invariant map is $q : x \mapsto x^{\otimes m}$, where $x^{\otimes m}$ is the m -th outer product tensor of x . The invariant space is the vector space of symmetric tensors of degree m .

An elementary computation shows that there is an invariant kernel $\iota(x, y) = \langle x^{\otimes m}, y^{\otimes m} \rangle = \langle x, y \rangle^m$, which coincides with k_G^E and the homogenous polynomial kernel of degree m . Note that the invariant map q needs not to be evaluated in computation of the invariant kernel ι . Using the invariant kernel trick, one further obtains $k_G^\pi(x, y) = (\langle x, y \rangle^m + 1)^d$ and $k_G^\gamma(x, y) = \exp\left(-\frac{\|x\|^{2m} + \|y\|^{2m} - \langle x, y \rangle^m - \langle y, x \rangle^m}{2\sigma^2}\right)$.

2.5.2. Phase invariance Phase invariance means that the data are in \mathbb{C}^n , with $v \equiv \exp(2\pi i \cdot \varphi) v$ for any $\varphi \in \mathbb{R}$ (and no further equivalences). The group is $G = \mathbb{R}/\mathbb{Z}$, acting on \mathbb{C}^n as $\varphi \cdot v = \exp(2\pi i \cdot \varphi) \cdot v$. One can show that the quotient map is $w : v \mapsto vv^*$, where v^* denotes Hermite transpose of v . This leads to the invariant kernel $\iota(x, y) = \langle x, y \rangle \cdot \langle y, x \rangle$. The phase invariant kernels are in complete analogy to the sign invariant kernels.

2.5.3. Scale invariance For data in $\mathbb{R}^n/\{0\}$, scale invariance or projective invariance means that two data points are considered the same if they differ by a common scale factor, i.e., $v \equiv \alpha v$ for any $\alpha \in \mathbb{R}^+$. The group is (additive) $G = \mathbb{R}$, with the (multiplicative) action $\alpha \cdot v = \exp(\alpha) \cdot v$. One shows that the quotient map is $q : x \mapsto \frac{x}{\|x\|}$, obtaining a scalar product $\iota(x, y) = \frac{\langle x, y \rangle}{\|x\| \|y\|}$. Interestingly, this is indeed a positive definite kernel, related to the so-called correlation kernel of Jiang and Ching (2012). The invariant kernel trick yields $k_G^E = \iota$, $k_G^\pi(x, y) = (\|x\|^{-1} \|y\|^{-1} \langle x, y \rangle + 1)^d$ and $k_G^\gamma(x, y) = \exp\left(\frac{\langle x, y \rangle}{\sigma^2 \|x\| \|y\|} - \sigma^{-2}\right)$.

2.5.4. Scale and sign/phase invariance or projective invariance The scale invariant feature map still leaves a sign/phase ambiguity. Having sign and scale invariance means $v \equiv \alpha v$ for any $\alpha \in \mathbb{K} \setminus \{0\}$, the group is $G = \mathbb{K}/\{0\}$ with action $\alpha \cdot v = \alpha \cdot v$. One way to obtain the invariant kernels is via direct computation of the invariants - or one can apply the invariant kernel trick twice, applying either scale invariance to the sign/phase invariant kernel or sign/phase invariance to the scale invariant kernels. Either way, one finds the scale and sign/phase invariant kernel $\iota(x, y) = \frac{\langle x, y \rangle \cdot \langle y, x \rangle}{\langle x, x \rangle \cdot \langle y, y \rangle}$. Applying the invariant kernel trick with this invariant kernel (thus effectively the original kernel trick thrice) yields $k_G^E = \iota$, $k_G^\pi(x, y) = (\|x\|^{-2} \|y\|^{-2} |\langle x, y \rangle|^2 + 1)^d$ and $k_G^\gamma(x, y) = \exp\left(\frac{\langle x, y \rangle \langle x, y \rangle}{\sigma^2 \|x\|^2 \|y\|^2} - \sigma^{-2}\right)$.

2.5.5. Multiple invariances In general, as demonstrated in the example of scale plus sign/phase invariance one can combine multiple invariances by repeating the invariant kernel trick with different invariances. If actions are compatible, the sequence of groups G_1, G_2, \dots, G_k does

not matter, as the final quotient will be the quotient of the group generated by the G_i . Care should be taken only for invariances where the group action has a different presentation, as the equivalence relation can trivialize, thus mapping all points to a single one in the quotient.

