

Directional features in online handwriting recognition[☆]

Claus Bahlmann*

Siemens Corporate Research, Inc., 755 College Road East, Princeton, NJ 08540, USA

Received 21 November 2003; received in revised form 24 March 2005; accepted 11 May 2005

Abstract

The selection of valuable features is crucial in pattern recognition. In this paper we deal with the issue that part of features originate from *directional* instead of common linear data. Both for directional and linear data a theory for a statistical modeling exists. However, none of these theories gives an integrated solution to problems, where linear *and* directional variables are to be combined in a *single*, multivariate probability density function. We describe a general approach for a unified statistical modeling, given the constraint that variances of the circular variables are small. The method is practically evaluated in the context of our online handwriting recognition system *frog on hand* and the so-called *tangent slope angle* feature. Recognition results are compared with two alternative modeling approaches. The proposed solution gives significant improvements in recognition accuracy, computational speed and memory requirements.

© 2005 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.

Keywords: Feature selection; Directional data; Distribution on a circle; Multivariate semi-wrapped Gaussian distribution; Online handwriting recognition; UNIPEN online handwriting database

1. Introduction

In statistical pattern recognition the modeling of an abstract *feature space* with parametric probability density functions (PDFs) is very common. Often the Gaussian (or normal) PDF

$$p(\mathbf{x}) = \mathcal{N}_{\mu_{\mathbf{x}}, \Sigma_{\mathbf{x}}}(\mathbf{x}) = (|2\pi\Sigma_{\mathbf{x}}| \exp((\mathbf{x} - \mu_{\mathbf{x}})^T \Sigma_{\mathbf{x}}^{-1} (\mathbf{x} - \mu_{\mathbf{x}})))^{-1/2} \quad (1)$$

or a mixture of it is used to describe the probability density of a random vector $\mathbf{x} \in \mathbb{R}^F$. For *linear* data, i.e., data that are distributed on the real line \mathbb{R} , sensible definitions of the Gaussian parameters *mean* and *covariance* exist, namely

$$\mu_{\mathbf{x}} = \mathcal{E}[\mathbf{x}] \quad \text{and} \quad (2)$$

$$\Sigma_{\mathbf{x}} = \mathcal{E}[(\mathbf{x} - \mu_{\mathbf{x}})(\mathbf{x} - \mu_{\mathbf{x}})^T], \quad (3)$$

with $\mathcal{E}[\cdot]$ the expectation value of its argument.

[☆] Most of this work has been composed while the author was with the Computer Science Department, Albert-Ludwigs-University Freiburg, Germany.

* Tel.: +1 6097343616; fax: +1 6097346565.

E-mail address: claus.bahlmann@siemens.com.

However, not all data in real world applications are of this particular linear type. In some situations [1–5] data originate from *directions*. Directions, in general, may be visualized as points on the surface of a hypersphere, in two dimensions on the circumference of a circle. In the latter situation we will talk about circular data. A directional variable is inherently cyclic and the common statistical modeling used for linear data—including Eqs. (1)–(3)—are not appropriate here, as will be shown later.

Also in (online) handwriting recognition (HWR) we are faced with the problem of directional data, since a valuable feature of this particular circular type exists, namely the *tangent slope angle* θ . Details about θ will follow. The answer of many HWR systems [6–8] to the problem of modeling circular features with parametric PDFs is somewhat defensive. There, a direct modeling of θ is avoided by transforming it into the representation $(\cos \theta, \sin \theta)$. Contrary to θ , the quantities $\cos \theta$ and $\sin \theta$ themselves are not circular. In this respect, previous systems take $\cos \theta$ and $\sin \theta$ for a linear feature and model $(\cos \theta, \sin \theta)$ instead of θ as part of a linear feature vector by Eqs. (1)–(3).

However, there are some drawbacks in this strategy. *First*, the dimension of the resulting feature space is unnecessarily

increased (by one), since the strategy uses two dimensions to describe one degree of freedom. Among others, computational speed and memory requirements are negatively affected. *Second*, the $(\cos \theta, \sin \theta)$ representation includes high dependencies: the feature pair $(\cos \theta, \sin \theta)$ lies on the circumference of the unit circle. In situations when parametric basis functions are used to model probability densities, those dependencies have to be addressed. This, however, is very difficult to achieve since basis functions generally assume a restrictive shape.

On the other hand, statisticians have developed methodologies that deal with directional data, in last decades especially influenced by the work of Mardia [1]. Remarkably, it seems that this work has not found its way into the pattern and handwriting recognition community.

This unfortunate situation might be explained by the following reason. While the directional methodologies developed so far are well suited for distributions of solely directional variables (as they appear in physical, biological, geological, geographical or medical applications) they still lack a clear description how they can be advantageously applied for multivariate PDFs of both linear *and* directional variables. Contrary to the applications listed above, where mostly a two- or three-dimensional space of spatial or physical coordinates is to be modeled, in pattern recognition we are faced with the problem of modeling an abstract feature space. Mostly, in these cases the situation of mixed linear and directional data exist.

The intention of this paper is twofold. First, statistics of directional data shall be introduced to the pattern and handwriting recognition community. Second, we want to propose an approach to the already addressed problem of integrating directional *and* linear data into a multivariate PDF that aims to model an abstract feature space: the *multivariate semi-wrapped Gaussian* PDF. In order to reduce analytic complexity, considerations are confined to a special constraint: it is assumed that the circular feature has a small variance.

Note that the emphasis of this paper concerns a generic approach to the modeling of semi-directional data with multivariate semi-wrapped Gaussian modeling rather than the description of an online HWR system. However, to put the theory into context, the effectiveness of the *multivariate semi-wrapped Gaussian* modeling will be demonstrated and evaluated in the context of our online HWR system *frog on hand*. For this, a brief description of *frog on hand* is included in this paper; a more thorough presentation and comparison in the context to other state-of-the-art HWR systems is given elsewhere [9–11].

We shall start with a brief description of our online handwriting recognition application in the section that follows. Section 3 first motivates and reviews basic concepts from the statistics of directional data. Section 4 introduces a distribution for directional data (the *wrapped Gaussian distribution*) and proposes an approximation of it. Following this, Section 5 introduces the formulation of a *multivariate semi-wrapped Gaussian distribution* and transfers the deliberations made

about approximation issues from Section 4 to this construct. In Section 6 experimental results of the proposed concept in the context of our online handwriting recognition *frog on hand* [9] and the UNIPEN [12] online handwriting database are presented. Section 7 concludes this contribution.

2. Online handwriting recognition

Handwriting recognition (HWR) is a task of transforming a language represented in its spatial form of graphical marks into its symbolic representation [13]. *Online* HWR refers to the situation where the recognition is performed concurrently to the writing process.

In the following, we give a brief overview of our online handwriting recognition system *frog on hand*, which is an acronym for “*f*reiburg *r*ecognition of *o*nline *h*andwriting”. For details please refer to original work [10,9,14,15]. Readers with no particular interest in the handwriting recognition application can skip the rest of this section and directly proceed to Section 3.

2.1. Data and features

Online HWR is a prominent example for a pattern recognition problem, where data are of variable-length sequential type. Online handwriting data is typically represented by the sequence

$$\not\phi = [\mathbf{p}_1, \dots, \mathbf{p}_N],$$

where the vectors $\mathbf{p}_i = (x_i, y_i)^T$ denote the horizontal and vertical coordinates, sampled from the writer’s pen movement. Indeed, this representation is much different from offline HWR, where data are typically image matrices.

In this paper we focus on the situation where a writing represents an isolated character.

It is widely acknowledged that the selection of meaningful features is crucial in pattern and handwriting recognition. In the *frog on hand* feature selection, the writing $\not\phi = [\mathbf{p}_1, \dots, \mathbf{p}_N]$ is transformed into a sequence $\ell = [\mathbf{t}_1, \dots, \mathbf{t}_N]$ of feature vectors $\mathbf{t}_i = (t_{i1}, \dots, t_{iF})^T \in \mathbb{R}^F$.

We experienced best performance with the following local features:

Normalized horizontal and vertical coordinates $t_{i1} = \tilde{x}_i = (x_i - \mu_x)/\sigma_x$ and $t_{i2} = \tilde{y}_i = (y_i - \mu_y)/\sigma_y$, are the pen coordinates normalized by the sample mean $\boldsymbol{\mu} = (\mu_x, \mu_y)^T = 1/N \sum_{i=1}^N \mathbf{p}_i$ and (vertical) y standard deviation $\sigma_y = \sqrt{1/(N-1) \sum_{i=1}^N (\mu_y - y_i)^2}$ of the character’s sample points.

