Uncertainty-Aware PCA Revisited

Lukas Friesecke, Christian Braune, Christian Rössl, and Holger Theisel

Abstract—Principal Component Analysis (PCA) is perhaps the most popular linear projection technique for dimensionality reduction. We consider PCA under the assumption that the high-dimensional data points are equipped with Gaussian uncertainty. Several approaches to such uncertainty-aware PCA have been developed recently in the visualization community. Since PCA is a discontinuous map, a small uncertainty in the data points can result in a huge uncertainty in the projected points. We show that the uncertainty of the data points also creates uncertainty in the eigenvectors of the covariance matrix that defines the PCA projection. We present a closed-form expression to quantify eigenvector uncertainty. Based on this, we propose a 3D glyph that supports the decision whether existing solutions for uncertainty-aware PCA are sufficient, or whether a more expensive sampling-based approach is required. We apply our approach to several test data sets.

Index Terms—PCA, dimensionality reduction, uncertainty visualization

1 Introduction

Dimensionality reduction, i.e., finding "good" projections from a high-dimensional data space to, e.g., a 2D screen, is a standard problem in visualization and other areas of data analysis. A variety of techniques has been developed. Existing techniques can be distinguished into two groups: linear vs. non-linear methods. Perhaps the most popular linear dimensionality reduction technique is the *principal component analysis* (PCA) [16]. Given the covariance matrix of the mean-centered high-dimensional data, the PCA projects the data into the linear subspace spanned by the major eigenvectors of the covariance matrix. For visualization, we usually select the first two or three major eigenvectors.

Recently, Görtler et al. [13] studied uncertainty in the data. They show that if the data points are uncertain – with uncertainty defined by a local mean and covariance matrix for each data point – then the global covariance matrix for the PCA projection depends on both: the individual local covariance matrices of each data point and the covariance of their means. The projected uncertain 2D points have then an uncertainty described by a 2D covariance matrix that is an orthographic projection of the high-dimensional covariance of each data point.

The research presented in this paper is driven by the realization that PCA is a *discontinuous map*: A small perturbation of the high-dimensional input data may lead to a large change in the eigenvalues and thus to a completely different projection, which depends on the order of eigenvalues and associated eigenvectors. This behavior is due to the instability of eigenvectors of covariance matrices with similar eigenvalues. This makes the consideration of uncertainty in PCA a challenging task: A small perturbation of the input data can be modeled by adding a small uncertainty to the data. This small uncertainty can, however, result in a huge uncertainty in the projected data. This behavior is not covered by Görtler et al., as for their approach the projected uncertainty of a data point is always smaller than (or equal to) the uncertainty of the points in data space. Similarly, the use of derivatives of eigenvectors as proposed recently by Zabel et al. [41] can be problematic as eigenvectors change discontinuously as two eigenvalues become equal.

The main insight of our paper is that the uncertainty of the input data points does not only influence the global covariance matrix and induce uncertainty to the projected points. It also introduces uncertainty to the global covariance matrix itself. We argue and show in examples that it is necessary to incorporate this uncertainty of the global covariance matrix into the PCA projection. As a result we present a sampling-based technique for the visual analysis of uncertainty-aware PCA.

Since our sampling-based technique is significantly more expensive than the approach by Görtler et al. [13], the question arises when it is necessary to use our technique, and when it is safe to stick to theirs. We address this question by introducing the *covariance stability glyph*, a 3D glyph that encodes how uncertain the two major eigenvectors of an uncertain high-dimensional covariance matrix are. We apply our new techniques to a number of test data sets and show that this glyph provides a simple method to decide which of the two methods should be used.

The main contributions of this paper are:

- We present a simple example where the uncertainty-aware PCA in [13] gives incorrect results.
- For uncertain input data points, with each from a multivariate normal distribution, we present a closed-from solution of the uncertain global covariance matrix.
- We introduce a sampling-based technique for uncertainty-aware PCA that considers the uncertainty of the global covariance matrix.
- We present a closed-form solution to compute a measure that quantifies the likelihood that an arbitrary vector is an eigenvector of an uncertain covariance matrix.
- Based on this, we introduce the covariance stability glyph that encodes the stability of the major eigenvectors of an uncertain covariance matrix.

2 Related Work

There exist many techniques and different taxonomies for dimension reduction. We refer to the following surveys [5, 22, 24, 36, 40] for an overview.

Automatic projection techniques aim to find projections that are optimal in some sense, i.e., which minimize certain criteria. Linear projection methods yield projections that are linear maps, examples are PCA [16], LDA [10], and a variety of variants of them. Examples of nonlinear techniques are Classical MDS [33], LLE [28], MVU [38], LSP [26], LAMP [27], SNE [15], t-SNE [35], and UMAP [20]. Each of them comes again with several extensions and variants.

For the PCA there exist extensions and improvements (see, e.g., [6, 19]): The Kernel PCA [30] tries to capture nonlinear patterns in the data by first applying possibly nonlinear map to a higher dimensional space in clusters can be linearly separated. Bayesian PCA [3, 21, 23, 29] focuses on estimating the dimensionality of the reduced space, while robust PCA methods [2, 34, 37] address the presence of outliers. Extensions to PCA have also been proposed in the context of fuzzy systems such as [8], in which PCA was adapted for fuzzy numbers by training an artificial neural network that accounts for the range of

• *Lukas Friesecke, Christian Braune, Christian Rössl, and Holger Theisel are with Otto von Guericke University Magdeburg. E-mail: {lukas.friesecke, christian.braune, christian.roessl, theisel}@ovgu.de*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

possible realizations for each fuzzy value. In [12] a review of methods for applying PCA to interval data is given.

Probabilistic PCA [32] defines PCA by using a Gaussian density estimation framework, which enables statistical testing and integration with Bayesian methods. An expectation-maximization algorithm is proposed to iteratively estimate the principal subspace, offering computational benefits for large data sets. Contrary to the approach in [13] and our new approach, an unknown isometric measurement error is assumed and estimated. This makes the computation of the model much faster at the price of being overly simplistic.

The approach most related to ours is the uncertainty-aware PCA by Görtler et al. [13]. In fact, our approach can be considered as an extension and generalization of their approach as will be detailed in Section 4. This approach has been applied to biological visualization [39], and was combined with other approaches to uncertainty-aware projections [14, 25].

Another recent approach considers the sensitivity of eigenvectors of the covariance matrix in uncertain PCA by Zabel et al. [41]. Their method constructs a local linearization of the eigenvectors and the output data from a linearization of the uncertain input data, where the required Jacobians are computed by automatic differentiation. While this allows to better model the nonlinearity of constructing the PCA, i.e., solving an eigenvalue problem, it does not consider its inherent discontinuity. In Section 4.3 we show a simple example where such linearization is not sufficient, such that a more general treatment of discontinuities in the PCA is necessary. Furthermore, Zabel et al. [41] present an animation-based approach for visualizing uncertain PCA: a random closed curve in the space of all data points (on an equipotential hyper-surface of the corresponding PDF) is mapped to the screen and animated. While this gives interesting animations, the results depend on the random choice of the closed input curve.

Our approach is based on the definition of the representation of uncertain covariance matrices. Similar definitions – but reduced to 3×3 matrices – are given in the context of uncertain tensor visualization by Gerrits et al. [11].

As part of our approach, we propose a 3D glyph to encode uncertainty of the two major eigenvectors of covariance matrix for a high-dimensional data space. Glyph design is a well-studied problem in visualization [4]. Several glyphs have been developed for the visualization of the uncertainty of 3D tensors, including cones of uncertainty [18], HiFiVE glyphs [31], SIP glyphs [17], multiple glyphs [1], or interactive approaches [42]. All of them come from 3D tensor visualization and are therefore restricted to uncertain 3×3 tensors.

3 Notation and Basic Concepts

3.1 Mandel Notation of a Symmetric Matrix

We use the Mandel notation that maps a symmetric $n\times n$ matrix $\mathbf{C}$ into an r-vector $\mathbf{v}(\mathbf{C})$ with $r=\frac{1}{2}n(n+1)$ such that $\mathbf{v}(\mathbf{C})$ contains all entries of $\mathbf{C}$. We define this map by an auxiliary matrix $\mathbf{T}\in\mathbb{N}^{r\times 2}$ that has entries from the set $\{1,\ldots,n\}$ such that every pair $(i,j)\in\{1,\ldots,n\}^2$ with $i<j$ is contained in exactly one row of $\mathbf{T}$. Further, $\mathbf{T}$ defines the r-vector $\mathbf{d}$

$$\mathbf{d}[i]=\begin{cases}1 & \text{for } \mathbf{T}[i,1]=\mathbf{T}[i,2]\\ \sqrt{2} & \text{otherwise}\end{cases}. \tag{1}$$

Then

$$\mathbf{v}(\mathbf{C})[i]=\mathbf{d}[i]\cdot\mathbf{C}[\mathbf{T}[i,1],\mathbf{T}[i,2]] \tag{2}$$

for $i=1,\ldots,r$, where the brackets $[\cdot]$ denote the index of matrix or vector's component. For example, for $n=3$, the definition of

$$\mathbf{T}=\begin{bmatrix}1&2&3&2&1&1\\1&2&3&3&3&2\end{bmatrix}^{\mathrm{T}} \tag{3}$$

gives $\mathbf{d}=(1,1,1,\sqrt{2},\sqrt{2},\sqrt{2})^{\mathrm{T}}$ such that

$$\begin{aligned}\mathbf{v}(\mathbf{C}) &= \mathbf{v}\left(\begin{bmatrix}c_{11}&c_{12}&c_{13}\\c_{12}&c_{22}&c_{23}\\c_{13}&c_{23}&c_{33}\end{bmatrix}\right) &(4)\\ &= (c_{11},c_{22},c_{33},\sqrt{2}c_{23},\sqrt{2}c_{13},\sqrt{2}c_{12})^{\mathrm{T}}. &(5)\end{aligned}$$

We remark the definition of $\mathbf{T}$ is not unique and may use any permutation of rows. Given any concrete and fixed $\mathbf{T}$ the map $\mathbf{v}$ is bijective and the inverse $\mathbf{v}^{-1}$ maps an r-vector to a symmetric $n\times n$ matrix.

Note that $\mathbf{v}(\cdot)$ preserves norms (due to the $\sqrt{2}$ entries), i.e., $||\mathbf{C}||_F=||\mathbf{v}(\mathbf{C})||$, where $||\cdot||_F$ denotes the Frobenius norm of a matrix, and $||\cdot||=||\cdot||_2$ is the standard Euclidean norm. We use the operator $\mathbf{v}(\cdot)$ to describe the uncertainty of matrices in terms of standard matrix and vector operations instead of non-standard higher order tensor operations.