2.5.6. Matrix invariances Another class of invariances which is practically relevant are invariances of matrix groups acting on $\mathbb{K}^{mr} \cong \mathbb{K}^{m \times r}$, for example the group of $(r \times r)$ permutation matrices or unitary matrices acting as $U.X = XU$. Permutation invariances give rise to bag-of-features representations, while unitary action is related to the empirical moments of the data rows. Row-wise translation is \mathbb{K}^m acting on $\mathbb{K}^{m \times r}$ through $\mu.X = X + \mathbb{1}\mu^\top$ with $\mathbb{1}$ being an m -vector of ones, leading to centered invariant kernels.

As quotient spaces for matrix invariances are more difficult to characterize than the ones presented, and so are the invariant kernels, we postpone the presentation to a longer version of this report.

2.6. Related concepts and literature

Both G -invariant and diagonally G -invariant kernels have been to our knowledge first defined and studied by Haasdonk and Burkhardt (2007), where they are called *totally invariant* and *simultaneously invariant* (as naming is not consistent throughout literature, we have decided to adopt a notation more closely inspired by invariant theory). Proposition 7 in (Haasdonk and Burkhardt, 2007) asserts existence of the invariant feature space for G -invariant kernels. The mathematical concepts described in the paper differ from our approach in our opinion by being less general and less explicit: (a) The two main contributions in (Haasdonk and Burkhardt, 2007) are two specific kernels which are G -invariant the translation integration (TI) kernel, and the invariant distance substitution (IDS) kernel. The TI kernel requires a Lie group endowed with Haar measure, and is given in terms of an invariant Haar integral; the IDS kernel on first sight looks similar to our invariant kernel trick, but instead of applying the kernel trick twice, IDS kernels require a metricized group and substitutes hard-to-compute infimal distances into an RBF kernel; the IDS-type generalization of inner product further requires the choice of an (possibly arbitrary) origin. (b) In (Haasdonk and Burkhardt, 2007), the existence of an G -invariant feature space is proven, but no further explicit statements on its structure are made. In particular, a result such as Theorem 3 which explicitly describes the form of kernels in terms of canonical invariants, given by the quotient map, is not available. On the other hand, our results not only show that the invariant kernel trick is universal, but also that any TI and IDS kernels as constructed in (Haasdonk and Burkhardt, 2007) must obey the specific form of Theorem 3, implying that they could also be obtained by the invariant kernel trick, potentially avoiding computation of integrals and infimal distances on Lie groups.

One further related reference is (Kondor, 2008), chapter 4.4 and following, where *diagonally* G -invariant kernels are studied, which behave quite different from G -invariant kernels (see above). Definition 4.4.1 refers to a definition of diagonal G -invariance in the reference [Krein, 1950], which we were not able to obtain at time of submission. Theorem 4.4.3 is a direct application of the Reynolds formula to diagonal invariance - it implicitly leads to the definition of a (non-diagonal) G -invariant kernel k^G . As we understand, k^G is defined only for finite G and furthermore needs a diagonally G -invariant kernel (which can be more difficult to obtain) to construct a G -invariant kernel. Moreover, k^G is not used in the remainder of the manuscript, as the intention of 4.4.3 seems to be the relation of what in our terminology would be the diagonally G -invariant kernel and the feature space of the (non-diagonally) G -invariant kernel.

There is some further literature on diagonally G -invariant kernels, see e.g. (Walder and Chapelle, 2007). As G -invariant kernels are diagonally G -invariant, the respective results also hold for G -invariant kernels.

There is also an approach outlined by Chapelle and Schölkopf (2001) which aims at incorporating invariances with differential methods. However, the invariances discussed there are motivated by vertical/horizontal translation of a pixel image encoded as an element of $\mathbb{R}^{m \times n}$, which is *not* an invariance in our sense, since translation by pixels does not give rise to a group action on $\mathbb{R}^{m \times n}$. Therefore our findings on the constructions of G -invariant kernels do not contradict the introductory statement of Chapelle and Schölkopf (2001) that globally invariant kernels do not exist, simply because the concept of invariance in (Chapelle and Schölkopf, 2001) is different and specific to the pixels image application, as Walder and Chapelle (2007) already correctly note in their introduction.