Tangent slope angle $t_{i3} = \theta_i = \arg((x_{i+1} - x_{i-1}) + J \cdot (y_{i+1} - y_{i-1}))$ with $J^2 = -1$, the imaginary unit and “arg” the phase of the complex number above, is an approximation of the tangent slope angle at point i .

To summarize, a feature vector sequence is defined as $\ell = [\mathbf{t}_1, \dots, \mathbf{t}_N]$, each vector of it as $\mathbf{t}_i = (\tilde{x}_i, \tilde{y}_i, \theta_i)^T$. Fig. 1(a)

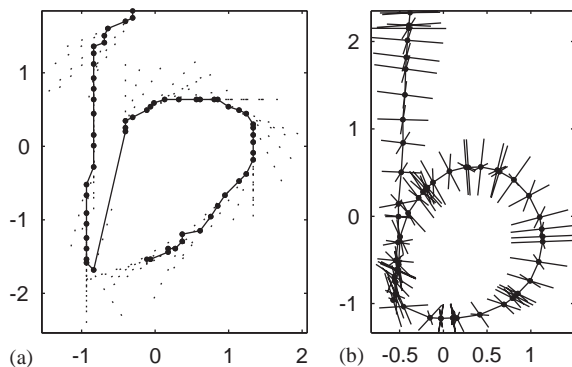


Fig. 1. (a) Shows an illustration of a feature sequence ℓ and the features \tilde{x}_i , \tilde{y}_i and θ_i : \tilde{x}_i and \tilde{y}_i are plotted according to their value in the \tilde{x} - \tilde{y} -plane. The dotted lines illustrate θ_i by the direction of the tangent. In (b) the reader can see a connected sequence of dots, each one with two additional lines attached. As a reference model is represented by a sequence of Gaussian PDFs $\beta_j(\mathbf{x}) = \mathcal{N}_{\mu_j, \Sigma_j}(\mathbf{x})$, $j = 1, \dots, N_{\mathcal{R}}$, each dot illustrates the mean of the Gaussian and the lines the covariance matrix. The direction of the two lines match the projection of the first two eigenvectors onto the \tilde{x} - \tilde{y} -plane, their length the square root of the corresponding eigenvalues. Thus, the lines indicate the orientation and width of the projected Gaussian. Note that α_j is not illustrated within this figure.

gives a graphical illustration of the feature representation of a character “b”.

Note that the feature θ_i codes a direction in two dimensions and thus is an example for circular data.

2.2. Classification framework

We give a brief summary of *frog on hand*'s generative classification framework CS-DTW, which abbreviates the term *cluster generative statistical dynamic time warping* [9]. CS-DTW is a general, scalable training and classification framework for variable-sized, sequential data. It supplements dynamic time warping (DTW) [16] by a holistic combination of cluster analysis and statistical sequence modeling.

Similar to hidden Markov modeling (HMM) [16], CS-DTW statistically represents a reference model for a handwritten character by a sequence

$$\mathcal{R} = [(\alpha_1, \beta_1), \dots, (\alpha_{N_{\mathcal{R}}}, \beta_{N_{\mathcal{R}}})], \quad (4)$$

where $\alpha_j(\Delta\phi)$ statistically models transitions $\Delta\phi$ between sequence elements and $\beta_j(\mathbf{x}) = \mathcal{N}_{\mu_j, \Sigma_j}(\mathbf{x})$ the PDF of a feature vector $\mathbf{x} \in \mathbb{R}^F$ by a unimodal multivariate Gaussian, each for the sequence index j . $N_{\mathcal{R}}$ is the length of \mathcal{R} .

The particularity of CS-DTW compared to other HMM based HWR classifiers is an advanced, statistical modeling of explicit *sub-classes*. A sub-class corresponds to a “compact” cluster in the feature space of training examples. In the HWR context, it is meant to model a particular style or shape of a character, which is also named *allograph*.

In this respect, a CS-DTW classifier for an L -class problem comprises a set $\mathfrak{R} = \{\mathcal{R}^{lk}\}_{l \in \{1, \dots, L\}, k \in \{1, \dots, K_l\}}$ of allograph

reference models, with l the class, k the sub-class index and K_l the number of sub-classes in class l . Each \mathcal{R}^{lk} is a generative statistical model of the type of Eq. (4) for one particular sub-class. A graphical illustration for an example of \mathcal{R} —a reference model of a character “b”—is shown in Fig. 1(b). There, the sequence of the Gaussian PDFs are represented by a connected sequence of dots, each one with two additional lines attached. Dots and lines illustrate mean and covariance of the underlying Gaussians. For details, it is referred to the figure caption.

The allograph reference models \mathcal{R}^{lk} are automatically generated by the CS-DTW training from data with class labels in two steps. Step *one* employs an unsupervised, *agglomerative hierarchical clustering* algorithm that aims to reveal clusters, each of them corresponding to a sub-class. The diversity of the hierarchical clustering can be controlled by the use of two threshold parameters. Step *two* employs the *Viterbi training* in order to estimate the statistical parameters of Eq. (4) for each cluster.

The CS-DTW classification of an unknown test pattern ℓ is solved by computing

$$\hat{l} = \operatorname{argmin}_{l \in \{1, \dots, L\}} \min_{k \in \{1, \dots, K_l\}} \{D^*(\ell, \mathcal{R}^{lk})\}, \quad (5)$$

where D^* is a specifically adopted distance, based on the maximum-likelihood (ML) principle, Markov assumptions and dynamic programming.

In order to account for a holistic treatment, clustering and statistical modeling are embedded in a single feature space and use a closely related dissimilarity measure (cf. the original publication [9] for details).

3. Statistics of directional data

Directional data may be visualized as points on the surface of a hypersphere, in two dimensions on the circumference of a circle. Fig. 2 illustrates an example. In Fig. 2(a), data points of a directional variable $\vartheta \in (-\pi, \pi]$ are shown as white dots on the unit circle.

One problem, which arises when directional data is used in combination with “conventional”, i.e., linear statistics, is illustrated in the following.

3.1. Linear statistics and directional data

Consider a linear variable x and a transformation $t_s(x) = \tilde{x} = x - v$. The transformation t_s represents a shift of the coordinate system origin. Valuable properties of the statistics with linear data, in particular of *mean* and *variance*, can be expressed by the equations

$$\mu_{\tilde{x}} = \mu_x - v, \quad (6)$$

$$\sigma_{\tilde{x}}^2 = \sigma_x^2. \quad (7)$$

Eq. (6) implies that the relative position of the mean remains invariant under a shift of the coordinate system

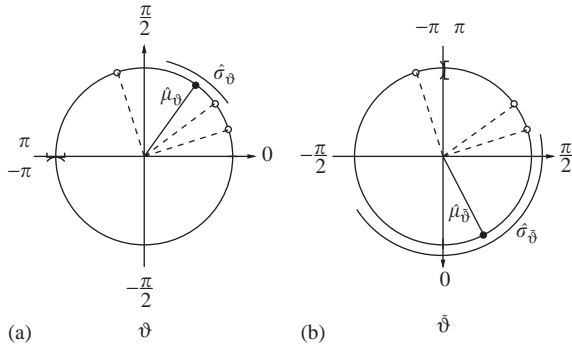


Fig. 2. Data points of a circular variable are shown as white dots on the unit circle. This figure illustrates that linear definitions of mean and variance violate the invariance of location and variance for directional data under a shift of zero direction $\tilde{\vartheta} = (\vartheta - \nu) \bmod 2\pi$: (a) For the example observations $\Theta = \{0.1\pi, 0.2\pi, 0.6\pi\}$, unbiased ML estimates for mean and standard deviation can be calculated to $\hat{\mu}_\vartheta = 0.3\pi$ (corresponding to the black dot) and $\hat{\sigma}_\vartheta \approx 0.26\pi$ (corresponding to the length of the arc). (b) The observations are shifted to $\tilde{\Theta} = \{0.6\pi, 0.7\pi, -0.9\pi\}$, corresponding to $\nu = -0.5\pi$. In the figure this corresponds to a rotation of the coordinate axes about 0.5π clockwise. Estimates for mean and standard deviation of the data points in the new coordinates can be calculated to $\hat{\mu}_{\tilde{\vartheta}} \approx 0.13\pi$ and $\hat{\sigma}_{\tilde{\vartheta}} \approx 0.90\pi$. Obviously, $\hat{\mu}_{\tilde{\vartheta}} \neq (\hat{\mu}_\vartheta - \nu) \bmod 2\pi$ and $\hat{\sigma}_{\tilde{\vartheta}}^2 \neq \hat{\sigma}_\vartheta^2$, thus neither the location nor the variance are invariant with respect to a shift of the origin.

origin. Eq. (7) refers to the invariance of the variance. In other words, the validity of Eqs. (6) and (7) guarantees a statistical behavior which is essentially independent of chosen coordinate system.