3.2 Random Variables, Distribution Functions, and Normal Distributions

A vector of random variables (or random vector) $\mathbf{x}\in\mathbb{R}^n$ describes a possible n-dimensional realization of a random process. The probability of such an outcome is described by a multivariate *probability density function* (PDF) $p(\mathbf{x})$ fulfilling $p(\mathbf{x})\geq 0$ and

$$\int_{\mathbb{R}^n} p(\mathbf{x})\,d\mathbf{x}=1 \tag{6}$$

where $p(\mathbf{x})$ describe the likelihood that any sample drawn from the random process would be the random vector $\mathbf{x}$. We also call a random vector $\mathbf{x}$ following the PDF p a *realization* of the distribution p. Note that while the infinite integral of $p(\mathbf{x})$ is always 1, $p(\mathbf{x})$ can be larger than 1 at some locations $\mathbf{x}\in\mathbb{R}^n$.

Of particular interest are n-dimensional normal distributions, for which holds

$$p(\mathbf{x})=\frac{1}{\sqrt{(2\pi)^n\det\mathbf{C}}}\,e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^{\mathrm{T}}\mathbf{C}^{-1}(\mathbf{x}-\mathbf{m})} \tag{7}$$

where $\mathbf{m}$ is a n-vector describing the mean and $\mathbf{C}$ is a positive definite symmetric $n\times n$ covariance matrix. If a realization $\mathbf{x}$ is drawn from such a distribution we write $\mathbf{x}\sim\mathcal{N}_n(\mathbf{m},\mathbf{C})$.

3.3 Regular PCA

We consider the regular principal component analysis (PCA) as a linear projection technique from a finite number of N data points $\mathbf{x}_i$ in an n-dimensional data space to the projected points $\mathbf{y}_i$ in an m-dimensional projection space with $n\geq m$. In general we would choose $m\ll n$. However, for visualization purposes we usually consider $m\leq 3$. In its regular setup, PCA can be formulated as follows: Given are N points $\mathbf{x}_1,\ldots,\mathbf{x}_N\in\mathbb{R}^n$. We search for their projected points $\mathbf{y}_1,\ldots,\mathbf{y}_N\in\mathbb{R}^m$. We compute mean $\mathbf{m}$ and covariance $\mathbf{C}$ of the points $\mathbf{x}_i$ as

$$\mathbf{m} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i \tag{8}$$

$$\mathbf{C} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i-\mathbf{m})(\mathbf{x}_i-\mathbf{m})^{\mathrm{T}}=\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}-\mathbf{m}\,\mathbf{m}^{\mathrm{T}} \tag{9}$$

where $\mathbf{m}$ is an n-dimensional point and $\mathbf{C}$ is a symmetric positive semi-definite $n\times n$ matrix.
Let $\mathbf{U}:=\mathbf{U}(\mathbf{C})\in\mathbb{R}^{n\times m}$ be the orthogonal matrix with the first m major eigenvectors of $\mathbf{C}$ sorted by descending eigenvalues as columns, then the PCA can be expressed as

$$\mathbf{y}_i=\mathbf{y}_i(\mathbf{x}_1,\ldots,\mathbf{x}_N)=\mathbf{U}^{\mathrm{T}}\,(\mathbf{x}_i-\mathbf{m}) \tag{10}$$

after which the images $\mathbf{y}_i$ of $\mathbf{x}_i$ reside in $\mathbb{R}^m$.

4 Uncertainty-Aware PCA

Görtler et al. [13] formulate the problem of uncertainty-aware PCA by considering uncertainty of the data points. Instead of the data point $\mathbf{x}_i$, an uncertain data point is represented by an n-dimensional PDF

$$p_i(\mathbf{x})\quad\text{for}\quad i=1,\ldots,N\,,\ \mathbf{x}\in\mathbb{R}^n \tag{11}$$

where we assume that the PDFs p_i are independent of each other. The searched unknown projected uncertain points are m-dimensional PDFs

$$q_i(\mathbf{y})\quad\text{for}\quad i=1,\ldots,N\,,\ \mathbf{y}\in\mathbb{R}^m. \tag{12}$$

Of particular interest is the case where the input distributions p_i are normal distributions, i.e.,

$$p_i(\mathbf{x}) : \mathbf{x} \sim \mathcal{N}_n(\mathbf{m}_i, \mathbf{C}_i) \quad \text{for} \quad i = 1, \ldots, N. \tag{13}$$

This means that the input data are N mean vectors $\mathbf{m}_i$ and N covariance matrices $\mathbf{C}_i$.
In the following we will write $p(\mathbf{x}) = \mathcal{N}_n(\mathbf{m}, \mathbf{C})(\mathbf{x})$ to indicate, that p is a PDF and its $\mathbf{x} \sim \mathcal{N}_n(\mathbf{m}, \mathbf{C})$.

4.1 Existing Uncertainty-Aware PCA

[13] present a solution for uncertainty-aware PCA for the case that $p_i(\mathbf{x})$ are normally distributed by (13). We call $\mathbf{m}_i$ and $\mathbf{C}_i$ the *local* mean and covariance, respectively, as they are given for each data point. From them [13] calculates the mean of the local means and the mean of the local covariance as

$$\overline{\mathbf{m}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{m}_i \quad , \quad \overline{\mathbf{C}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{C}_i \tag{14}$$

and the covariance of the local means as (cf. Equation (9))

$$\mathbf{C}_\mathbf{m} \quad = \quad \frac{1}{N} \sum_{i=1}^{N} (\mathbf{m}_i - \overline{\mathbf{m}})(\mathbf{m}_i - \overline{\mathbf{m}})^\mathrm{T} \tag{15}$$

Then the global mean $\mathbf{m}$ and the global covariance $\mathbf{C}$ that define the PCA projection are

$$\mathbf{m} = \overline{\mathbf{m}} \tag{16}$$

$$\mathbf{C} = \mathbf{C}_\mathbf{m} + \overline{\mathbf{C}}. \tag{17}$$

Note that (17) is the main theoretical contribution of [13]. Based on $\mathbf{m}$ and $\mathbf{C}$, each PDF $p_i(\mathbf{x})$ is projected to the PDF $q_i(\mathbf{y})$ by

$$q_i(\mathbf{y}) = \mathcal{N}_m(\mathbf{m}'_i, \mathbf{C}'_i)(\mathbf{y}) \tag{18}$$

with

$$\mathbf{m}'_i \quad = \quad \mathbf{U}^\mathrm{T} (\mathbf{m}_i - \mathbf{m}) \tag{19}$$

$$\mathbf{C}'_i \quad = \quad \mathbf{U}^\mathrm{T} \mathbf{C}_i \mathbf{U} \tag{20}$$

with $\mathbf{U}$ given in (10). Note, that the main theoretical contribution of [13] lies in the fact that we can use the covariance of the means and the mean of the covariances to derive a new covariance matrix from which the uncertainty-aware PCA can be calculated. Görtler et al. [13] get these results by applying summary statistics but mention that it can also be obtained *"by integrating over the deviation of all possible realizations of each probability distribution"*. Further, they state that these results do not only hold for normal distributions but for all distributions having a mean and covariance.

4.2 Existing Uncertainty-Aware PCA under Integration over Realizations

In order to analyze the solution of [13] we rewrite their approach in terms of integration over realizations. Let $\mathbb{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ be a realization of $\mathbb{P} = (p_1, \ldots, p_N)$ with each $\mathbf{x}_i \sim p_i$, i.e., $\mathbf{x}_i$ is a random point following the distribution p_i. They give a mean for this particular realization as

$$\widetilde{\mathbf{m}} \; = \; \widetilde{\mathbf{m}}(\mathbb{X}) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i. \tag{21}$$

The global mean is then obtained as integral over the means of all realizations of $\mathbb{P}$:

$$\mathbf{m} = \int_{\mathbb{R}^n} \cdots \int_{\mathbb{R}^n} (p_1(\mathbf{x}_1) \cdot \ldots \cdot p_N(\mathbf{x}_N) \cdot \widetilde{\mathbf{m}}) \; d\mathbf{x}_1 \; \ldots \; d\mathbf{x}_N. \tag{22}$$

For computing the global covariance, one first considers the covariance for the particular realization $(\mathbf{x}_1, \ldots, \mathbf{x}_N)$ of $(p_1, \ldots, p_N)$

$$\widetilde{\mathbf{C}} \; = \; \widetilde{\mathbf{C}}(\mathbb{X}) \quad = \quad \frac{1}{N} \sum_{i=1}^{N} (\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^\mathrm{T} \tag{23}$$

$$= \quad \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^\mathrm{T} \; - \; \mathbf{m}\mathbf{m}^\mathrm{T} \tag{24}$$

which gives the global covariance by integration over all realizations

$$\mathbf{C} = \int_{\mathbb{R}^n} \cdots \int_{\mathbb{R}^n} \left(p_1(\mathbf{x}_1) \cdot \ldots \cdot p_N(\mathbf{x}_N) \cdot \widetilde{\mathbf{C}} \right) \; d\mathbf{x}_1 \; \ldots \; d\mathbf{x}_N. \tag{25}$$

With this, each realization $\mathbb{X}$ of $\mathbb{P}$ is mapped to a realization $\mathbb{Y}$ of $\mathbb{Q}$ by

$$\mathbf{y}_i = \mathbf{y}_i(\mathbb{X}) = \mathbf{U}^\mathrm{T}(\mathbf{x}_i - \mathbf{m}) \tag{26}$$

for $i = 1, \ldots, N$, and $\mathbf{U}$ the orthogonal matrix with the major eigenvectors of $\mathbf{C}$ as columns. Then the final projected distributions are

$$q_i(\mathbf{y}) = \int \cdots \int_{D_i(\mathbf{y})} (p_1(\mathbf{x}_1) \cdot \ldots \cdot p_N(\mathbf{x}_N)) \; d\mathbf{x}_1 \; \ldots \; d\mathbf{x}_N \tag{27}$$

where

$$D_i(\mathbf{y}) = \{\mathbb{X} : \mathbf{y}_i(\mathbb{X}) = \mathbf{y}\} \tag{28}$$

is the set of all realizations $\mathbb{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ of $\mathbb{P} = (p_1, \ldots, p_N)$ that map $\mathbf{x}_i$ to $\mathbf{y}$.