3. Algorithms

It goes without saying that the proposed invariant kernels can be used in *any* kernel-based learning algorithm, see (Schölkopf and Smola, 2002, Schölkopf et al., 1998). In order to study the performance of our method we focus on a particular unsupervised learning technique: kernel-based spectral clustering.

Kernel-based spectral clustering

Spectral clustering methods have been pioneered by Meila and Shi (2001), Ng et al. (2002) and have been studied and extended by e.g. Dhillon et al. (2004), Sugiyama et al. (2014), von Luxburg (2007), Zelnik-manor and Perona (2005).

In recent work it has been shown that kernel entropy component analysis (kernel ECA) can be used to cluster data (Jenssen, 2010). As kernel PCA, kernel ECA is based on an eigen decomposition of the kernel matrix K but identifies those kernel principal axes that contribute most to the Rényi entropy estimate of the data. It has been demonstrated that selected kernel ECA components provide information about the clustering structure of the data by analysing the eigenvalues and the eigenvectors of the kernel matrix (Jenssen, 2010).

In order to demonstrate that any existing kernel algorithm can in principle profit from the concepts introduced above, we will in the following use the kernel-based spectral clustering algorithm described in (Jenssen, 2010) with Gaussian kernels invariant w.r.t. different symmetries.

4. Experiments

In section 2.1 we introduced a novel class of Mercer kernels that are invariant to sign flips of the input data.

As an example, in this section we consider the problem of spectral clustering under sign invariances. For our experiments we used the MATLAB implementation of kernel ECA by Robert Jenssen ¹ and our MATLAB implementation of the sign-invariant Gaussian kernel in Eq. (1).

¹<http://ansatte.uit.no/robert.jenssen/software.html>

4.1. Spectral clustering of handwritten digits

In order to illustrate clustering under sign invariance we choose 98 images of handwritten zeros and ones from the USPS dataset (Hull, 1994) where we randomly switched the signs of the 16×16 pixel images. The input data consists of 98 vectors in 256 dimensions. We apply spectral clustering using kernel ECA with (a) a standard Gaussian kernel and (b) the sign-invariant Gaussian Kernel. The results in figure 2 show that in the invariant case a successful grouping into groups of zeros and ones, respectively, occurs, i.e. ignoring the black vs white groups caused by random sign flips.