Now consider a circular variable ϑ . For ϑ , an addition “ $a+b$ ” becomes “ $(a+b) \bmod 2\pi$ ”. Here and in the remaining part, we assume a period of 2π and adopt the convention of angles represented in the interval $(-\pi, \pi]$. Note that under this assumption the mod operator also maps to $(-\pi, \pi]$. Let the variables μ_ϑ^c and V_ϑ^c denote the circular counterparts of mean and variance. Reasonable definitions for μ_ϑ^c and V_ϑ^c should have a similar behavior as Eqs. (6) and (7) under a shift of the zero direction which is expressed by $\tilde{\vartheta} = (\vartheta - \nu) \bmod 2\pi$. In this respect, equivalent invariances for a circular variable are

$$\mu_{\tilde{\vartheta}}^c = (\mu_\vartheta^c - \nu) \bmod 2\pi, \quad (8)$$

$$V_{\tilde{\vartheta}}^c = V_\vartheta^c. \quad (9)$$

However, it can easily be verified that with the linear definitions of mean and variance, given in Eqs. (2) and (3), the desired invariance is violated, i.e.,

$$\mu_{\tilde{\vartheta}} \neq (\mu_\vartheta - \nu) \bmod 2\pi \quad (10)$$

$$\sigma_{\tilde{\vartheta}}^2 \neq \sigma_\vartheta^2 \quad (11)$$

in general. Fig. 2 gives an example for this misbehavior, employing a simple set of circular observations

$\Theta = \{0.1\pi, 0.2\pi, 0.6\pi\}$ (cf. Fig. 2(a)) and $\tilde{\Theta} = \{0.6\pi, 0.7\pi, -0.9\pi\}$ (cf. Fig. 2(b)), which corresponds to $\nu = -0.5\pi$. For these observations, unbiased maximum likelihood (ML) estimates for mean and variance can be computed to $\hat{\mu}_\vartheta = 0.3\pi$, $\hat{\sigma}_\vartheta \approx 0.26\pi$, $\hat{\mu}_{\tilde{\vartheta}} \approx 0.13\pi$ and $\hat{\sigma}_{\tilde{\vartheta}} \approx 0.90\pi$, which are obviously not in agreement with Eqs. (8) and (9). In the figure, values of mean and variance are illustrated by the location of the black dot and the length of the accompanying arc, respectively.

As a concluding remark, we note that for circular data the linear definitions of mean and variance are highly dependent on the zero direction, which is an inappropriate behavior and demands for a suitable handling.

Physicists and statisticians have developed a methodology for dealing with statistics of directional data. Original publications lead back to the early 20th century [17–20]. Most recent comprehensive description can be found in the book of Mardia [1]. We briefly summarize the basics in the following.

3.2. Circular mean direction and circular variance

Assume a circular random variable ϑ with a PDF $p(\vartheta)$. In agreement with standard statistical properties, the PDF should satisfy $p(\vartheta) \geq 0$ and $\int_{-\pi}^{\pi} p(\vartheta) d\vartheta = 1$. Mardia [1] represents ϑ as a complex number $e^{J\vartheta}$ (with $J^2 = -1$) and employs the notation of *circular mean direction* μ_ϑ^c and *circular variance* V_ϑ^c . They are defined by

$$\rho_\vartheta e^{J\mu_\vartheta^c} = \mathcal{E}[e^{J\vartheta}] \quad (12)$$

with

$$V_\vartheta^c = 1 - \rho_\vartheta. \quad (13)$$

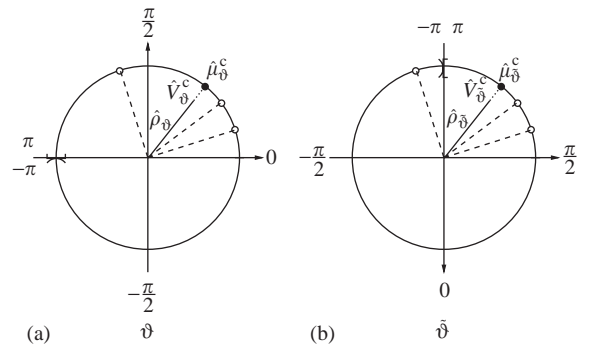


Fig. 3. For the same observations Θ (a) and $\tilde{\Theta}$ (b) as in Fig. 2. This figure shows estimates of the circular mean direction $\hat{\mu}_\vartheta^c$, the resultant length $\hat{\rho}_\vartheta$ and the circular variance \hat{V}_ϑ^c . $\hat{\mu}_\vartheta^c$ corresponds to the phase of the complex number associated to the black dot, $\hat{\rho}_\vartheta$ to the solid line and $\hat{V}_\vartheta^c = 1 - \hat{\rho}_\vartheta$ to the dotted line towards $\hat{\mu}_\vartheta^c$. The figures also express the validity of Eqs. (8) and (9).

The quantity ρ_ϑ is called the *resultant length*. Figuratively speaking, μ_ϑ^c is the expected phase and ρ_ϑ the expected length of $e^{J\vartheta}$. $V_\vartheta^c \in [0, 1]$ measures the amount of *dispersion*.

It can be shown [1] that, contrary to the linear definitions of mean and variance, μ_ϑ^c and V_ϑ^c fulfill the desired invariance of Eqs. (8) and (9) and can be utilized as suitable counterparts for the linear mean and variance. Fig. 3 graphically motivates this invariance, and further gives a figurative interpretation of μ_ϑ^c , V_ϑ^c , and ρ_ϑ estimates for the same samples as used in Fig. 2.

4. Wrapped Gaussian distribution

Based on μ_ϑ^c and V_ϑ^c , Mardia describes two *circular normal distributions* that should serve as appropriate substitutes for the univariate linear normal distribution. One is the *wrapped Gaussian (or normal) distribution*, and the other the *von Mises distribution*. Both have particular benefits and drawbacks compared to each other. Among others, the wrapped Gaussian has theoretical advantages and the von Mises distribution has practical benefits including the parameter estimation [1, Section 3.4.10]. However, it can be shown that they can be made to approximate each other closely. In this respect, a concrete assessment for one of the alternatives is practically not too restrictive.

As previously explained, our aim is to set up a *multivariate* Gaussian distribution of both linear *and* circular variables. In this context it appears that due to its apparent closeness to the linear Gaussian, the wrapped Gaussian (a definition will follow shortly) is the more natural choice for the present problem. Thus, in spite of the practical drawbacks in the parameter estimation, the wrapped Gaussian distribution, which will briefly be reviewed in the remaining part of this section, has been chosen.

4.1. General wrapped distribution

Any given PDF $p(x)$ of a linear variable x on the line can be “wrapped” around the circumference of a circle of unit radius. That is, the PDF $p_w(\vartheta)$ of the wrapped variable

$$\vartheta = x_w = x \bmod 2\pi \quad \in (-\pi, \pi] \quad (14)$$

is

$$p_w(\vartheta) = \sum_{k=-\infty}^{\infty} p(\vartheta + 2\pi k). \quad (15)$$

4.2. Wrapped Gaussian distribution

In particular, for $p(x)$ being a univariate Gaussian distribution $\mathcal{N}_{\mu_x, \sigma_x}(x)$ the *wrapped univariate Gaussian*

distribution is defined as

$$\begin{aligned} \mathcal{N}_{\mu_\vartheta^c, V_\vartheta^c}^w(\vartheta) \\ = \sum_{k=-\infty}^{\infty} \left(2\pi\sigma_x^2 \exp\left(\frac{(\vartheta - \mu_x + 2\pi k)^2}{\sigma_x^2}\right) \right)^{-1/2}. \end{aligned} \quad (16)$$

It can be shown [1] that for the circular mean direction μ_ϑ^c and the circular variance V_ϑ^c the equations

$$\mu_\vartheta^c = \mu_x \bmod 2\pi \quad \text{and} \quad (17)$$

$$\sigma_x^2 = -2 \log(1 - V_\vartheta^c) \quad (18)$$

hold. Further, $\mathcal{N}_{\mu_\vartheta^c, V_\vartheta^c}^w(\vartheta)$ is unimodal (i.e., has a single local maximum) and symmetric about μ_ϑ^c . With the relations of Eqs. (17) and (18) in mind we can use the notations of $\mathcal{N}_{\mu_\vartheta^c, V_\vartheta^c}^w$ and $\mathcal{N}_{\mu_x, \sigma_x}^w$ interchangeably in the rest of the paper.