For the special case that the input PDF $p_i(\mathbf{x})$ are normally distributed, the integrals in this section have a closed from solution. In fact, the closed-form solutions for global mean (22), global covariance (25), and projected distribution (27) are identical to the solutions in [13] by (16), (17), (18), respectively. For a derivation of these, see appendix A.2.

4.3 A Simple Example

In order to analyze the existing uncertainty-aware PCA under normal distribution, we consider a simple example with $N = 3, n = 2, m = 1$. Let the distributions p_1, p_2, p_3 be the normal distributions $\mathcal{N}_2(\mathbf{m}_1, \mathbf{C}_1)$, $\mathcal{N}_2(\mathbf{m}_2, \mathbf{C}_2)$, $\mathcal{N}_2(\mathbf{m}_3, \mathbf{C}_3)$ with

$$\mathbf{m}_1 = \begin{pmatrix} 0 \\ -1 \end{pmatrix}, \; \mathbf{m}_2 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \; \mathbf{m}_3 = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \tag{29}$$

$$\mathbf{C}_1 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \; \mathbf{C}_2 = \begin{pmatrix} 7/2 & 0 \\ 0 & 0 \end{pmatrix}, \; \mathbf{C}_3 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

Note that for distributions p_1 and p_3 the drawn realizations are always $\mathbf{m}_1$ and $\mathbf{m}_3$ respectively since there is no uncertainty. Only p_2 will show actual randomness. Applying existing uncertainty-aware PCA from Section 4.1 gives by (14) … (17)

$$\mathbf{m} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \; \mathbf{C} = \begin{pmatrix} 7/6 & 0 \\ 0 & 2/3 \end{pmatrix}. \tag{30}$$

This gives

$$\mathbf{U} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{31}$$

which results by applying (19), (20) in output normal distributions $\mathcal{N}_1(\mathbf{m}'_1, \mathbf{C}'_1)$, $\mathcal{N}_1(\mathbf{m}'_2, \mathbf{C}'_2)$, $\mathcal{N}_1(\mathbf{m}'_3, \mathbf{C}'_3)$ with

$$\mathbf{m}'_1 = \mathbf{m}'_2 = \mathbf{m}'_3 = \mathbf{0} \tag{32}$$

$$\mathbf{C}'_1 = \mathbf{0}, \; \mathbf{C}'_2 = (7/2), \; \mathbf{C}'_3 = \mathbf{0}.$$

This means that all points $\mathbf{m}_i$ are projected onto the same point $\mathbf{0}$, and for $\mathbf{m}_1$, $\mathbf{m}_3$ this projection has zero uncertainty. Figure 1 illustrates this. The top part of Figure 1 shows the three input 2D data points by ellipses representing the local covariance matrices. The bottom part of Figure 1 shows the PCA projection of the means and covariances to 1D following [13], with a small vertical offset added to $\mathbf{m}'_1$ and $\mathbf{m}'_3$ to make them distinguishable. All the means are projected to $\mathbf{0}$, and the projected covariances $\mathbf{C}'_1$ and $\mathbf{C}'_3$ are zero as well. The result of the existing uncertainty-aware PCA tells us that for any realization $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$ of p_1, p_2, p_3, the PCA projects both $\mathbf{x}_1$ and $\mathbf{x}_3$ to exactly the point $\mathbf{0}$ without any uncertainty. Unfortunately, this is not true for most realizations, not even for the realization with the highest expectation: $\mathbf{x}_1 = \mathbf{m}_1, \mathbf{x}_2 = \mathbf{m}_2, \mathbf{x}_3 = \mathbf{m}_3$. In this case, PCA projects $\mathbf{x}_1$ and $\mathbf{x}_3$ to $-\mathbf{1}$, and $\mathbf{1}$, respectively.

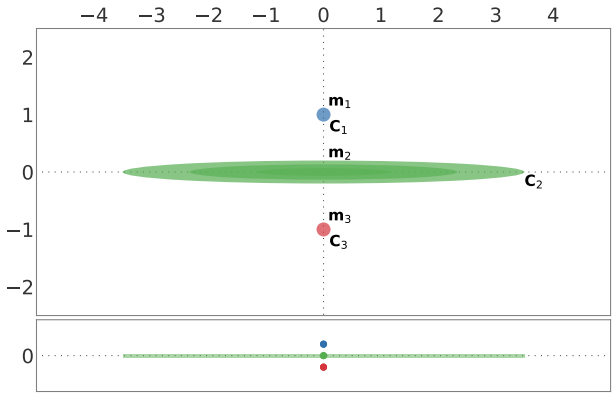


Fig. 1: Top: Three uncertain input points represented by their means $\mathbf{m}_1, \mathbf{m}_2$, and $\mathbf{m}_3$ and their respective covariance ellipses (cf. Eq. (29));Bottom: Result of using the uncertainty-aware PCA by Görtler et al. [13] to project to a single dimension. $\mathbf{m}_1$ and $\mathbf{m}_3$ are both mapped to $\mathbf{0}$ with zero uncertainty, while $\mathbf{m}_2$ still has uncertainty. To make the point distinguishable a small vertical jitter has been addded.

Let $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$ be a realization of the distributions (p_1, p_2, p_3). Due to (29), we know that

$$\mathbf{x}_1 = \begin{pmatrix} 0 \\ -1 \end{pmatrix}, \; \mathbf{x}_2 = \begin{pmatrix} a \\ 0 \end{pmatrix}, \; \mathbf{x}_3 = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \tag{33}$$

where a is a realization of the PDF

$$p_2(x) = \frac{1}{\sqrt{\frac{7}{2}}\sqrt{2\pi}} \, e^{-\frac{1}{2}\frac{2}{7}(x-0)^2} \tag{34}$$

Applying regular PCA to $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$ gives by (8), (9)

$$\mathbf{m} = \begin{pmatrix} a/3 \\ 0 \end{pmatrix}, \; \mathbf{C} = \begin{pmatrix} 2/9\, a^2 & 0 \\ 0 & 2/3 \end{pmatrix}. \tag{35}$$

In the case that a fulfills

$$-\sqrt{3} < a < \sqrt{3}, \tag{36}$$

we have $\mathbf{U} = (0,1)^{\mathrm{T}} \in \mathbb{R}^{2\times 1}$ and by applying (10) resulting in

$$\mathbf{y}_1 = (-1), \; \mathbf{y}_2 = \mathbf{0}, \; \mathbf{y}_3 = (1). \tag{37}$$

This projection is not covered by (32) at all. In fact, (32) tells us that is absolutely impossible (with zero uncertainty) that the point $\mathbf{x}_1$ in (33) is mapped to the point $\mathbf{y}_1$ in (37). The probability that (36) is true is

$$\int_{-\sqrt{3}}^{\sqrt{3}} p_2(x)\, dx \approx 0.65. \tag{38}$$

This means that in almost $2/3$ of all realizations of p_1, p_2, p_3 in (29), the result is not covered by existing uncertainty-aware PCA [13]. Figure 2 illustrates different realizations of (29) and their PCA projections. Note that none of these realizations is covered by the uncertainty-aware PCA in [13] in Figure(29).

We can use the same example (29) to analyze the approach by Zabel et al. [41]. Here, the linearization of the uncertainty of the data points leads to a zero Jacobian of the eigenvectors of the covariance matrix: a small perturbation of the data points around their mean gives a zero perturbation of the eigenvectors and therefore a projection

$$\mathbf{m}_1' = (-1), \; \mathbf{m}_2' = \mathbf{0}, \; \mathbf{m}_3' = (1) \tag{39}$$

with some linear uncertainty around $\mathbf{m}_2'$. This means that [41] captures only all data points fulfilling (36) but not the remaining ones. In other words: for this example, [13] captures only approx. $1/3$ of all possible input data, [41] covers the remaining $2/3$, but none of them yields the correct result for all possible input data.

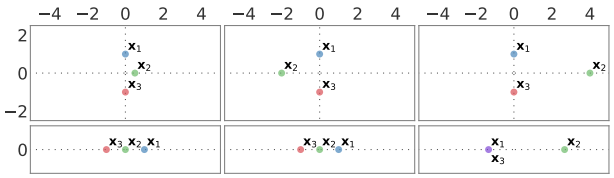


Fig. 2: Example (29): Different realizations and their PCA projections

4.4 Analysis of Existing Uncertainty-Aware PCA

Under some circumstances, the existing uncertainty-aware PCA may produce the wrong result even for a very simple example. The uncertainty-aware PCA by [13] computes a global mean and a global covariance by (16), (17) that are the basis for the projection of each realization. In other words, every realization $\mathbb{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ of $\mathbb{P} = (p_1, \ldots, p_N)$ is projected by the same transformation given by $\mathbf{m}$ and the eigenvectors of $\mathbf{C}$. This means that [13] treats PCA as a continuous map: a small perturbation of the data points (represented by a small uncertainty) always results in a small uncertainty of the projections. While this simplification has the benefit of allowing simple closed-form solutions for the unknown projected distributions, it can be an over-simplification, since every realization $\mathbb{X}$ creates its own mean and covariance from which the projection is computed.

The main insight from this analysis is: *The uncertainty of the data points also introduces uncertainty to the global mean and covariance* $\mathbf{m}$ *and* $\mathbf{C}$ *from which the PCA projection is computed.*

In the following, we propose a new approach for uncertainty-aware PCA that takes the uncertainty of $\mathbf{m}$ and $\mathbf{C}$ into consideration.

5 New Uncertainty-Aware PCA

The main idea of our approach is to consider all possible realizations of the PDFs $(p_1, \ldots, p_N)$. For each realization (i.e., for each N-tuple of data points $(\mathbf{x}_1, \ldots, \mathbf{x}_N)$ where each $\mathbf{x}_i$ follows the PDF p_i), we compute a regular PCA projection, resulting in the projected points $(\mathbf{y}_1, \ldots, \mathbf{y}_N)$ in 2D. Since we know the relative likelihood of the realization $(\mathbf{x}_1, \ldots, \mathbf{x}_N)$, we can compute the distributions of the projected points by integration over all possible input realizations $(\mathbf{x}_1, \ldots, \mathbf{x}_N)$, i.e., by integration over an $(n \cdot N)$-dimensional space. In the following, we describe this approach in detail.

We introduce our new approach first in a general integration-based formulation, before we present closed-form solutions for normal distributions. Similar to existing uncertainty-aware PCA, the data is represented as n-dimensional PDFs $p_i(\mathbf{x})$ for $i = 1, \ldots, N$, and we search for unknown projected m-dimensional PDFs $q_i(\mathbf{y})$.