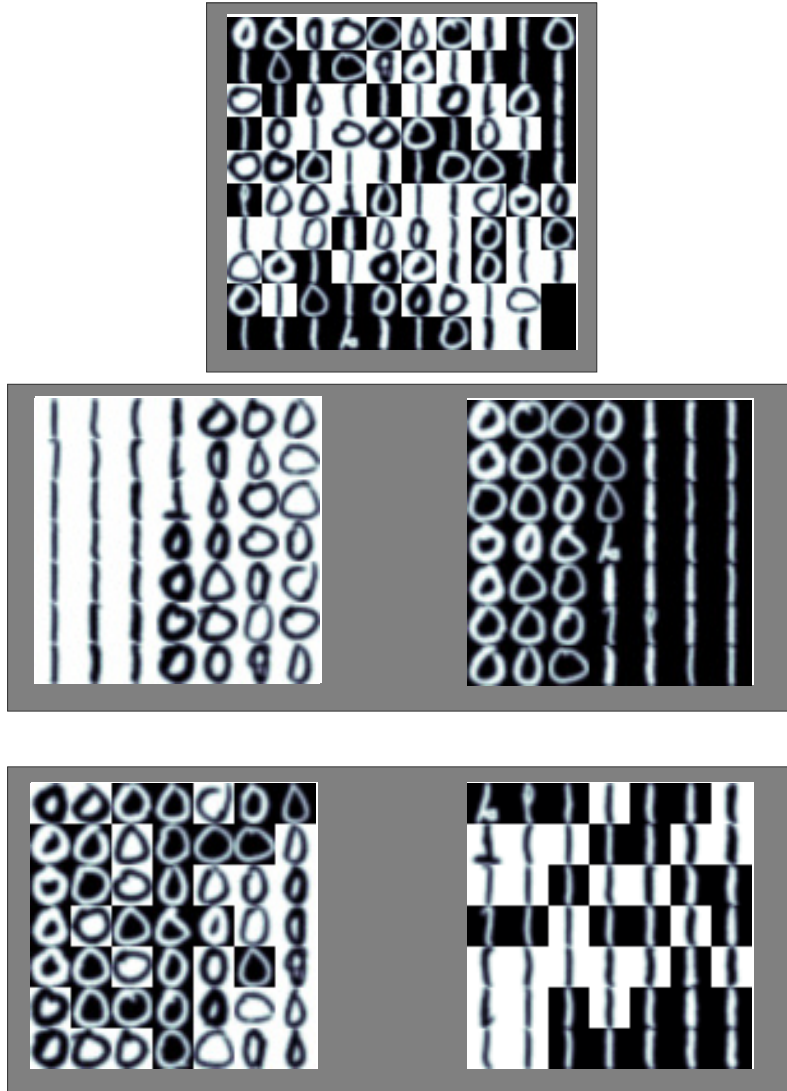


Figure 2: (Top) 98 Handwritten digits from the USPS dataset with randomly switched signs. (Middle) Spectral Clustering using a Gaussian Kernel ($\sigma = 22$) and two clusters. (Bottom) Spectral Clustering using the sign-invariant Gaussian Kernel ($\sigma = 22$) and two clusters.

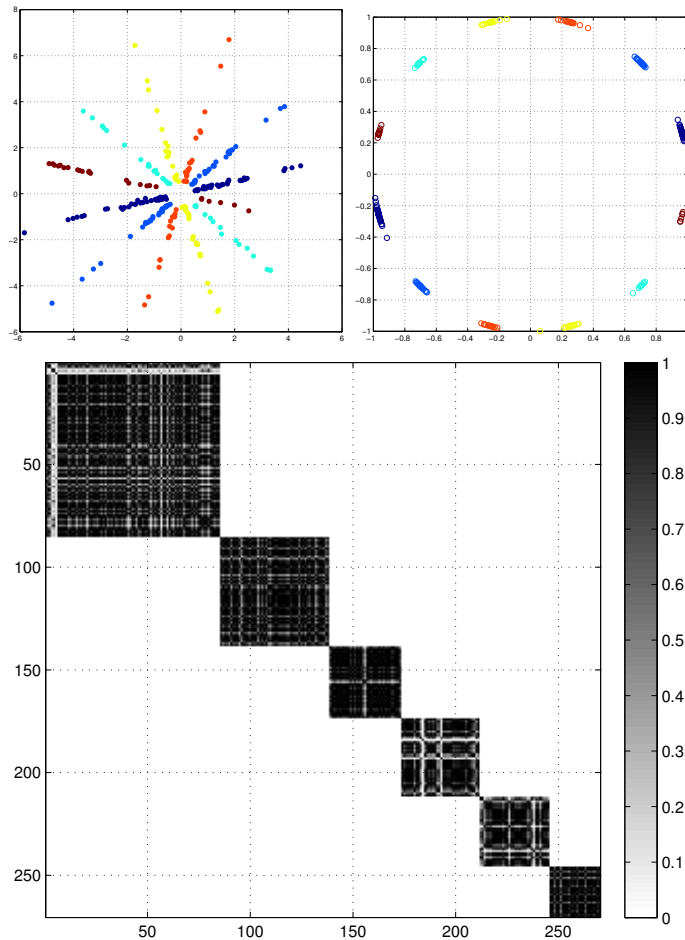


Figure 3: Here we applied spectral clustering using our sign-invariant Gaussian kernel (1) with $\sigma = 0.1$ and 6 clusters. (Top left panel) Clustering of sparse wavelet coefficients to solve an overcomplete ICA problem. There are six directions corresponding to six sources measured with two sensors. (Top right panel) Normalized data to obtain scaling invariance. (Bottom panel) Rearranged kernel matrix shows a clear block structure.

4.2. Clustering for overcomplete ICA

The main motivation for clustering with sign invariance comes from matrix factorization methods such as ICA (Hyvärinen et al., 2001) where the solutions are only determined up to scaling (which can easily be addressed by normalization) and sign or phase factors (which are more difficult to address otherwise). We note that the task of clustering of ICA components is important in group studies of EEG or MEG (Spadone et al., 2012).

Beyond the task of grouping ICA components, it is also possible to perform ICA by clustering of sparse representations of the observed mixtures obtained by e.g. short-time Fourier or wavelet-decompositions. In this scenario the clustering approach enables the identification and recovery of more sources than sensors, i.e. it solves an overcomplete ICA problem (Bofill and Zibulevsky, 2001, Chen et al., 1998, Hyvärinen et al., 2001, Meinecke et al., 2005, Zibulevsky and Pearlmutter, 2001).