Fig. 4(a) shows an example of the wrapped Gaussian distribution with parameters $\mu_x = \pi/2$ and $\sigma_x = 1.50$. The dashed lines show three contributing Gaussian terms, corresponding to $k = -1$, $k = 0$ and $k = 1$ of Eq. (16), the solid line shows their superposition.

4.3. An approximation to the wrapped Gaussian distribution

The wrapped Gaussian and von Mises distributions have been successfully applied to a variety of problems [1]. These problems can be assigned to one of the following categories:

- (1) The PDF is one-dimensional and the random variable corresponds to a direction in two dimensions.
- (2) The PDF is two-dimensional and the random variables correspond to a direction in three dimensions.

However, contrary to these problems, in many pattern recognition applications we are faced with the situation that a multivariate (> 2) PDF is to be modeled where only one (or a few) dimension(s) correspond to circular and the rest to linear variables. A suitable transfer of the mentioned directional distributions (the wrapped Gaussian or the von Mises distribution) to these “semi-directional” situations is not straightforward. In order to cope with this difficulty the remaining deliberations will be confined to an approximation of the wrapped Gaussian distribution: it is assumed that the wrapped Gaussian can be approximated by only one, but the most meaningful wrap of it. It will turn out in Section 5 that with this confinement a semi-directional PDF can be modeled directly.

In this respect, the following considerations shall be restricted to situations in which mainly one wrap of Eq. (16) contributes to the computation of values $\mathcal{N}_{\mu_\vartheta^c, V_\vartheta^c}^w(\vartheta)$. As can be verified from Fig. 4(a) this is the case, if the overlap of neighboring Gaussian wraps is negligible. In this case, it is permissible to approximate $\mathcal{N}_{\mu_\vartheta^c, V_\vartheta^c}^w(\vartheta)$ by mostly

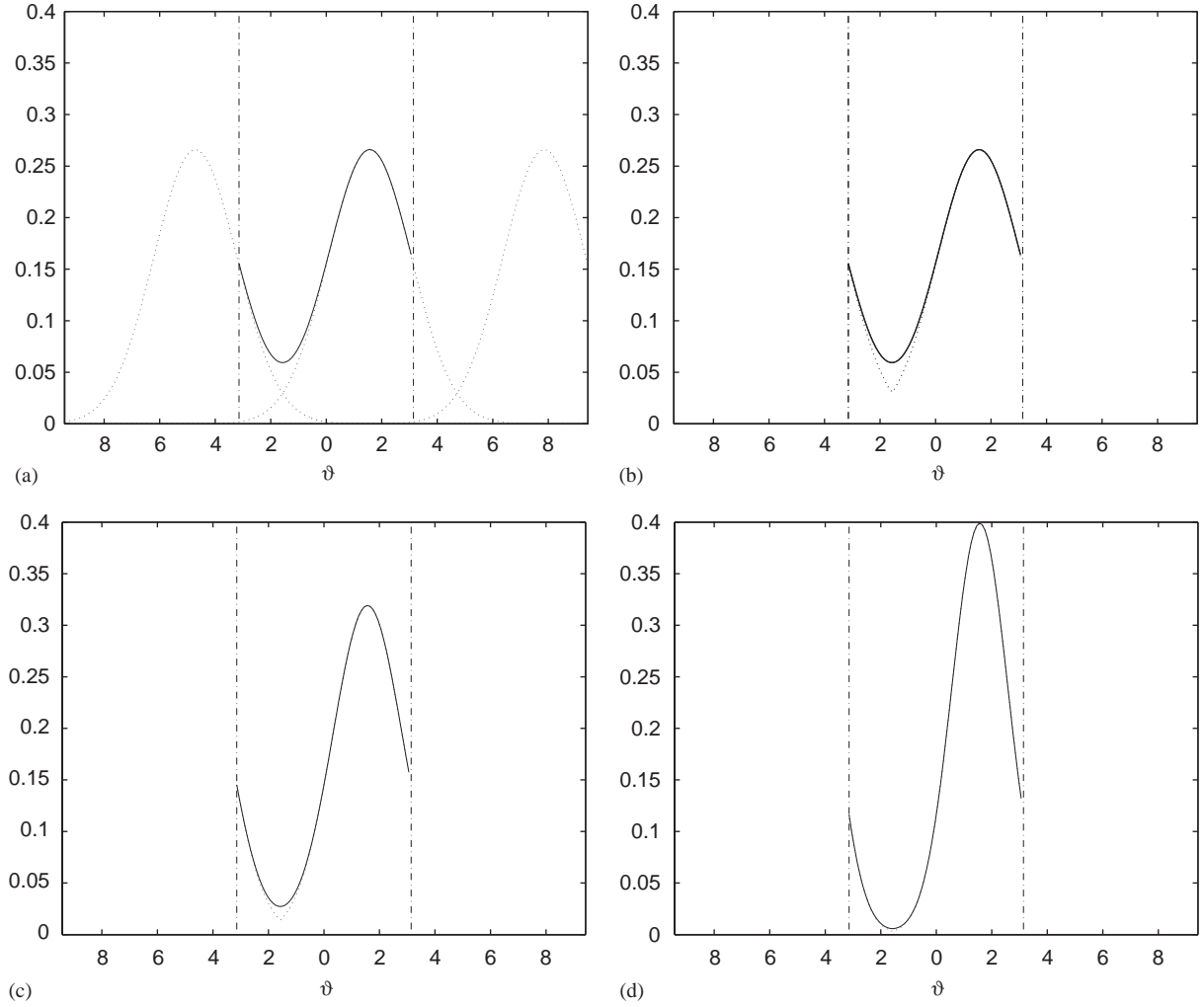


Fig. 4. (a): An example of the wrapped Gaussian distribution $\mathcal{N}_{\mu_x, \sigma_x}^w(\vartheta)$ with parameters $\mu_x = \pi/2$ and $\sigma_x = \sqrt{-2 \log(1 - V_\vartheta^c)} = 1.50$ (solid line). The dashed lines show three contributing terms in Eq. (16), corresponding to $k = -1, k = 0$ and $k = 1$. (b)–(d): Wrapped Gaussian $\mathcal{N}_{\mu_x, \sigma_x}^w(\vartheta)$ (solid line) together with their approximation $\mathcal{N}_{\mu_x, \sigma_x}^{aw}(\vartheta)$ (dotted line) with parameters $\mu_x = \pi/2$ and $\sigma_x \in \{1.50, 1.25, 1.00\}$. It can be seen that for $\sigma_x = 1.00$ approximation errors are small.

contributing wrap for $\vartheta \in (-\pi, \pi]$. This approximation can be summarized by the formulation

$$\mathcal{N}_{\mu_\vartheta^c, V_\vartheta^c}^{aw}(\vartheta) = \left(2\pi\sigma_x^2 \exp\left(\frac{((\vartheta - \mu_x) \bmod 2\pi)^2}{\sigma_x^2}\right) \right)^{-1/2}, \quad (19)$$

where the upper index “aw” shall indicate the term “approximated wrapped”. Again, the notations $\mathcal{N}_{\mu_\vartheta^c, V_\vartheta^c}^{aw}$ and $\mathcal{N}_{\mu_x, \sigma_x}^{aw}$ shall be used interchangeably, with Eqs. (17) and (18) giving the transformations between the respective parameters.