Let $\mathbb{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ be a realization of $\mathbb{P} = (p_1, \ldots, p_N)$. Similar to existing uncertainty-aware PCA, the mean $\widetilde{\mathbf{m}}$ of the realization is given by (21). The new local covariance of this realization is computed as (cf. Eq. (9))

$$\widetilde{\mathbf{C}} = \widetilde{\mathbf{C}}(\mathbb{X}) = \frac{1}{N}\sum_{i=1}^{N} (\mathbf{x}_i - \widetilde{\mathbf{m}})(\mathbf{x}_i - \widetilde{\mathbf{m}})^{\mathrm{T}} \tag{40}$$

$$= \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} - \widetilde{\mathbf{m}}\widetilde{\mathbf{m}}^{\mathrm{T}}. \tag{41}$$

Note the difference of (40), (41) to the covariance (23), (24) considered in existing uncertainty-aware PCA. In order to further analyze the uncertainty of $\widetilde{\mathbf{C}}$, we write it in vector form in Mandel notation

$$\widetilde{\mathbf{c}} = \mathbf{v}(\widetilde{\mathbf{C}}) \tag{42}$$

where $\widetilde{\mathbf{c}}$ is an r-vector with $r = \frac{1}{2}n(n+1)$ containing the entries of $\widetilde{\mathbf{C}}$.

To compute the global uncertain mean and covariance, $\mathbf{m}$ is replaced by a n-dimensional PDF $m(\mathbf{x})$, and $\mathbf{v}(\mathbf{C})$ is replaced by an r-dimensional PDF $c(\mathbf{z})$ that are computed in the following way:

$$m(\mathbf{x}) = \int \cdots \int_{D_{\mathbf{m}}(\mathbf{x})} (p_1(\mathbf{x}_1) \cdot \ldots \cdot p_N(\mathbf{x}_N))\, d\mathbf{x}_1 \ldots d\mathbf{x}_N \tag{43}$$

where

$$D_{\mathbf{m}}(\mathbf{x}) = \{\mathbb{X} : \widetilde{\mathbf{m}}(\mathbb{X}) = \mathbf{x}\} \tag{44}$$

is the set of all realizations $\mathbb{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ of $\mathbb{P} = (p_1, \ldots, p_N)$ that have a mean of $\mathbf{x}$. Further,

$$c(\mathbf{z}) = \int \cdots \int_{D_{\mathbf{c}}(\mathbf{z})} (p_1(\mathbf{x}_1) \cdot \ldots \cdot p_N(\mathbf{x}_N))\, d\mathbf{x}_1 \ \ldots \ d\mathbf{x}_N \tag{45}$$

where

$$D_{\mathbf{c}}(\mathbf{z}) = \{\mathbb{X} : \widetilde{\mathbf{c}}(\mathbb{X}) = \mathbf{z}\} \tag{46}$$

is the set of all realizations $\mathbb{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ of $\mathbb{P} = (p_1, \ldots, p_N)$ that have a covariance (in Mandel notation) of $\mathbf{z}$. We call $m(\mathbf{x})$ the *global uncertain mean* and $c(\mathbf{z})$ the *global uncertain covariance*. Finally, for computing the projected distributions $q_i(\mathbf{y})$, each realization $\mathbb{X}$ is projected to $\mathbb{Y}$ by

$$\mathbf{y}_i = \mathbf{y}_i(\mathbb{X}) = \widetilde{\mathbf{U}}^{\mathrm{T}}(\mathbf{x}_i - \widetilde{\mathbf{m}}) \tag{47}$$

for $i = 1, \ldots, N$, where $\widetilde{\mathbf{U}}$ is the orthogonal matrix consisting of the major eigenvectors of $\widetilde{\mathbf{C}}$ sorted by descending eigenvalues, and the calculations of $\widetilde{\mathbf{m}}$ and $\widetilde{\mathbf{C}}$ given in (21) and (40) respectively. With this, the distributions $q_i(\mathbf{y})$ are computed - similar to existing uncertainty-aware PCA - by (27) and (28).

To summarize, the difference of the new uncertainty-aware PCA to the existing uncertainty-aware PCA by Görtler et al. [13] is to

- replace the global mean $\mathbf{m}$ in (22) by the global uncertain mean $m(\mathbf{x})$ in (43),
- replace the global covariance $\mathbf{C}$ in (25) by the global uncertain covariance $c(\mathbf{z})$ in (45),
- replace (26) by (47) for the computation of the projected distributions $q(\mathbf{y})$, ensuring that each realization is projected by its own mean and covariance.

5.1 New Uncertainty-Aware PCA under Normal Distribution

For the special case that the input PDF p_i are normal distributions by (13), both the uncertain global mean $m(\mathbf{x})$ and the uncertain global covariance $c(\mathbf{z})$ are normal distributions by

$$m(\mathbf{x}) = \mathcal{N}_n(\mathbf{m}, \mathbf{M})(\mathbf{x}) \tag{48}$$

$$c(\mathbf{z}) = \mathcal{N}_r(\mathbf{m}_{\mathbf{C}}, \mathbf{C}_{\mathbf{C}})(\mathbf{z}) \tag{49}$$

where $\mathbf{m}$ is a n-vector, $\mathbf{M}$ is an $n \times n$ covariance matrix, $\mathbf{m}_{\mathbf{C}}$ is an r-vector and $\mathbf{C}_{\mathbf{C}}$ is a $n \times n$ covariance matrix that are computed as

$$\mathbf{m} = \int_{\mathbb{R}^n} \cdots \int_{\mathbb{R}^n} (p_1(\mathbf{x}_1) \cdot \ldots \cdot p_N(\mathbf{x}_N) \cdot \widetilde{\mathbf{m}})\, d\mathbf{x}_1 \ \ldots \ d\mathbf{x}_N \tag{50}$$

$$\mathbf{m}_{\mathbf{C}} = \int_{\mathbb{R}^n} \cdots \int_{\mathbb{R}^n} (p_1(\mathbf{x}_1) \cdot \ldots \cdot p_N(\mathbf{x}_N) \cdot \widetilde{\mathbf{c}})\, d\mathbf{x}_1 \ \ldots \ d\mathbf{x}_N \tag{51}$$

$$\widetilde{\mathbf{M}} = (\widetilde{\mathbf{m}} - \mathbf{m})(\widetilde{\mathbf{m}} - \mathbf{m})^{\mathrm{T}} \tag{52}$$

$$\widetilde{\mathbf{C}_{\mathbf{C}}} = (\widetilde{\mathbf{c}} - \mathbf{m}_{\mathbf{C}})(\widetilde{\mathbf{c}} - \mathbf{m}_{\mathbf{C}})^{\mathrm{T}} \tag{53}$$

$$\mathbf{M} = \int_{\mathbb{R}^n} \cdots \int_{\mathbb{R}^n} \left(p_1(\mathbf{x}_1) \cdot \ldots \cdot p_N(\mathbf{x}_N) \cdot \widetilde{\mathbf{M}}\right) d\mathbf{x}_1 \ \ldots \ d\mathbf{x}_N \tag{54}$$

$$\mathbf{C}_{\mathbf{C}} = \int_{\mathbb{R}^n} \cdots \int_{\mathbb{R}^n} \left(p_1(\mathbf{x}_1) \cdot \ldots \cdot p_N(\mathbf{x}_N) \cdot \widetilde{\mathbf{C}_{\mathbf{C}}}\right) d\mathbf{x}_1 \ \ldots \ d\mathbf{x}_N \tag{55}$$

Note that (50), (51), (54), (55) have closed-form solutions

$$\mathbf{m} = \overline{\mathbf{m}} \tag{56}$$

$$\mathbf{M} = \frac{1}{N}\overline{\mathbf{C}} \tag{57}$$

$$\mathbf{C} = \mathbf{C}_{\mathbf{m}} + \frac{N-1}{N}\overline{\mathbf{C}} \tag{58}$$

$$\mathbf{m}_{\mathbf{C}} = \mathbf{v}(\mathbf{C}). \tag{59}$$

Further, $\mathbf{C}_{\mathbf{C}}$ is an $r \times r$ matrix for which the closed form solution is given in Appendix A.1 and its derivation in Appendix A.2.

Unfortunately, the projected distributions $q_i(\mathbf{y})$ are not normal distributions and in general seem to have no closed-form solution. We compute them by a Monte-Carlo sampling described in the next section.

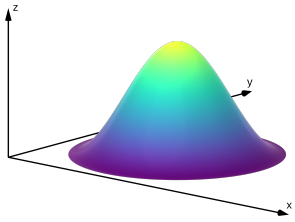


Fig. 3: 2D radial Hann function $h_{\mathbf{y}_0,R}(\mathbf{y})$

6 Sampling-Based Computation of Projected Distributions

We compute a (large) number U of realizations $\mathbb{X}_k = (\mathbf{x}_{1,k}, \ldots \mathbf{x}_{N,k})$ of the input distributions $\mathbb{P}_k = (p_1, \ldots, p_N)$ for $k = 1, \ldots, U$. For each realization, we perform a regular PCA with

$$\widetilde{\mathbf{m}}_k = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_{i,k} \quad , \quad \widetilde{\mathbf{C}}_k = \frac{1}{N}\sum_{i=1}^{N} (\mathbf{x}_{i,k} - \widetilde{\mathbf{m}}_k)(\mathbf{x}_{i,k} - \widetilde{\mathbf{m}}_k)^{\mathrm{T}}$$

and compute the m major eigenvectors of $\widetilde{\mathbf{C}}_k$ as columns of $\mathbf{u}_{k,j}$ of

$$\widetilde{\mathbf{U}}_k = \mathbf{U}_k(\widetilde{\mathbf{C}}_k) \tag{60}$$

for $j = 1, \ldots, m$. Since the orientation of eigenvectors is arbitrary, we make them orientation consistent by the following algorithm:

Input : An orthogonal matrix $\widetilde{\mathbf{U}}_k = \mathbf{U}_k(\widetilde{\mathbf{C}}_k)$ (cf. (60)).
Output : An orthogonal matrix $\widetilde{\mathbf{U}}_k$ with the eigenvectors all oriented consistently.

```
ū ← ũ_{k,1}
for j ← 2 to U do
    if ū^T ũ_{k,j} < 0 then
        ũ_{k,j} ← −ũ_{k,j}
    end
    ū ← ((j−1)/j) ū + (1/j) ũ_{k,j}
end
```

that is computed for $k = 1, \ldots, m$. Then the projected points to 2D are

$$\mathbf{y}_{i,k} = \widetilde{\mathbf{U}}_k^{\mathrm{T}}(\mathbf{y}_{i,k} - \widetilde{\mathbf{m}}_k) \quad \text{for } i = 1, \ldots, N\,. \tag{61}$$