In this example we use the “six-flutes” dataset of Bofill and Zibulevsky (2001) which consists of two instantaneous linear mixtures $x(t) = As(t)$ of recordings of six different notes

played with a flute. The mixing matrix A has dimensions (2×6) .

Applying a cosine packet tree (Chen et al., 1998) for sparsification and selecting 270 points with largest norm yields the dataset shown in the left panel of figure 3. The columns of the mixing matrix A correspond to 6 equally spaced directions in the two dimensional space of the mixed signals, i.e. there are six directions representing six sources measured with two sensors. Clustering this data is equivalent to a “blind” identification of the mixing matrix A up to scale and sign indeterminacies. In the bottom panel of figure 3 we show the entries of the kernel matrix computed with the sign-and-scale invariant Gaussian kernel. There is a clear block structure with six blocks corresponding to six sign-and-scale invariant clusters.

5. Conclusion

Algorithms for invariant pattern recognition have received continuous attention over the last decades.

In this paper we propose a theoretical framework for kernel methods that allows the systematic and constructive incorporation of algebraic invariance structure into a Mercer Kernel. In this manner known structural invariances can be implemented by applying the kernel trick twice: first a nonlinear inner kernel is constructed that hard codes the invariance and then a second nonlinear kernel is applied to the result. The additional computational load involved is negligible, but the gain in performance and meaningfulness of the learned result is substantial. While the proposed framework is general – we showed how to code a variety of invariances – also its potential practical applications are numerous; in fact, any kernel algorithm can be made invariant in this manner.

We have limited ourselves to the application of spectral clustering and demonstrated exemplarily how to incorporate sign invariance and sign-and-scaling invariance; our experiments have validated the usefulness of our approach.

We remark that being able to construct sign-invariant, or sign-and-scale invariant clustering algorithms is actually a pertinent issue in biomedical applications where decomposition methods such as PCA, ICA, CSP etc. yield components that are to be grouped with low computational cost, since clearly, combinatorial heuristic solutions are infeasible in the context of the now commonly available multi-channel imaging systems.

Here our novel algorithms will already be of practical use. Note, however, that the spirit of this contribution is mainly considered foundational and conceptual.

Future work will be devoted to incorporate even more general invariance structure and to the broader application of the proposed framework in the fields of Bioinformatics and Biomedical Engineering.

Acknowledgements

KRM and AZ gratefully acknowledge funding by DFG and BMBF. KRM thanks for partial funding by the National Research Foundation of Korea funded by the Ministry of Education, Science, and Technology in the BK21 program. This research was carried out at MFO, supported by FK’s Oberwolfach Leibniz Fellowship.

Appendix

In the following, we provide proofs for several of the statements in the corpus regarding invariances.

Quotient rings and quotient varieties In the following, commutative rings will always contain the one-element.

Definition 5.1. Let R be a commutative ring, and G a group acting on R . Then, we define

$$R^G := \{r \in R : g.r = r \text{ for all } g \in G\}.$$

and call R^G the invariant ring of R w.r.t G .

A classical characterization of the invariant ring is by its universal property:

Lemma 5.2. Let $\phi : S \rightarrow R$ be a homomorphism of rings, and G a group acting on R . Let $\iota : R^G \rightarrow R$ be the canonical embedding. If $g \circ \phi = \phi$ for all $g \in G$, then there exists a unique homomorphism $\varphi : S \rightarrow R^G$, such that $\iota \circ \varphi = \phi$.

Proof. It suffices to show that the image of ϕ is contained in R^G . But $g \circ \phi = \phi$ for all $g \in G$ implies $g.\phi(s) = \phi(s)$ for all $g \in G, s \in S$, which implies that $\phi(s) \in R^G$ for all $s \in S$, which proves the claim. \square

Definition 5.3. Let V, W be algebraic varieties, let G be a group acting on V . A homomorphism $f : V \rightarrow W$ is called G -invariant if $f \circ g = f$ for all $g \in G$.

We proceed with the geometric analogue of Lemma 5.2, which is the analogue universal property for the scheme-theoretic quotient, which we formulate in a form less general than usual in algebraic geometry, but which fits the specific setting outlined in the main corpus.