Figs. 4(b)–(d) compare $\mathcal{N}_{\mu_x, \sigma_x}^w(\vartheta)$ and $\mathcal{N}_{\mu_x, \sigma_x}^{aw}(\vartheta)$ for the particular parameter settings $\mu_x = \pi/2$ and $\sigma_x \in \{1.50, 1.25, 1.00\}$. For $\sigma_x = 1.50$ and $\sigma_x = 1.25$ the reader can verify approximation errors, centered at the anti-mode of $\mathcal{N}_{\mu_x, \sigma_x}^w(\vartheta)$. However, for $\sigma_x = 1.00$ only small deviations of the two functions can be found.

Thus, it will be assumed in the remainder of this contribution that for $\sigma_x \lesssim 1$ errors were small, if we use $\mathcal{N}_{\mu_x, \sigma_x}^{aw}(\vartheta)$ instead of $\mathcal{N}_{\mu_x, \sigma_x}^w(\vartheta)$ to model the PDF of a circular variable. A quantitative statement with respect to the approximation error can be specified in terms of the integrated error

$$E_{\text{int}}(\sigma_x) = \int_{-\pi}^{\pi} |\mathcal{N}_{\mu_x, \sigma_x}^{aw}(\vartheta) - \mathcal{N}_{\mu_x, \sigma_x}^w(\vartheta)| d\vartheta, \quad (20)$$

which corresponds to the area between the solid and the upper dotted lines in Fig. 4(a). If only directly adjacent wraps have a considerable overlap (which is the case for small variances), the mentioned area is approximately equivalent to the area below the lower dotted lines. Further, as the intersection of two adjacent wraps is at the position $(\mu_x - \pi) \bmod 2\pi = (\mu_x + \pi) \bmod 2\pi$, $E_{\text{int}}(\sigma_x)$ corresponds to the area of $\mathcal{N}_{\mu_x, \sigma_x}(x)$ that falls outside the interval $(\mu_x - \pi, \mu_x + \pi]$. Due to the symmetry of \mathcal{N} , the integrated error for $\sigma_x = 1$

can be computed to

$$E_{\text{int}}(\sigma_x = 1) = 2 \int_{-\infty}^{-\pi} \mathcal{N}_{0,1}(x) dx = \text{erfc}\left(\frac{\pi}{\sqrt{2}}\right) \approx 0.001721 \quad (21)$$

with

$$\text{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^{\infty} \exp(-t^2) dt, \quad (22)$$

the complementary error function.

4.4. Parameter estimates

In statistical pattern recognition the estimation of the parameters μ_{ϑ}^c and V_{ϑ}^c for $\mathcal{N}_{\mu_{\vartheta}^c, V_{\vartheta}^c}^{\text{aw}}(\vartheta)$ from a set $\Theta = \{\vartheta^{(1)}, \dots, \vartheta^{(M)}\}$ of circular observations is of practical importance.

A maximum likelihood estimate $\hat{\mu}_{\vartheta}^c$ can rather straightforwardly be computed. Mardia [1] derives the formula

$$\hat{\mu}_{\vartheta}^c = \arg\left(\frac{1}{M} \sum_{m=1}^M e^{J\vartheta^{(m)}}\right). \quad (23)$$

An estimate for V_{ϑ}^c can be obtained similarly by

$$\hat{V}_{\vartheta}^c = 1 - \left\| \frac{1}{M} \sum_{m=1}^M e^{J\vartheta^{(m)}} \right\|. \quad (24)$$

However, when the assumption $\sigma_x \lesssim 1$ holds, another approximative solution is valid. Then, as argued above, $\mathcal{N}_{\mu_x, \sigma_x}^w(\vartheta)$ becomes close to $\mathcal{N}_{\mu_x, \sigma_x}^{\text{aw}}(\vartheta)$. Under this assumption, one can think of $\mathcal{N}_{\mu_x, \sigma_x}^{\text{aw}}$ being a single Gaussian centered at $\hat{\mu}_{\vartheta}^c$ (corresponding to only one wrap) and only a small accuracy is sacrificed with a linear like estimate for V_{ϑ}^c via σ_x by means of

$$\hat{\sigma}_x^2 \approx \frac{1}{M-1} \sum_{m=1}^M ((\vartheta^{(m)} - \hat{\mu}_{\vartheta}^c) \bmod 2\pi)^2. \quad (25)$$

Note that Eq. (25) employs the circular mean direction estimate $\hat{\mu}_{\vartheta}^c$ instead of the linear mean estimate $\hat{\mu}_{\vartheta}$. As it has been shown in Section 3.1, the latter is inappropriate for directional data.

In the present case, the advantage of Eq. (25) over Eq. (24) is that it can be straightforwardly extended when covariances of directional and linear data are to be computed (cf. Section 5.5).

5. Multivariate semi-wrapped Gaussian distribution

In the following, the formulation of a combination of wrapped and non-wrapped Gaussian distributions for multivariate situations will be introduced. The resulting distribu-

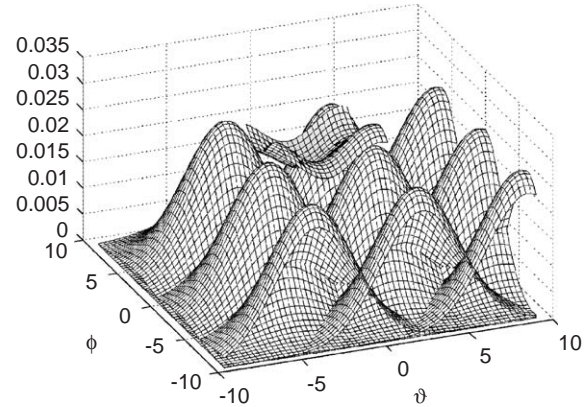


Fig. 5. An example for a bivariate wrapped PDF $\mathcal{N}_{\mu_x, \Sigma_x}^w$ is shown by the small patch above the (comparably small) interval $(-\pi, \pi] \times (-\pi, \pi]$. In correspondence with Fig. 4 (a), the figure additionally gives a plot of nine summands of Eq. (26), corresponding to $k_1 = -1, 0, 1$ and $k_2 = -1, 0, 1$ in Eq. (26). The parameter settings in this figure are $\mu_x = (2.5, -2.0)^T$ and $\Sigma_x = \begin{bmatrix} 6 & 0.7 \\ 0.7 & 6 \end{bmatrix}$.

tion will be formulated as *multivariate semi-wrapped Gaussian distribution*. To start with, the *multivariate wrapped distribution* and the *multivariate semi-wrapped distribution* shall be defined first.

5.1. Multivariate wrapped distribution

The concept of a univariate wrapped distribution can be extended to the multivariate context by an extension of the simple sum in Eq. (15) to a number of F sums that cover all dimensions in the feature space:

$$p_w(\mathbf{x}) = \sum_{k_1=-\infty}^{\infty} \cdots \sum_{k_F=-\infty}^{\infty} p(\mathbf{x} + 2\pi k_1 \mathbf{e}_1 + \cdots + 2\pi k_F \mathbf{e}_F). \quad (26)$$

In this equation, $\mathbf{e}_k = (0, \dots, 0, 1, 0, \dots, 0)^T$ is the k th Euclidean basis vector (with an entry of 1 at the k th element and 0 elsewhere). Fig. 5 illustrates an example of a bivariate wrapped Gaussian PDF. There, in correspondence to Fig. 4(a), the reader can see nine Gaussian summands (corresponding to $k_1 = -1, 0, 1$ and $k_2 = -1, 0, 1$) as well as their sum (the small patch, restricted to the interval $(-\pi, \pi] \times (-\pi, \pi]$). Fig. 6(a) shows another bivariate wrapped Gaussian as a color plot.