From this, we compute the projected distributions as

$$q_i(\mathbf{y}) = \frac{1}{U}\sum_{j=1}^{U} h_{\mathbf{y}_{i,k},R}(\mathbf{y}) \tag{62}$$

where h is the radial Hann function [9] with center $\mathbf{y}_0$ and radius of the support $R > 0$ that is in 2D defined as

$$h_{\mathbf{y}_0,R}(\mathbf{y}) = \begin{cases} \frac{2\pi}{R^2(\pi^2-4)} \cos^2\left(\frac{\pi\|\mathbf{y}-\mathbf{y}_0\|}{2R}\right) & \text{for} \quad \|\mathbf{y}-\mathbf{y}_0\| < R \\ 0 & \text{o.w.} \end{cases}\,. \tag{63}$$

Figure 3 illustrates this. Any other radial basis function with local support, C^1 continuity, and integral 1 over its whole domain could be chosen as well. Note that the $q_i(\mathbf{y})$ in (62) are C^1-continuous PDFs. For a fast computation of q_i, an efficient lookup is necessary to decide which projected points $\mathbf{y}_{i,k}$ are in the R-neighborhood of a point $\mathbf{y}$, as only these points influence $q_i(\mathbf{y})$. This can be done by a binning of the points $\mathbf{y}_{i,k}$.

Our computation of q_i depends on two parameters: the number of samples U and the radius R of the support of the radial basis function. We discuss their influence on performance and accuracy in Section 9.

7 Uncertain Eigenvalues and Eigenvectors

Our sampling-based approach in Section 6 is significantly more expensive than the approach by Görtler et al. [13]. On the other hand, we identified a simple example where their method gives wrong results, and thus a sampling-based approach should be used. This raises the following question: We present an approach to decide a priori, i.e., *without* carrying out the sampling, whether the approach by Görtler et al. is sufficient, or whether the more expensive sampling is necessary. This decision requires an analysis of the uncertainty of the global covariance matrix, which is developed in this section. In particular, we quantify the *uncertainty of the eigenvectors and eigenvalues of the global covariance matrix*, which is crucial as they steer the PCA projection. If this uncertainty is small, i.e., the eigenvectors are rather certain, the result produced by Görtler et al. is reliable. The more uncertain the eigenvectors, the more unreliable are their results. In general, we expect the uncertainty of the global covariance to depend on two factors: the uncertainty of the data points, i.e., the local covariance matrices, and the dissimilarity of the eigenvalues of the global covariance matrix: if the eigenvalues are close to each other, a small perturbation of the data, i.e., a small local uncertainty, can change the (selection of) eigenvectors and therefore the PCA projection drastically. Based on the definition of uncertain eigenvectors, we develop a glyph-based visualization, which supports a user to answer the initial question on requiring or not requiring a sampling approach.

Given the global uncertain covariance matrix $c(\mathbf{z})$, we measure the uncertainty of its eigenvectors and eigenvalues by two distributions

$$v(\mathbf{x}) = \int_{E(\mathbf{x})} c(\mathbf{z})\, d\mathbf{z} \quad \text{and} \quad e(\mathbf{x},\lambda) = \int_{E(\mathbf{x},\lambda)} c(\mathbf{z})\, d\mathbf{z}\,, \tag{64}$$

where $\mathbf{x} \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$, and $E(\mathbf{x})$ (and $E(\mathbf{x},\lambda)$) denotes the set of all symmetric matrices that have $\mathbf{x}$ as eigenvector (with λ as associated eigenvalue). The $v(\mathbf{x})$ (and $e(\mathbf{x},\lambda)$) measure the likelihood that $\mathbf{x}$ is an eigenvector of the global covariance matrix (with eigenvalue λ). Note that they are *not* probability functions, as will be discussed below. We call $v(\mathbf{x})$ and $e(\mathbf{x},\lambda)$ *uncertain eigenvectors* and *uncertain eigenvalues*.

In the following, we describe the sets $E(\mathbf{x})$ and $E(\mathbf{x},\lambda)$ and derive a closed-form representation of the functions $v(\mathbf{x})$ and $e(\mathbf{x},\lambda)$ for the case that $c(\mathbf{z})$ is a normal distribution.

7.1 Symmetric Matrices with Eigenvector x

Let S denote the r-dimensional vector space of symmetric $n \times n$ matrices, i.e., the set of $n \times n$ covariance matrices are a subset of S, where $r = \frac{1}{2}n(n+1)$. Let

$$E(\mathbf{x}) = \{\, \mathbf{A} \in S : \exists \lambda \in \mathbb{R} : \mathbf{A}\mathbf{x} = \lambda \mathbf{x} \,\}$$

denote the set of all $\mathbf{A} \in S$ that have $\mathbf{x}$ as eigenvector.

Lemma 1. *$E(\mathbf{x})$ is an $(r-n+1)$-dimensional linear subspace of S.*

Proof. We write symmetric matrices in $\mathbf{A} \in S$ as

$$\mathbf{A} = \mathbf{U}\Lambda\mathbf{U}^{\mathrm{T}} = \sum_{i=1}^{n} \mathbf{u}_i \mathbf{u}_i^{\mathrm{T}} \lambda_i \;=:\; \mathbf{x}\mathbf{x}^{\mathrm{T}} w_1 + \mathbf{Q} \sum_{i=2}^{n} \mathbf{u}_i \mathbf{u}_i{}^{\mathrm{T}} w_i\,. \tag{65}$$

This is the spectral decomposition with the orthogonal matrix $\mathbf{U}$ that has eigenvectors $\mathbf{u}_i$ as columns and the diagonal matrix Λ of eigenvalues $\lambda_i \in \mathbb{R}$. We change the orthonormal basis as follows: fix $\mathbf{u}_1 := \mathbf{x}$ and chose $\mathbf{Q}$ as an orthogonal transformation in the $(n-1)$-dimensional hyper-plane that has $\mathbf{x}$ as normal vector, i.e. $\mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{I}$ and $\mathbf{Q}\mathbf{x} = \mathbf{x}$ such that $\mathbf{x}, \mathbf{Q}\mathbf{u}_2, \ldots, \mathbf{Q}\mathbf{z}_n$ are orthonormal vectors. This leaves n degrees of freedom for the weights w_i and $\frac{1}{2}(n-2)(n-1)$ for $\mathbf{Q}$, which results in $r-n+1$ degrees of freedom in total. Finally, the condition $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$, i.e., $\mathbf{x}$ is an eigenvector of $\mathbf{A}$, is linear in the entries of $\mathbf{A}$. □

We can easily compute an orthogonal basis for $E(\mathbf{x})$ as well as an implicit representation of the subspace: Choose a rank-r matrix $\mathbf{R} \in \mathbb{R}^{n\times n}$ as follows: such that its first column equals $v(\mathbf{x}\mathbf{x}^{\mathrm{T}})$ and all remaining columns are of the form $v(\mathbf{z}_j \mathbf{z}_j{}^{\mathrm{T}})$ with $\mathbf{x}^{\mathrm{T}}\mathbf{z}_j = 0$. Full rank is achieved almost certainly by a random choice of vectors $\mathbf{z}_j$. Orthogonality can be achieved by applying one Gram-Schmidt step $\mathbf{z}_j \leftarrow (\mathbf{I} - \mathbf{x}\mathbf{x}^{\mathrm{T}}/||\mathbf{x}||^2)\mathbf{z}_j$ to each vector $\mathbf{z}_j$. Finally compute the singular value decomposition $\mathbf{R} = \mathbf{U}\Sigma\mathbf{V}^{\mathrm{T}}$: The first $r-n+1$ columns of $\mathbf{U}$ provide an orthonormal basis for $E(\mathbf{x})$, where each basis vector represents a symmetric matrix in Mandel notation. Let $\mathbf{B} \in \mathbb{R}^{r\times(r-n+1)}$ denote the orthonormal matrix that has these basis vectors as columns. The remaining $n-1$ columns of $\mathbf{U}$ provide the orthogonal complement. We collect them in the orthogonal matrix $\mathbf{K} \in \mathbb{R}^{r\times(n-1)}$ and get an implicit representation such that

$$E(\mathbf{x}) = \{\mathbf{B}\mathbf{w} : \mathbf{w} \in \mathbb{R}^{r-n+1}\} = \{\mathbf{z} \in \mathbb{R}^r : \mathbf{K}^{\mathrm{T}}\mathbf{z} = \mathbf{0}\}\,. \tag{66}$$

It becomes obvious that with fixing $w_1 = \lambda$ in (65), $E(\mathbf{x},\lambda)$ can be represented as an $(r-n)$-dimensional *affine* subspace of S.

7.2 Uncertain Eigenvectors

In the following, we use the basis representation with $\mathbf{B}$ of $E(\mathbf{x})$ and its implicit representation with the kernel $\mathbf{K}$ as given in (66) for integration in this subspace. As this integration generalizes to any subspace V defined by $\mathbf{B}$ or $\mathbf{K}$, respectively, and any multivariate normal distribution, we use the generic symbols in this section. For the specific application of computing the uncertainty of eigenvectors as in (64), we set $V = E(\mathbf{x})$ and $\mathbf{C} = \mathbf{C}_C$, and $\mathbf{m} = \mathbf{m}_C$ as defined in Section 5.1.