Proposition 5.4. Let $V \subseteq \mathbb{K}^n$ be an affine algebraic variety, let G be a group algebraically acting on V . Then, there is a family of algebraic invariant maps

$$q : V \rightarrow V/G \subseteq \mathbb{K}^I, \text{ i.e., } (q_i : \mathbb{K}^n \rightarrow \mathbb{K})_{i \in I},$$

such that for all varieties $W \subseteq \mathbb{K}^m$ and G -invariant homomorphisms $f : V \rightarrow W$ there is a unique homomorphism $g : V/G \rightarrow W$ such that $f = g \circ q$. Furthermore, I is countable, and if G is finite, then so is I .

Proof. By the algebra-geometry duality (contravariant equivalence of commutative rings and affine schemes), the homomorphism of varieties $f : V \rightarrow W$ is dual to a homomorphism of rings $\phi : S \rightarrow R$, with $W = \text{Spec} S$ and $V = \text{Spec} R$, and G canonically acting on R . By Lemma 5.2, ϕ factors uniquely as $\iota \circ \varphi = \phi$, into $\varphi : S \rightarrow R^G$, and the canonical quotient morphism $\iota : R^G \rightarrow R$. Taking $V/G := \text{Spec} R^G$ and $q := \text{Spec}(\iota), g := \text{Spec}(\varphi)$ yields the first claim on existence of q and factorization of f . For the second statement, note that V is contained in \mathbb{K}^n , thus R is finitely generated over \mathbb{K} , thus R is a \mathbb{K} -vector space of countable dimension. As a sub-vector space, R^G is also of countable dimension over \mathbb{K} . As R^G is generated by any vector space basis, it is countably generated over \mathbb{K} , thus I can be taken countable. Moreover, if G is finite, then by Hilbert's invariant theorem, R^G is finitely generated over \mathbb{K} as well. \square

We would like to note that if G is infinite, it can both happen that V/G is of finite type or not. For example, an infinite but linearly reductive G gives rise to a finite type quotient, as well as any G in the case $n = 1$ or 2 , see (Zariski, 1954). But V/G can also be not of finite type, as any counterexample to Hilbert's 14-th problem shows, such as the example of Totaro (2008) in the case $n = 16$ which can be invoked with a real or complex ground field \mathbb{K} .

The universal property of the quotient We proceed with the proof of Theorem 2 which is essentially a minor variant of Proposition 5.4 above and translates the universal property of the quotient to a setting more suitable for kernels.

Proof of Theorem 2. Consider first the case $\mathcal{F} = \mathbb{K}$.

Note that G acts canonically on the \mathbb{K} -vector space $C(W, \mathcal{F})$, with an vector space of invariant functions $C(W, \mathcal{F})^G$. Now, by the Stone-Weierstraß-Theorem, the algebraic/polynomical functions $\text{Pol}(W, \mathcal{F})$ are dense in the continuous functions $C(W, \mathcal{F})$. Taking the intersection with $C(W, \mathcal{F})^G$, it follows that the G -invariant algebraic/polynomical functions $\text{Pol}(W, \mathcal{F})^G$ are dense in $C(W, \mathcal{F})^G$.

Now (by categorical equivalence of evaluation homomorphisms and polynomials over the characteristic zero field \mathbb{K}), the space $\text{Pol}(W, \mathcal{F})$ is canonically isomorphic to the ring $R = \mathbb{K}[X_1, \dots, X_n]/I(W)$, while $\text{Pol}(W, \mathcal{F})^G$ is isomorphic to R^G . Note that $I(W) = I(Z)$, where Z is the Zariski closure of W in \mathbb{K}^n . Thus, $\text{Pol}(W, \mathcal{F})$ is isomorphic to $\text{Pol}(Z, \mathcal{F})$, and $\text{Pol}(W, \mathcal{F})^G$ is isomorphic to $\text{Pol}(Z, \mathcal{F})^G$, showing that in the above argumentation W can be replaced by the variety Z .

Proposition 5.4 yields, for every $f \in \text{Pol}(Z, \mathcal{F})^G$, a decomposition of the desired kind, thus by isomorphism for any $f \in \text{Pol}(W, \mathcal{F})^G$, thus by denseness for any $f \in C(W, \mathcal{F})^G$.