5.2. Multivariate semi-wrapped distribution

As it has been indicated previously, in some applications only a subset of variables in a feature vector originates from directional data, the remaining variables may be of linear type. For these situations, a suitable modeling should employ a distribution that is wrapped in the directional and non-wrapped in the linear dimensions. A multivariate distribution with this property shall be named *multivariate semi-*

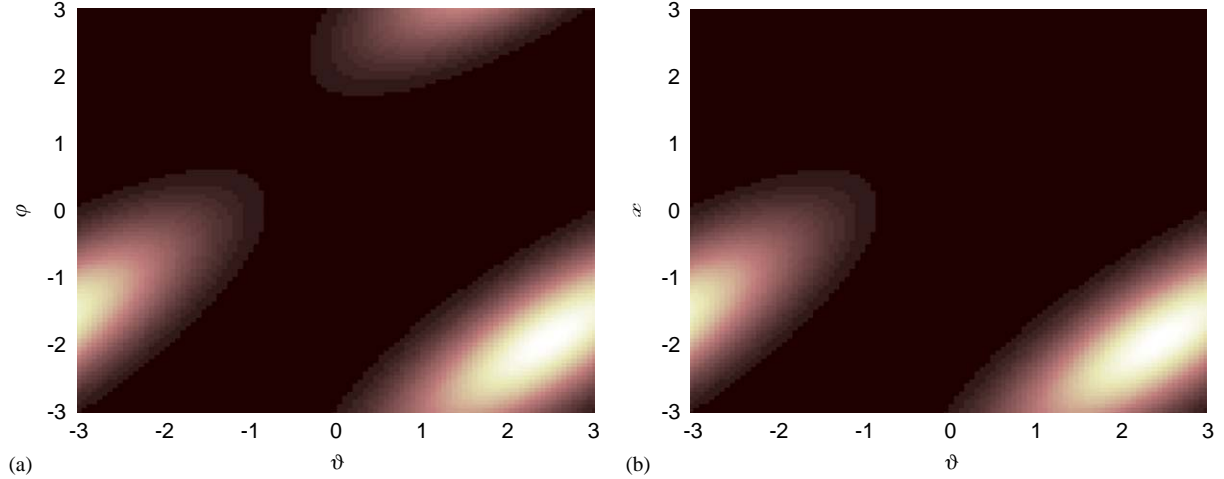


Fig. 6. Pseudo-color plots of (a) a wrapped $\mathcal{N}_{\mu_{\mathbf{x}}, \Sigma_{\mathbf{x}}}^w(\mathbf{x})$ and (b) a semi-wrapped $\mathcal{N}_{\mu_{\mathbf{x}}, \Sigma_{\mathbf{x}}}^{sw}(\mathbf{x}, 1)$ multivariate Gaussian PDF with the parameters $\mu_{\mathbf{x}} = (2.5, -2.0)^T$ and $\Sigma_{\mathbf{x}} = \begin{bmatrix} 1 & 0.7 \\ 0.7 & 0.8 \end{bmatrix}$. Bright colors denote high values. In both cases the abscissa corresponds to a circular variable. The ordinate corresponds to a circular variable in (a), however to a linear in (b). Note that for the chosen values $(\Sigma_{\mathbf{x}})_{i,i} \leq 1$, $i = 1, 2$ the assumption of a “small variance” for the circular quantity is valid and an approximation of $\mathcal{N}_{\mu_{\mathbf{x}}, \Sigma_{\mathbf{x}}}^{sw}(\mathbf{x}, 1)$ by $\mathcal{N}_{\mu_{\mathbf{x}}, \Sigma_{\mathbf{x}}}^{asw}(\mathbf{x}, 1)$ is justified.

wrapped distribution. For a simpler notation, it is assumed that the directional variable refers to only one dimension. Let f_w denote the dimension index of it. Then, the multivariate semi-wrapped distribution p_{sw} of an F -dimensional random vector \mathbf{x} can be defined as

$$p_{sw}(\mathbf{x}, f_w) = \sum_{k=-\infty}^{\infty} p(\mathbf{x} + 2\pi k \mathbf{e}_{f_w}). \quad (27)$$

For the sake of completeness, the wrapping index f_w is included as a function argument in p_{sw} .

5.3. Multivariate semi-wrapped Gaussian distribution

For p being the Gaussian PDF, that is, $p(\mathbf{x}) = \mathcal{N}_{\mu_{\mathbf{x}}, \Sigma_{\mathbf{x}}}(\mathbf{x})$, Eq. (27) becomes the *multivariate semi-wrapped Gaussian PDF*

$$\mathcal{N}_{\mu_{\mathbf{x}}, \Sigma_{\mathbf{x}}}^{sw}(\mathbf{x}, f_w) = \sum_{k=-\infty}^{\infty} (|2\pi \Sigma_{\mathbf{x}}| \exp(((\mathbf{x} - \mu_{\mathbf{x}} + 2\pi k \mathbf{e}_{f_w})^T \Sigma_{\mathbf{x}}^{-1} (\mathbf{x} - \mu_{\mathbf{x}} + 2\pi k \mathbf{e}_{f_w})))^{-1/2}. \quad (28)$$

Fig. 6(b) shows a plot of a bivariate semi-wrapped Gaussian PDF $\mathcal{N}_{\mu_{\mathbf{x}}, \Sigma_{\mathbf{x}}}^{sw}(\mathbf{x}, 1)$. The abscissa corresponds to the circular variable ϑ , the ordinate to a linear variable x . The reader can verify the wrap on the abscissa.

5.4. An approximation to the multivariate semi-wrapped Gaussian distribution

A practical handling of $\mathcal{N}_{\mu_{\mathbf{x}}, \Sigma_{\mathbf{x}}}^{sw}$ is rather involved. In particular, the computation of the infinite sum and the estimation of its parameters is a complex task. In order to cope with these problems and in agreement with the approximation derived in Section 4.3, a transition from a multivariate

semi-wrapped Gaussian distribution to an approximation of it shall be approached.

Again, the assumption for the approximation is a small variance in the circular variable in the sense that neighboring Gaussian terms of Eq. (28) have only a small overlap. For the multivariate situation this is the case, when $\sqrt{(\Sigma_{\mathbf{x}})_{f_w, f_w}} \lesssim 1$.

Under these conditions, $\mathcal{N}_{\mu_{\mathbf{x}}, \Sigma_{\mathbf{x}}}^{sw}(\mathbf{x})$ can be approximated by only one wrap—in correspondence to Eq. (19):

$$\mathcal{N}_{\mu_{\mathbf{x}}, \Sigma_{\mathbf{x}}}^{asw}(\mathbf{x}, f_w) = (|2\pi \Sigma_{\mathbf{x}}| \exp(((\mathbf{x} - \mu_{\mathbf{x}}) \bmod_{f_w} 2\pi)^T \Sigma_{\mathbf{x}}^{-1} ((\mathbf{x} - \mu_{\mathbf{x}}) \bmod_{f_w} 2\pi)))^{-1/2}. \quad (29)$$

The function \bmod_{f_w} performs the modulo operation solely on the dimension f_w .

5.5. Parameter estimates

Parameter estimation refers to the estimation of $\mu_{\mathbf{x}}$ and $\Sigma_{\mathbf{x}}$ from a set $\mathbb{X} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(M)}\}$ of observations $\mathbf{x}^{(m)}$. The elements of $\mu_{\mathbf{x}}$ can be estimated independently, using the linear ML mean estimates for the non-wrapped, and Eq. (23) for the wrapped dimension. In this respect, an estimate of $\mu_{\mathbf{x}}$ is obtained from

$$(\hat{\mu}_{\mathbf{x}})_f = \begin{cases} \frac{1}{M} \sum_{m=1}^M x_f^{(m)} & \text{if } f \neq f_w, \\ \arg \left(\frac{1}{M} \sum_{m=1}^M e^{J x_f^{(m)}} \right) & \text{else.} \end{cases} \quad (30)$$

For a covariance estimate we recall the assumption of small variances in the directional variable(s) and present an approximate solution. Since the entity of linear and directional variables is modeled by a *single* Gaussian wrap in the mentioned approximation, the semi-wrapped covariance estimate can be derived from a combination of linear ML covariance

estimate and Eq. (25), that is,

$$\hat{\Sigma}_{\mathbf{x}} \approx \frac{1}{M-1} \sum_{m=1}^M \mathbf{x}^{(m)'} \mathbf{x}^{(m)T} \quad (31)$$

with

$$x_f^{(m)'} = \begin{cases} x_f^{(m)} - (\hat{\mu}_{\mathbf{x}})_f & \text{if } f \neq f_w, \\ (x_f^{(m)} - (\hat{\mu}_{\mathbf{x}})_f) \bmod 2\pi & \text{else.} \end{cases} \quad (32)$$

In the context of the online HWR application, Eqs. (30) and (31) define the counterparts for the linear ML estimates that are employed during training. Eq. (29) is used in order to evaluate the probability density $p_{\text{asw}}(\mathbf{t}, f_w)$ of an observed feature vector \mathbf{t} during classification.

6. Experiments

This section describes experiments using the multivariate semi-wrapped Gaussian PDF, applied to online HWR in the context of the learning framework CSOTW.