Let $\mathbf{C} \in S \subset \mathbb{R}^{n\times n}$ and $r = \frac{1}{2}n(n+1)$, and let $V \subset \mathbb{R}^r$ denote a linear subspace of $\mathbb{R}^r$. For a multivariate normal distribution $c(\mathbf{z}) = \mathcal{N}_n(\mathbf{m},\mathbf{C})(\mathbf{z})$ and an n-dimensional linear subspace $V \subseteq \mathbb{R}^r$ defined as above, the integral of c over V has the closed form

$$\int_V c(\mathbf{z})\, d\mathbf{z} = \frac{1}{\sqrt{(2\pi)^{\mathrm{rank}(\mathbf{K})}\det(\mathbf{K}^{\mathrm{T}}\mathbf{C}\mathbf{K})}}\, e^{-\frac{1}{2}s(\hat{\mathbf{z}})}\,, \tag{67}$$

with the squared Mahalanobis distance

$$s(\mathbf{z}) = (\mathbf{z}-\mathbf{m})^T\, \mathbf{C}^{-1}\, (\mathbf{z}-\mathbf{m})\,,$$

and

$$\hat{\mathbf{z}} = \arg\min_{\mathbf{z}} s(\mathbf{z}) \;\text{ subject to }\; \mathbf{z} \in V\,.$$

For a basis representation of V, we obtain for $\mathbf{x} \in \mathbb{R}^n$

$$\hat{\mathbf{z}} = \mathbf{B}\left(\mathbf{B}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{B}\right)^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{x}\,.$$

by substituting $\mathbf{x} = \mathbf{B}\mathbf{z}$ and solving the linear system $\frac{d}{d\mathbf{x}}\, s(\mathbf{z}) = \mathbf{0}$.

For an implicit representation of V, we consider the gradient $\frac{d}{d\mathbf{z}}L$ of the Lagrange function $L(\mathbf{z},\Lambda) = s(\mathbf{z}) + \Lambda^{\mathrm{T}}(K^{\mathrm{T}}\mathbf{z})$ with Lagrange multipliers Λ, which leads to solving the linear system

$$\begin{pmatrix}\hat{\mathbf{z}}\\ \Lambda\end{pmatrix} = \begin{pmatrix}\mathbf{C}^{-1} & \mathbf{K}\\ \mathbf{K}^{\mathrm{T}} & \mathbf{0}\end{pmatrix}^{-1}\begin{pmatrix}\mathbf{C}^{-1}\mathbf{m}\\ \mathbf{0}\end{pmatrix}\,.$$

Figure 4 illustrates the setting.

7.3 Properties of Uncertain Eigenvectors

The uncertain eigenvector $v(\mathbf{x})$ and eigenvalues $e(\mathbf{x},\lambda)$ are *not* normally distributed, even if $C(\mathbf{z})$ is normally distributed. Indeed, the measures $v(\mathbf{x})$ and $e(\mathbf{x},\lambda)$ are not even probability functions as they do generally not integrate to 1 over the whole domain because multiple distinct vectors $\mathbf{x}$ can be eigenvectors of $\mathbf{C}$ at the same time.

We consider integrals over spaces of symmetric matrices in $E(\mathbf{x}) \subset S \subset \mathbb{R}^{n\times n}$. While these spaces include all covariance matrices, they also include negative definite matrices, which are impossible covariance matrices. The rationale for this is that the integration in a subspace yields the closed form (67), whereas the integration over half-spaces (see Section 7.1) is considerably more difficult. Furthermore, since $\mathbf{v}^{-1}(\mathbf{m}_\mathbf{C})$ is positive definite by definition and $c(\mathbf{z})$ has an exponential decay away from $\mathbf{m}_\mathbf{C}$, we can assume that $c(\mathbf{z})$ is close to zero for $\mathbf{z}$ in

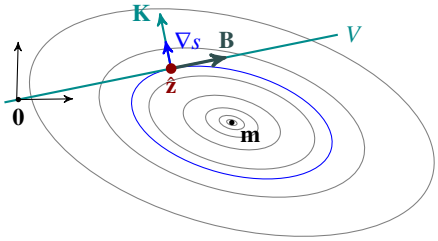


Fig. 4: The figure shows contour lines of the squared Mahalanobis distance $s(\mathbf{z})$. The evaluation point $\hat{\mathbf{z}}$ in (67) is the minimum of $s(\mathbf{z})$ restricted to the subspace V that is defined either explicitly by basis vectors $\mathbf{B}$ or implicitly by the orthogonal complement $\mathbf{K}$, which is here depicted as normal vector.

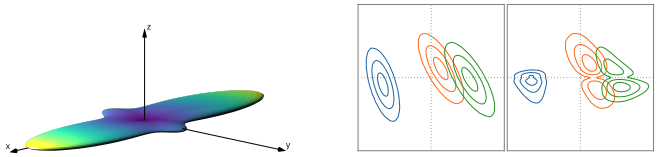


Fig. 5: The covariance stability glyph for IRIS data set shows strong deviations from distinct peaks in xy-plane (left), the projection by Görtler et al. [13] (center) and our sampling-based method (right).

regions of negative-definite matrices. This is a common assumption in many applications of normal distributions. For example, height or birth weight of people are commonly assumed as approximately normally distributed, neglecting that this gives a non-zero likelihood of negative heights or birth weights. A similar assumption was used in [11] for uncertain diffusion tensor visualization. We analyzed this assumption empirically and compared the evaluation of the closed form on the right-hand-side of (67) with a numerical approximation of the integral on the left-hand-side. We restricted the integration domain to the open half-space of the positive semi-definite matrices and applied a Monte Carlo integration. (Note that *sampling* the half-space is straightforward for a suitable basis representation of S, see Section 7.1.) We observed a significant difference only if $\mathbf{m}_\mathbf{C}$ was close to the origin. And even then, the evaluations were qualitatively very close in a sense that they showed the same number of extrema at approximately the same locations.

We finally remark that while we are able to measure whether $\mathbf{x}$ is an eigenvector of $c(\mathbf{z})$ by $v(\mathbf{x})$, we are not aware of any closed form to compute a measure of $\mathbf{x}$ being the eigenvector associated with the *largest* eigenvalue of $c(\mathbf{z})$.

7.4 Covariance Stability Glyphs

We propose *covariance stability glyphs* as a visual representation of uncertain eigenvalues and eigenvectors under normal distribution. The functions $v(\mathbf{x})$ and $e(\mathbf{x},\lambda)$ are defined w.r.t. $\mathbf{x}\in\mathbb{R}^n$ in n-dimensional data space. Their visualization is straightforward if n is sufficiently small: For $n=2$, we can represent eigenvectors in polar coordinates as $\mathbf{x}(\alpha)=(\cos\alpha,\sin\alpha)^T$. This leaves us with the visualization of radial 1D functions $v(\mathbf{x}(\alpha))$. The examples in Figure 5 and 8 show a polar plot, i.e., a 2D curve

$$v(\mathbf{x}(\alpha))\,\mathbf{x}(\alpha)\;. \tag{68}$$

Similarly, $e(\mathbf{x}(\alpha),\lambda)$ is a bivariate function that can be visualized, e.g., as a heat map. For $n=3$, writing $\mathbf{x}$ in spherical coordinates

$$\mathbf{x}(\alpha,\beta)\;=\;\begin{pmatrix}\sin\beta\\ \sin\alpha\cos\beta\\ \cos\alpha\cos\beta\end{pmatrix}$$

results in a spherical function $v(\mathbf{x})$.

Covariance Stability Glyph. In general, however, $v(\mathbf{x})$ and $e(\mathbf{x}\lambda)$ map from higher-dimensional data spaces, which renders their visualization challenging. Fortunately, we are not interested in all eigenvectors of $c(\mathbf{z})$ but only the major eigenvectors associated with the largest eigenvalues. For $m=2$, i.e., PCA projection to the 2D screen, we restrict $v(\mathbf{x})$ to the space spanned by the 3 major eigenvectors of $\mathbf{v}^{-1}(\mathbf{m}_\mathbf{C})=\mathbf{C}$. Let $\mathbf{U}=\mathbf{U}(\mathbf{C})\in\mathbb{R}^{n\times 3}$ be the orthogonal matrix with the major eigenvectors sorted by descending eigenvalues as columns, and let $\mathbf{x}(\alpha,\beta)$ be parametrized in spherical coordinates. Then we consider the spherical function

$$v(\alpha,\beta):=v(\mathbf{U}\mathbf{x}(\alpha,\beta))\;.$$

Then the subject of interest for visualization is a nonnegative spherical function $v(\alpha,\beta)$. This gives several options for visual encoding such as color, shape, texture, or combinations of these [4]. We follow the common approach to encode properties of uncertain tensors by shape, and propose a 3D *covariance stability glyph* as a closed 3D parametric surface in spherical coordinates as

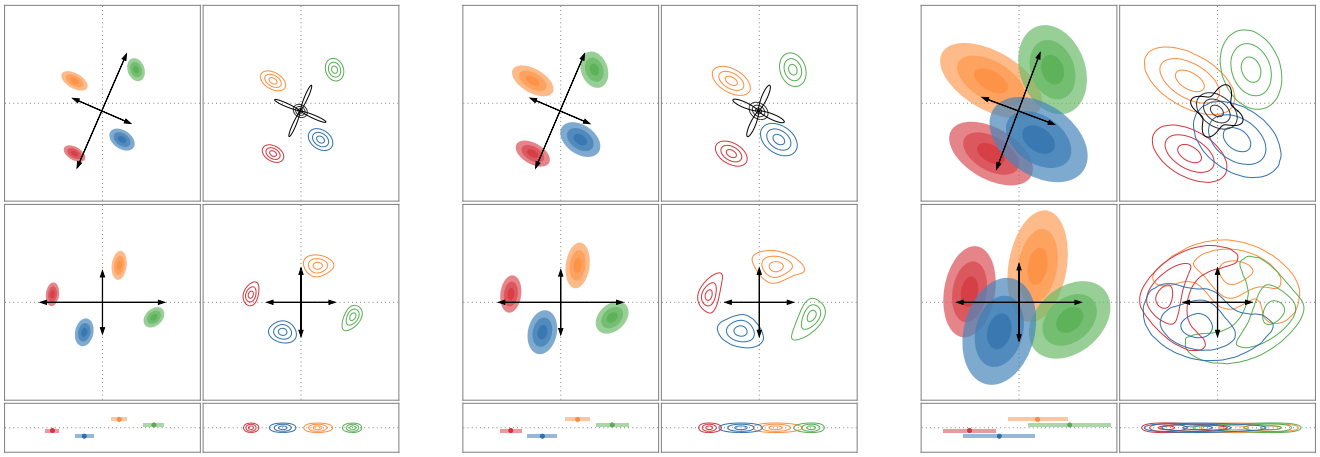
$$\mathbf{g}(\alpha,\beta)=v(\alpha,\beta)\,\mathbf{x}(\alpha,\beta)\;. \tag{69}$$

How to Read and Interpret the Glyph. The glyph $\mathbf{g}$ has the following interpretation: a strong local peak in the direction of the z-axis ($(0,0,1)^T$, i.e., $\alpha=0,\beta=0$)) shows that the first major eigenvector, i.e., the one associated with the largest eigenvalue of $c(\mathbf{z})$, has a low uncertainty. A low uncertainty of the second major eigenvector of $c(\mathbf{z})$ is encoded by a peak of $\mathbf{g}$ in y-direction ($(0,1,0)^T$, i.e., $\alpha=\pi/2,\beta=0$). Finally, a low uncertainty of the third major eigenvector is encoded by a peak in x-direction ($(1,0,0)^T$, i.e., $\alpha=\pi/2,\beta=\pi/2$).