For general \mathcal{F} , one obtains the analogue statement by passing to a basis representation $\mathcal{F} = \text{span}(f_i, i \in I)$, yielding a decomposition $\text{Pol}(W, \mathcal{F}) = \bigoplus_{i \in I} \text{Pol}(W, \mathbb{K})$.

Note that the Hilbert space \mathcal{J} in the statement is therefore isomorphic to the dual of $\text{Pol}(W, \mathcal{F})^G$.

Isometric invariants

Lemma 5.5. Consider the orthogonal/unitary group $G = O(n)$ resp. $G = U(n)$ (consisting of orthogonal/unitary matrices) which acts on $V = \mathbb{K}^n$, i.e, $A \cdot x = A \cdot x$.

Then the invariant map is $q : V \rightarrow V/G \subseteq \mathbb{R}$ given by $x \mapsto \langle x, x \rangle$.

Proof. Observe that any $x \in \mathbb{K}^n$ is G -equivalent to $\|x\| \cdot e_1$, where e_1 is a standard normal basis vector, and the orbits of $\|x\| \cdot e_1$ are distinct. Alternatively, instantiate the diagonal invariant map from Lemma 5.6 on vectors $(x, 0)$. \square

Lemma 5.6. Consider the orthogonal/unitary group $G = O(n)$ resp. $G = U(n)$ (consisting of orthogonal/unitary matrices) which acts diagonally on $V = \mathbb{K}^n \times \mathbb{K}^n$, i.e, $A \cdot (x, y) = (A \cdot x, A \cdot y)$.

Then the invariant map is $q : V \rightarrow V/G \subseteq \mathbb{R} \times \mathbb{K} \times \mathbb{R}$ given by

$$(x, y) \mapsto (\langle x, x \rangle, \langle x, y \rangle, \langle y, y \rangle).$$

Proof. This is shown in Lemma 6.1.1 of (Crook, 2009). \square

Lemma 5.7. Consider the translational group $G = T(n) \cong \mathbb{K}^n$ which acts diagonally on $V = \mathbb{K}^n \times \mathbb{K}^n$, i.e, $v \cdot (x, y) = (x + v, y + v)$.

Then the invariant map is $q : V \rightarrow V/G \subseteq \mathbb{K}^n$ given by $(x, y) \mapsto (x - y)$.

Proof. The action is linear, thus the group quotient as variety equals the vector space quotient, which is the image of the linear map q . \square

Lemma 5.8. Consider the Euclidean group $G = \text{Euc}(n)$ generated by translations and orthogonal/unitary rotations which acts diagonally on $V = \mathbb{K}^n \times \mathbb{K}^n$.

Then the invariant map is $q : V \rightarrow V/G \subseteq \mathbb{R}$ given by $(x, y) \mapsto \|x - y\|^2 = \langle x - y, x - y \rangle$.

Proof. Recall that G is generated by the translations $T(n)$ and orthogonal/unitary matrices $H = O(n)$ resp. $H = U(n)$, that $T(n)$ is a normal subgroup of $G = \text{Euc}(n)$, and the corresponding factor group is $\text{Euc}(n)/T(n) = H$. Thus, $V/G = (V/T(n))/H$. By Lemma 5.7, there is a canonical quotient map

$$q_T : V \rightarrow V/T(n) = \mathbb{K}^n, (x, y) \mapsto (x - y),$$

thus $(V/T(n))/H = V/G = \mathbb{K}^n/H$. By Lemma 5.5, there is a further canonical quotient map

$$q_H : V/T(n) \rightarrow V/G = \mathbb{R}, x \mapsto \langle x, x \rangle = \|x\|^2,$$

which after composition $q := q_H \circ q_T$ yields the claim. An alternative proof can be obtained in analogy to that of Lemma 5.5, by observing and verifying in an elementary way that any $(x, y) \in \mathbb{K}^n \times \mathbb{K}^n$ is diagonally G -equivalent to $\|x - y\| \cdot e_1$, and two such vectors are distinct if and only if their absolute value is. \square