6.1. Data

The experiments are based on Sections 1a, b and c (digits, upper and lower case characters, respectively) of the UNIPEN [12] Train-R01/V07 online handwriting database. For these sections, the number of classes are 10, 26, and 26, respectively, the overall data set size is approximately 16 000, 28 000 and 61 000 characters, respectively, and the sample set size of each class varies from approximately 600 to 6000 samples. Examples from the UNIPEN Section 1c are illustrated in Fig. 7. It should be noted that UNIPEN

consists of very difficult data due to the variety of writers and noisy or mislabeled data. We used the database without cleaning in order to be as comparable as possible to other classification reports.

6.2. Results

All of the results presented in the following were determined with a fivefold hold-out method, that is, the entire data set has been partitioned into disjoint training and test subsets (of a ratio 2 : 1) independently in five experiments. The divisions were completely random, thus one writer was allowed to be present in both of the sets. For a quantization of the statistical significance we include the standard deviation of the tests in the results presented below.

In order to assess the impact of the directional feature representation proposed, we have studied three different feature selections.

- (1) The first shall help judging about the discriminative power of the tangent slope angle at all. One might argue that the tangent slope angle θ is just a redundant representation of \tilde{x} and \tilde{y} and is thus useless, since it is directly computed from the other two features \tilde{x} and \tilde{y} . We want to disprove these considerations by our experiments. To this end, we include a selection comprising only the normalized horizontal and vertical coordinates $\mathbf{t}_i = (\tilde{x}_i, \tilde{y}_i)^T$ in our experiments.
- (2) A second selection follows the commonly taken approach [6–8] and uses the indirect vector modeling, $\mathbf{t}_i = (\tilde{x}_i, \tilde{y}_i, \cos \theta_i, \sin \theta_i)^T$.
- (3) The third approach sets $\mathbf{t}_i = (\tilde{x}_i, \tilde{y}_i, \theta_i)^T$ and applies the described methodology, summarized by Eqs. (29), (30) and (31).

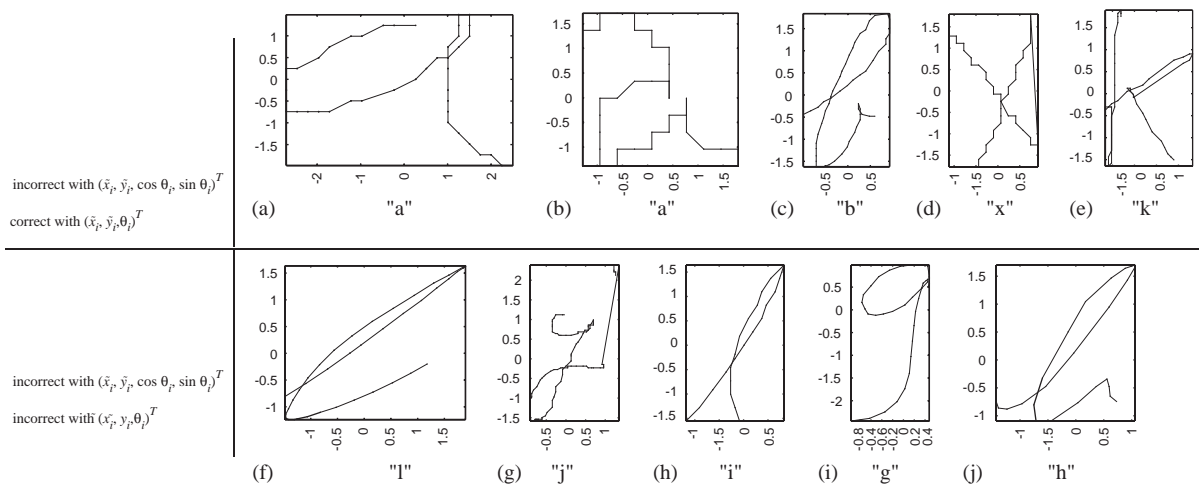


Fig. 7. Classification results of some exemplary UNIPEN handwriting samples: The first row shows samples, which are correctly classified with the proposed $(\tilde{x}_i, \tilde{y}_i, \theta_i)^T$ modeling, but incorrectly with the $(\tilde{x}_i, \tilde{y}_i, \cos \theta_i, \sin \theta_i)^T$ modeling. The last row gives examples for samples misclassified by both representations.

Table 1

Experiments on the UNIPEN Sections 1a/b/c (as indicated in the first column). The second column denotes the choice of the features. The third column shows the mean error rate \bar{E} of five different dataset partitionings, together with the respective standard deviation σ_E of the benchmark set, the fourth column the change of error rate relative to the feature set $(\tilde{x}_i, \tilde{y}_i, \theta_i)^T$. The results show that in all sections the proposed feature selection including a direct representation of the directional feature θ_i gives best recognition results

UNIPEN	Feature representation \mathbf{t}_i	Mean error rate $\bar{E} \pm \sigma_E$	Relative change of \bar{E}
1a (digits)	$(\tilde{x}_i, \tilde{y}_i)^T$	$4.2 \pm 0.2\%$	+45%
	$(\tilde{x}_i, \tilde{y}_i, \cos \theta_i, \sin \theta_i)^T$	$3.6 \pm 0.3\%$	+24%
	$(\tilde{x}_i, \tilde{y}_i, \theta_i)^T$	$2.9 \pm 0.2\%$	
1b (upper case)	$(\tilde{x}_i, \tilde{y}_i)^T$	$9.8 \pm 0.7\%$	+36%
	$(\tilde{x}_i, \tilde{y}_i, \cos \theta_i, \sin \theta_i)^T$	$7.5 \pm 0.3\%$	+4%
	$(\tilde{x}_i, \tilde{y}_i, \theta_i)^T$	$7.2 \pm 0.1\%$	
1c (lower case)	$(\tilde{x}_i, \tilde{y}_i)^T$	$13.1 \pm 0.2\%$	+40%
	$(\tilde{x}_i, \tilde{y}_i, \cos \theta_i, \sin \theta_i)^T$	$10.1 \pm 0.05\%$	+9%
	$(\tilde{x}_i, \tilde{y}_i, \theta_i)^T$	$9.3 \pm 0.05\%$	

It is worth noting, that all other recognition parameters were kept constant over the three feature extractions. The classifier complexity in terms of the overall number of allograph models were about 150, 270 and 600 for the digits, lower and upper case characters, respectively.

Table 1 summarizes mean classification error rates \bar{E} of the three scenarios, each of which is the average from five different dataset partitionings (of the ratio 2 : 1, as explained previously). It further gives the standard deviation σ_E of the respective fivefold benchmark and the change of error rate relative to the feature set $(\tilde{x}_i, \tilde{y}_i, \theta_i)^T$.

We can draw the following inferences from the results in Table 1.

1. Although the feature θ is computed from the other two features \tilde{x} and \tilde{y} , it is shown that the incorporation of θ significantly improves the recognition accuracy. The relative change of error rate varies from 36% to 45%.

2. A direct representation of θ instead of a detour over $(\cos \theta_i, \sin \theta_i)$ results in a higher accuracy (with our classifier). The proposed solution achieved lower error rates in all three UNIPEN sections. The relative change of error rate varies from 4% to 24%.

Further, the computational complexity and the memory requirements of the second feature selection are systematically higher, as evaluating Eq. (29) and the storage of $\mu_{\mathbf{x}}$ and $\Sigma_{\mathbf{x}}$ is of order $\mathcal{O}(F^2)$.

In Fig. 7, we show example classifications. Here, the first row illustrates characters, which are incorrectly classified with the $(\tilde{x}_i, \tilde{y}_i, \cos \theta_i, \sin \theta_i)^T$ feature set, and correctly with $(\tilde{x}_i, \tilde{y}_i, \theta_i)^T$. The second row shows examples misclassified by both feature selections. The ground truth label is shown below. While the characters from the first row can indeed easily and correctly be classified by a human expert, the second row contains rather ambiguous examples, which are also a challenge for the human expert (without the inclusion of knowledge about reference lines, punctuation, etc).

For a comparison of the achieved results in a broader context within the handwriting recognition community refer to previous publications [9,11].

7. Conclusions

In this contribution we have proposed a solution for a unified statistical modeling of linear and circular data with a Gaussian-like PDF. In order to reduce the analytic complexity, the approach has been confined to the constraint of small variances in the circular variables.