This means that strong distinct peaks in the major axis directions but small values of $\mathbf{g}(\alpha,\beta)$ in other directions indicate that the major eigenvectors of $c(\mathbf{z})$ have a low uncertainty. In the limit of zero uncertainty of the eigenvectors, i.e., the regular PCA, $\mathbf{g}(\alpha,\beta)$ shows 6 Dirac peaks along all coordinate axes in positive and negative orientation, and $\mathbf{g}$ is 0 in all other directions. The more $\mathbf{g}(\alpha,\beta)$ is nonzero away from the major axis directions, the more uncertain are the eigenvectors of $c(\mathbf{z})$, and the less reliable is the approach by Görtler et al. [13].

Even though we consider a projection into 2D, we have to consider the uncertainty of the *three* major eigenvectors and therefore consider a 3D glyph. There could be uncertainty on deciding which is the first and the second major eigenvector but as well there could be uncertainty between the second and the third ones (or, though unlikely, even among all three). Therefore, it is necessary to analyze not only the first two major eigenvectors but the first three, even though we ultimately apply a projection to the 2D screen. This is because uncertainty of the second major eigenvector that is caused similarly to the associated second and third largest eigenvalues of $c(\mathbf{z})$ affect the projection to the 2D screen: in this case, the order of these two eigenvectors changes under a small perturbation of the data, which results in a discontinuous projection. Note that this reflects the fact that we can well find derivatives of eigenvectors (as pursued in [41]), however, the norm of the derivative approaches infinity as one eigenvalue approaches another one, and the derivative is undefined if two eigenvalues are equal.

While it is necessary to consider the first three eigenvectors of $\mathbf{v}^{-1}(\mathbf{m}_\mathbf{C})$, it is not necessary to consider further eigenvectors. In fact, higher order eigenvectors can be unstable if the corresponding eigenvalues are similar. While this may be an interesting information, it does not influence the PCA projection. Since we project to 2D, only the stability of the first two eigenvectors is relevant, i.e., the first 2 Eigenvalues must be well-distinguishable. Because of this, we leave the higher-order eigenvectors unconsidered.



(a) For this data set, the approach of Görtler et al. is reliable.

(b) For this data set, the approach of Görtler et al. is still reliable.

(c) For this data set, the approach of Görtler et al. is *not* reliable, and sampling is unavoidable.

Fig. 6: 2D illustrating example with $N = 4, n = 2$. The left column of each subfigure shows Görtler et al. [13], the right column of each subfigure shows our approach. Top row: Uncertain data and global covariance matrix. Center row: PCA for $m = 2$. Bottom row: PCA for $m = 1$ with some height added for better readability.

8 Results

8.1 Illustrating 2D Data Set

We illustrate our approach on a simple data set with $n = 2, N = 4$, similar to the one used by Görtler et al. [13]. We use similar data and a similarly their style of presenting. In all comparisons, their method is shown in the left column and ours in the right column. Figure 6a (top left) shows the four uncertain normally distributed data points as colored ellipses together with the major eigenvectors. Figure 6a (center left) shows the alignment according to the major eigenvectors and is therefore the linear map from $n = 2$ to $m = 2$ determined by the PCA. Figure 6a (bottom left) shows the projection to 1D (i.e., $m = 1$) where a small vertical offset is added to avoid clutter. Figure 6a (top right) shows the same 4 uncertain data points in 2D data space, but now by colored contours of the corresponding PDF. In addition, we visualize the uncertain global mean $m(\mathbf{x})$ in (48) by thin black contours of the corresponding PDF, and we depict the uncertain eigenvectors $v(\mathbf{x})$ of the uncertain global covariance matrix as defined in (64) as 2D polar plot (68) using a thick black line. Note that the uncertain eigenvector plot has strong cone-shaped peaks in the directions of the eigenvectors of the global covariance matrix. This means that the results of Görtler et al. [13] are reliable.

To verify this, we show the projected PDF of each data point by our sampling-based approach in Figure 6a (center right). The contours are not ellipses any more, but still close to the ellipses in Figure 6a (center left). This confirms that [13] is reliable for this data set. Figure 6a (bottom right) shows the projection to 1D. Note that the center right and the bottom right parts of Figure 6a are the ones that are expensive to compute, as they are based on a sampling as described in Section 6. We used the parameters $R = 0.2, U = 250\,000$ for sampling for all examples in this paper. For all other parts of Figure 6a their computation is inexpensive, because they can be evaluated in closed form. This includes the uncertain global mean and uncertain eigenvectors in Figure 6a (top right). For showing contours of 2D PDFs, we used $c_1 = 0.97, c_2 = 0.78, c_3 = 0.30$ such that inside the outermost contour we have 97% of all realizations, and similarly for c_2, c_3.

Figure 6b shows a similar data set but with a scaled covariance for each uncertain data point. Here the uncertain eigenvalues still have distinguished peaks (thick black polar plot in Figure 6b (top right)) but not a sharp as in Figure 6a. Nevertheless, it still predicts that Görtler et al. is reliable: the ellipses in Figure 6a (center left) are still sufficiently similar to the PDF contours shown (center right).

In Figure 6c however, as the uncertainty of the input data is once more increased by scaling the local covariance matrices of the data points the uncertain eigenvectors (Figure 6c (top right)) now do not show strong peaks anymore. This means that the approach of Görtler et al. is not reliable any more: the ellipses in Figure 6c (center left) differ significantly from the PDF contours in Figure 6c (center right) which are not even convex anymore. In this case, our expensive sampling-based approach (center right) is unavoidable.

8.2 Example from Section 4.3

Figure 7 shows the result of our approach for the example in Section 4.3 consisting of 3 points in 2D. This is a special case because only one point exhibits uncertainty, and does so in only one direction. In fact, this example was constructed as the simplest possible example to illustrate the issues of not considering the uncertainty of the global covariance matrix. Here, the uncertainty of a single point causes a switch of first and second principal eigenvectors without any changes in their directions. Such a behavior cannot be detected by our uncertain eigenvector $v(\mathbf{x})$ in Figure 7 (top left) because it cannot distinguish between first and second principal eigenvector (see remark at the end of Section 7.3). However, the switch of first and second principal eigenvector by data uncertainty results in completely different projected PDFs for the data points (Figure 7 (center right)) than the elliptic projections by Görtler et al. [13] (center left).

8.3 Iris Data Set

We consider the well-known Iris data set [10] with $N = 150, n = 4$, which is labeled into 3 classes. Similar to Görtler et al. [13], we assume a high-dimensional normal distribution of the data points in each labeled class. This means that the input for their and for our approach are 3 uncertain data points in 4D. Figure 5 shows the covariance stability glyph (69) for the data set. This reveals strong deviations from distinct peaks in the xy-plane, which indicates that the second major eigenvector of the global covariance matrix is unstable. This means that we consider the projection by Görtler et al. as unreliable.

To verify this, we compare their projection to our sampling-based approach in Figure 5, which reveals significant differences. In particular, this is an example where the contours of the projected PDFs are not only not convex anymore but may even split into disconnected components. The multiple local maxima of the of the projected PDF are aligned across the x-axis, which confirms that the second major eigenvectors are unstable and we therefore have stronger deviations of the projection in the direction of the y-axis.

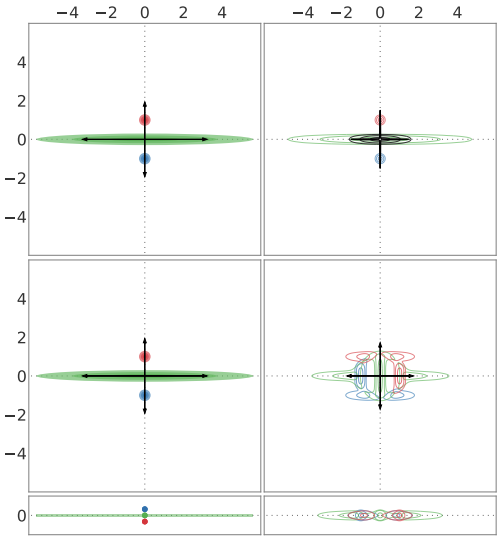


Fig. 7: 2D example from Section 4.3 with $n = 2$: Left column shows Görtler et al. [13]. Right column shows our approach. Top row: Uncertain data and global covariance matrix. Center row: PCA for $m = 2$. Bottom row: PCA for $m = 1$.

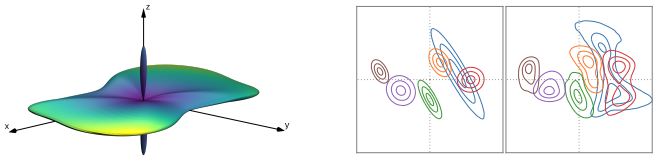


Fig. 8: Left: The covariance stability glyph for the STUDENT data set. Center: Projection by Görtler et al. [13]. Right: Our sampling-based method.

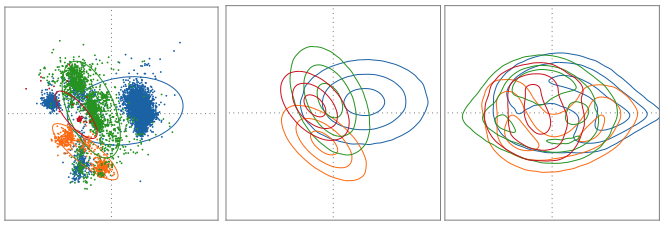


Fig. 9: ANURAN CALLS data set. Left: Clustered data, reproduced from Görtler et al. [13]. Center: Their projection as expected (missing in [13]). Right: Our sampling-based projection.