References

- Yoshua Bengio. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1):1–127, 2009. doi: 10.1561/2200000006. Also published as a book. Now Publishers, 2009.
- Pau Bofill and Michael Zibulevsky. Underdetermined blind source separation using sparse representations. *Signal Processing*, 81:2353–2362, 2001.
- Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pages 144–152. ACM Press, 1992.
- Olivier Chapelle and Bernhard Schölkopf. Incorporating invariances in non-linear support vector machines. In *Advances in neural information processing systems*, pages 609–616, 2001.
- Scott S. Chen, David L. Donoho, and Michael A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1998.
- Deborah Crook. *Polynomial Invariants of the Euclidean Group - Action on Multiple Screws*. PhD thesis, Victoria University of Wellington, 2009.
- Inderjit S. Dhillon, Yuqiang Guan, and Brian Kulis. Kernel k-means: spectral clustering and normalized cuts. In *Knowledge Discovery and Data Mining*, pages 551–556, 2004. doi: 10.1145/1014052.1014118.
- Bernard Haasdonk and Hans Burkhardt. Invariant kernel functions for pattern analysis and machine learning. *Machine learning*, 68(1):35–61, 2007.
- Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE*, 29(6):82–97, 2012.
- Jonathan J. Hull. A Database for Handwritten Text Recognition Research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16:550–554, 1994. doi: 10.1109/34.291440.

- Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent Component Analysis*. Wiley, 2001.
- Tony Jebara. Convex invariance learning. In *Artificial Intelligence and Statistics*, 2003.
- Robert Jenssen. Kernel entropy component analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):847–860, 2010.
- Hao Jiang and Wai-Ki Ching. Correlation kernels for support vector machines classification with applications in cancer data. *Computational and Mathematical Methods in Medicine*, 2012. doi: 10.1155/2012/205025.
- Risi Kondor. *Group Theoretical Methods in Machine Learning*. PhD thesis, Columbia University, 2008.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.
- Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- Marina Meila and Jianbo Shi. Learning segmentation by random walks. In *Advances in Neural Information Processing Systems 13*, pages 873–879. MIT Press, 2001.
- Frank Meinecke, Stefan Harmeling, and Klaus-Robert Müller. Inlier-based ICA with an application to super-imposed images. *International Journal of Imaging Systems and Technology (IJIST)*, 15(1):48–55, 2005.
- Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems 14*, pages 849–856. MIT Press, 2002.
- Bernhard Schölkopf and Alexander J Smola. *Learning with kernels*. MIT Press, 2002.
- Bernhard Schölkopf, Patrice Simard, Vladimir Vapnik, and Alexander Smola. Improving the accuracy and speed of support vector machines. *Advances in neural information processing systems*, 9:375–381, 1997.
- Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319, 1998.
- Patrice Y. Simard, Bernard Victorri, Yann A. LeCun, and John S. Denker. Tangent prop — a formalism for specifying selected invariances in an adaptive network. In J.E. Moody, S.J. Hanson, and R.P. Lippmann, editors, *Advances in Neural Information Processing Systems*, volume 4, pages 895–903, San Mateo, CA, 1992. Morgan Kaufmann.
- Sara Spadone, Francesco de Pasquale, Dante Mantini, and Stefania Della Penna. A k-means multivariate approach for clustering independent components from magnetoencephalographic data. *NeuroImage*, 62(3):1912 – 1923, 2012.
- Masashi Sugiyama, Gang Niu, Makoto Yamada, Manabu Kimura, and Hirotaka Hachiya. Information-maximization clustering based on squared-loss mutual information. *Neural Computation*, 26(1):84–131, 2014.

- Burt Totaro. Hilbert's 14th problem over finite fields and a conjecture on the cone of curves. *Compositio Mathematica*, 144(05):1176–1198, 2008.
- Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995.
- Ulrike von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.
- Christian Walder and Olivier Chapelle. Learning with transformation invariant kernels. In *Advances in Neural Information Processing Systems*, pages 1561–1568, 2007.
- Oscar Zariski. Interprétations algébriques-géométriques du quatorzième problème de Hilbert. *Bull. Sci. Math*, 78(2):155–168, 1954.
- Lihi Zelnik-manor and Pietro Perona. Self-tuning spectral clustering. In L.K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 1601–1608. MIT Press, 2005.
- Michael Zibulevsky and Barak A. Pearlmutter. Blind source separation by sparse decomposition in a signal dictionary. *Neural Computation*, 13:863–882, 2001.
- Alexander Zien, Gunnar Rätsch, Sebastian Mika, Bernhard Schölkopf, Thomas Lengauer, and K-R Müller. Engineering support vector machine kernels that recognize translation initiation sites. *Bioinformatics*, 16(9):799–807, 2000.