We have started with a brief review of directional data, its statistics, and the wrapped Gaussian distribution. Followed by this, the original contributions of the paper have been introduced: A scenario has been formulated in that the wrapped Gaussian distribution can be substituted by an approximation. Approximative solutions to the problem of parameter estimation were given. Further, extensions of the wrapped Gaussian distribution to multivariate and semi-wrapped situations were presented. As with the univariate case, we have given a complete framework for an approximative practical handling, including solutions for the tasks of parameter estimation and function evaluation.

We have shown that the proposed framework improves the recognition accuracy in our application of online handwriting recognition significantly. Compared to previous approaches of incorporating directional data into a statistical feature space modeling, errors were relatively reduced by 4–24%. Further benefits of the proposed solution are savings in computation time and memory.

The suggested solution is transferable to many existing handwriting recognition systems. It can be plugged into any system that uses a feature representation of the tangent slope angle (or other circular features) in combination with a parametric PDF modeling based on the Gaussian function class. It can be expected that in these systems the recognition accuracy can generically be increased and the time and memory complexity be decreased with the proposed modeling.

Further, the presented framework can be applied in other pattern recognition problems and areas such as biological and chemical engineering, robotics, etc.

We think that the proposed solution is a first, but not the final step to a unified handling of a semi-wrapped feature space in pattern recognition. For further research, in particular a non-approximative modeling of semi-directional situations based on the wrapped or the von Mises distribution is a worthwhile challenge.

8. Summary

The selection of valuable features is a crucial step in pattern recognition. In this paper we deal with the issue that part of features originates from *directional data*. Contrary to linear data, which are distributed on the real line, directional data are, in general, distributed on a surface of a hypersphere, in two dimensions on the circumference of a circle. Hence, directional data are inherently cyclic. Both

for linear and directional data a theory for the statistical modeling exists. Here, researchers often use Gaussians or wrapped Gaussians, respectively, in order to model probability densities. However, none of the theories gives a solution to problems, where *both* linear and directional data are to be modeled together in a *single*, multivariate probability density function. In this paper, we propose a general approach for an integrated statistical modeling of linear and circular data: the multivariate semi-wrapped Gaussian distribution. In order to reduce the analytic complexity we confine the considerations to the constraint of a small variance in the circular variable(s).

We apply the multivariate semi-wrapped Gaussian distribution to our application of the writer independent online handwriting recognition system *frog on hand*. Here, we are faced with the problem of directional data through the use of a valuable feature, the so-called *tangent slope angle*. We show experimental results and compare them with two alternative, commonly employed approaches. In the experiments we show, that our proposed modeling gives significant improvements in recognition accuracy and in computational speed compared to the alternative representations.

The suggested solution can be applied to any handwriting recognition system that uses a feature representation of the tangent slope angle (or other circular features) in combination with a parametric PDF modeling based on the Gaussian function class. It is further transferable to other applications that use directional data, such as biological and chemical engineering, robotics, etc.

Acknowledgements

The author would like to thank Nikos Canterakis from the Computer Science Department at the University of Freiburg for many valuable discussions, comments and corrections of the manuscript of this paper.

References

- [1] K.V. Mardia, *Statistics of Directional Data*, Academic Press, New York, 1972.
- [2] H. Shatkay, L.P. Kaelbling, *Heading in the right direction*, in: *Proceedings of the 15th International Conference on Machine Learning*, Morgan Kaufmann Publishers, San Francisco, USA, 1998, pp. 531–539.
- [3] R. Döker, T. Maurer, W. Kremer, K.-P. Neidig, H.R. Kalbitzer, *Determination of mean and standard deviation of dihedral angles*, *Biochem. Biophys. Res. Commun.* 257 (2) (1999) 348–350.
- [4] S.G. Hyberts, M.S. Goldberg, T.F. Havel, G. Wagner, *The solution structure of eglin c based on measurements of many NOEs and coupling constants and its comparison with X-ray structures*, *Protein Sci.* 1 (1992) 736–751.
- [5] B.C. Lovell, P.J. Kootsookos, R.C. Williamson, *The circular nature of discrete-time frequency estimates*, in: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 5, 1991, p. 3369.
- [6] I. Guyon, P. Albrecht, Y. LeCun, J.S. Denker, W. Hubbard, *Design of a neural network character recognizer for a touch terminal*, *Pattern Recog.* 24 (2) (1991) 105–119.
- [7] M. Schenkel, I. Guyon, D. Henderson, *On-line cursive script recognition using time delay neural networks and hidden Markov models*, in: R. Plamondon (Ed.), *Machine Vision and Applications* 8, Springer, Berlin, 1995, pp. 215–223.
- [8] S. Jäger, S. Manke, A. Waibel, *NPEN++: An on-line handwriting recognition system*, in: *Proceedings of the 7th International Workshop Front. Handwriting Recognition (IWFHR)*, Amsterdam, The Netherlands, 2000, pp. 249–260.
- [9] C. Bahlmann, H. Burkhardt, *The writer independent online handwriting recognition system frog on hand and cluster generative statistical dynamic time warping*, *IEEE Trans. Pattern Anal. Mach. Intell.* 26 (3) (2004) 299–310.
- [10] C. Bahlmann, *Advanced Sequence Classification Techniques Applied to Online Handwriting Recognition*, Ph.D. Thesis, Albert-Ludwigs-Universität Freiburg, Institut für Informatik, 2005. URL lmb.informatik.uni-freiburg.de/.
- [11] E.H. Ratzlaff, *Methods, report and survey for the comparison of diverse isolated character recognition results on the unipen database*, in: *Proceedings of the 7th International Conference on Document Analysis and Recognition (ICDAR)*, Edinburgh, Scotland, 2003.
- [12] I. Guyon, L.R.B. Schomaker, R. Plamondon, M. Liberman, S. Janet, *UNIPEN project of on-line data exchange and recognizer benchmarks*, in: *Proceedings of the 12th International Conference on Pattern Recognition (ICPR)*, Jerusalem, Israel, 1994, pp. 29–33. URL www.unipen.org/.
- [13] R. Plamondon, S.N. Srihari, *On-line and off-line handwriting recognition: A comprehensive survey*, *IEEE Trans. Pattern Anal. Mach. Intell.* 22 (1) (2000) 63–84.
- [14] C. Bahlmann, B. Haasdonk, H. Burkhardt, *On-line handwriting recognition with support vector machines—a kernel approach*, in: *Proceedings of the 8th International Workshop Front. Handwriting Recognition (IWFHR)*, Niagara-on-the-Lake, Canada, 2002, pp. 49–54.
- [15] C. Bahlmann, H. Burkhardt, *Measuring HMM similarity with the Bayes probability of error and its application to online handwriting recognition*, in: *Proceedings of the 6th International Conference on Document Analysis and Recognition (ICDAR)*, Seattle, WA, 2001, pp. 406–411.
- [16] L.R. Rabiner, B.H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, NJ, 1993.
- [17] R. von Mises, *Über die Ganzzahligkeit der Atomgewichte und verwandte Fragen*, *Phys. Z.* 19 (1918) 490–500.
- [18] Frits Zernike, *Wahrscheinlichkeitsrechnung und mathematische Statistik*, in: H. Thirring (Ed.), *Handbuch der Physik*, vol. 3, Springer, Berlin, 1928, pp. 419–492, (Chapter 12).
- [19] A. Wintner, *On the stable distribution laws*, *Am. J. Math.* 55 (1933) 335–339.
- [20] E.G. Gumbel, J.A. Greenwood, D. Durand, *The circular normal distribution: Theory and tables*, *J. Am. Stat. Assoc.* 48 (1953) 131–152.

About the Author—CIAUS BAHLMANN is currently postdoc researcher with Siemens Corporate Research Inc., in Princeton, NJ, since 2004. His research interests include various areas in pattern recognition, computer vision, and machine learning. From 1998–2003, he was with the Computer Science Department, University of Freiburg, Germany as a research associate, where he received his doctoral degree “with highest honor” based on new types of generative and discriminative classification for online handwriting recognition. He received the B.S. and M.S. in computer science from the University of Bielefeld, Germany in 1997. In 2002, his work “On-line Handwriting Recognition with Support Vector Machines—A Kernel Approach” was awarded the “Best Paper Presentation” prize at the IWFHR 2002 conference in Niagara-on-the-Lake, Canada. In 2005, his Ph.D. thesis “Advanced Sequence Classification Techniques Applied to Online Handwriting Recognition” was awarded the “Wolfgang-Gentner-Nachwuchsförderpreis” from the University of Freiburg, Germany.