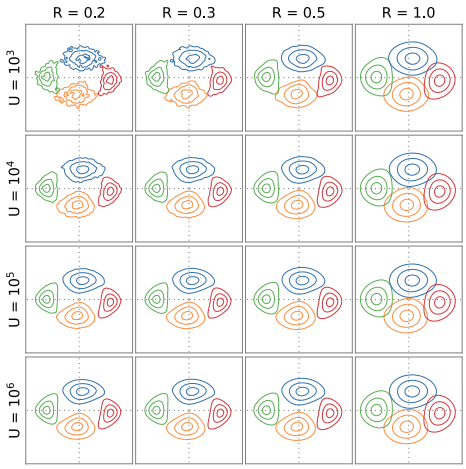


Fig. 10: Parameter study for sampling with radius R (varying horizontally) and number of samples U (varying vertically). Higher values of R give smoother reconstructions but may smooth out sharp detail. The higher U, the higher the quality of the sampled PDF.

8.4 Student Data Set

The STUDENT data set was introduced in [8]. We use an adopted version described in [13] and [41] that models all data uncertainties as normal distributions for $n = 4, N = 6$. The covariance stability glyph shown in Figure 8 (left) reveals a strong peak in the direction of the z-axis but no distinct peaks along other axes. This means that the first major eigenvector of the global covariance is rather stable, while the second one is not.

Figure 8 shows that because of this the uncertainty in the PCA by Görtler et al. [13] (center) appears much lower than in our approach (right), which considers the uncertainty of the global covariance.

8.5 Anuran Calls Data Set

The ANURAN CALLS data set [7] contains acoustic sound features from frog recordings. We are considering four different classes of frogs in the data set and model the features as normal distributions, which results in uncertain data with $N = 4$ and $n = 22$. Figure 9 shows the clustered input data (left) as similarly depicted in [13] as well as the result of their PCA projection (center). We remark that this visualization of the projection is missing and not shown in their paper neither for clustering by family as shown here nor for the clustering by genus, they only show the clustering (left). Our sampling-based projection (right) shows projected PDFs for the data points that are significantly more complex.

9 Discussion

Parameters for sampling Our sampling-based approach described in Section 6 depends on two parameters: the radius R of the local Hann function and the number of samples U. Figure 10 shows a parameter study for the data set in Figure 6b: For a small radius and low number of samples (top left), strong artifacts are visible. Increasing R under preservation of U provides a smoother reconstruction at the cost of removing sharp detail. The smaller R the more details can be encoded. This means that high-quality PDFs require large values for U and small values for R.

Also related to the quality of the sampled PDF is the dimensionality of the data space. Clearly, the higher-dimensional the data, the higher U must be to get acceptable samplings of PDFs. The dependence of dimensionality and U is discussed in [13] and [41]. The performance of our approach is directly related to the number U of samples. In this paper, we used U between 250 000 (for illustrating 2D data set and example in Section 4.3) and 1 000 000 (for IRIS, ANURAN CALLS, and STUDENT data sets), resulting in computation between a few minutes to an hour on an non-optimized Python implementation.

Non-normal distributions Our approach assumes a normal distribution of the uncertain data points. While this is a reasonable assumption for many applications, it cannot always be assumed. If normal distribution is not a realistic assumption for the input data, our sampling-based approach is still feasible as long as a sampling of the data distributions is available. However, in this case we are not able to give an inexpensive prediction of the reliability of existing approaches, as we derived the computation of uncertain eigenvectors only for data from normal distributions.

Acknowledgments

This work was partially supported by DFG grant TH 692/21-1.

Source Code Availability

The source code of our implementation is published under https://github.com/lfriesecke/uncertainty-aware-pca-revisited/.

References

[1] A. Abbasloo, V. Wiens, M. Hermann, and T. Schultz. Visualizing tensor normal distributions at multiple levels of detail. *IEEE TVCG*, 22:975–984, 2016. 2

[2] J.-H. Ahn and J.-H. Oh. A constrained em algorithm for principal component analysis. *Neural Comput.*, 15(1):57–65, 9 pages, Jan. 2003. doi: 10.1162/089976603321043694 1

[3] C. Bishop. Bayesian pca. In M. Kearns, S. Solla, and D. Cohn, eds., *Advances in Neural Information Processing Systems*, vol. 11. MIT Press, 1998. 1

[4] R. Borgo, J. Kehrer, D. H. S. Chung, E. Maguire, R. S. Laramee, H. Hauser, M. Ward, and M. Chen. Glyph-based Visualization: Foundations, Design Guidelines, Techniques and Applications. In M. Sbert and L. Szirmay-Kalos, eds., *Eurographics 2013 - State of the Art Reports*. The Eurographics Association, 2013. doi: /10.2312/conf/EG2013/stars/039-063 2, 7

[5] A. Buja, D. F. Swayne, M. L. Littman, N. Dean, H. Hofmann, and L. Chen. Data visualization with multidimensional scaling. *Journal of Computational and Graphical Statistics*, 17(2):444–472, 2008. doi: 10.1198/106186008X318440 1

[6] C. J. C. Burges. Dimension Reduction: A Guided Tour. 2(4):275–364. doi: 10.1561/2200000002 1

[7] J. S. Cañas, M. P. Toro-Gómez, L. S. M. Sugai, H. D. B. Restrepo, J. Rudas, B. P. Bautista, L. F. Toledo, S. Dena, A. H. R. Domingos, F. L. de Souza, S. Neckel-Oliveira, A. da Rosa, V. Carvalho-Rocha, J. V. Bernardy, J. L. M. M. Sugai, C. E. dos Santos, R. P. Bastos, D. Llusia, and J. S. Ulloa. Anuraset: A dataset for benchmarking neotropical anuran calls identification in passive acoustic monitoring. *Sci Data*, 110(1):771, 2023. doi: 10.1038/s41597-023-02666-2 9

[8] T. Denoeux and M.-H. Masson. Principal component analysis of fuzzy data using autoassociative neural networks. *IEEE Transactions on Fuzzy Systems*, 12(3):336–349, 2004. doi: 10.1109/TFUZZ.2004.825990 1, 9

[9] O. Essenwanger. *Elements of Statistical Analysis*. General climatology. Elsevier, 1986. 5

[10] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(7):179–188, 1936. 1, 8

[11] T. Gerrits, C. Rössl, and H. Theisel. Towards glyphs for uncertain symmetric second-order tensors. *Computer Graphics Forum (Proc. EuroVis)*, 38(3):325–336, 2019. 2, 7

[12] P. Giordani and H. Kiers. A comparison of three methods for principal component analysis of fuzzy interval data. *Computational Statistics & Data Analysis*, 51:379–397, 11 2006. doi: 10.1016/j.csda.2006.02.019 2

[13] J. Görtler, T. Spinner, D. Streeb, D. Weiskopf, and O. Deussen. Uncertainty-aware principal component analysis. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):822–831, jan 2020. doi: 10.1109/TVCG.2019.2934812 1, 2, 3, 4, 5, 6, 7, 8, 9

[14] D. Hägele, T. Krake, and D. Weiskopf. Uncertainty-aware multidimensional scaling. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):23–32, 2023. doi: 10.1109/TVCG.2022.3209420 2

[15] G. E. Hinton and S. T. Roweis. Stochastic neighbor embedding. In S. Becker, S. Thrun, and K. Obermayer, eds., *Advances in Neural Information Processing Systems 15*, pp. 857–864. MIT Press, 2003. 1

[16] H. Hotelling. Analysis of a complex of statistical variables with principal components. *Journal of Educational Psychology*, 24:417–441, 1933. 1

[17] F. Jiao, J. M. Phillips, Y. Gur, and C. R. Johnson. Uncertainty visualization in hardi based on ensembles of ODFs. In *Visualization Symposium (PacificVis)*, pp. 193–200, 2012. 2

[18] D. K. Jones. Determining and visualizing uncertainty in estimates of fiber orientation from diffusion tensor MRI. *Magnetic Resonance in Medicine*, 49:7–12, 2003. 2

[19] Y. Koren and L. Carmel. Robust linear dimensionality reduction. *IEEE Transactions on Visualization and Computer Graphics*, 10(4):459–470, 2004. doi: 10.1109/TVCG.2004.17 1

[20] L. McInnes and J. Healy. Umap: Uniform manifold approximation and projection for dimension reduction. *ArXiv e-prints 1802.03426*, 2018. 1

[21] S. Nakajima, M. Sugiyama, and D. Babacan. On bayesian pca: automatic dimensionality selection and analytic solution. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, 8 pages, p. 497–504. Omnipress, Madison, WI, USA, 2011. 1

[22] L. Nonato and M. Aupetit. Multidimensional projection for visual analytics: Linking techniques with distortions, tasks, and layout enrichment. *IEEE Transactions on Visualization and Computer Graphics*, 6 2018. doi: 10.1109/TVCG.2018.2846735 1

[23] M. N. Nounou, B. R. Bakshi, P. K. Goel, and X. Shen. Bayesian principal component analysis. *Journal of Chemometrics*, 16(11):576–595, 2002. doi: 10.1002/cem.759 1

[24] H. Osipyan, M. Kruliš, and S. Marchand-Maillet. A Survey of CUDA-based Multidimensional Scaling on GPU Architecture. In C. Schulz and D. Liew, eds., *2015 Imperial College Computing Student Workshop (ICCSW 2015)*, vol. 49 of *OpenAccess Series in Informatics (OASIcs)*, pp. 37–45. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 2015. doi: 10.4230/OASIcs.ICCSW.2015.37 1

[25] P. Paetzold, D. Hägele, M. Evers, D. Weiskopf, and O. Deussen. Uadapy: An uncertainty-aware visualization and analysis toolbox, 09 2024. doi: 10.48550/arXiv.2409.10217 2

[26] F. Paulovich, L. Nonato, R. Minghim, and H. Levkowitz. Least square projection: A fast high-precision multidimensional projection technique and its application to document mapping. *IEEE transactions on visualization and computer graphics*, 14:564–75, 05 2008. doi: 10.1109/TVCG.2007.70443 1

[27] F. V. Paulovich, C. T. Silva, and L. G. Nonato. Two-phase mapping for projecting massive data sets. *IEEE Trans. Vis. Comput. Graph.*, 16(6):1281–1290, 2010. 1

[28] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *SCIENCE*, 290:2323–2326, 2000. 1

[29] G. Sanguinetti, M. Milo, M. Rattray, and N. D. Lawrence. Accounting for probe-level noise in principal component analysis of microarray data. *Bioinformatics*, 21(19):3748–3754, 08 2005. doi: 10.1093/bioinformatics/bti617 1

[30] B. Schölkopf, A. J. Smola, and K.-R. Müller. Kernel principal component analysis. In *Proceedings of the 7th International Conference on Artificial Neural Networks*, ICANN '97, 6 pages, p. 583–588. Springer-Verlag, Berlin, Heidelberg, 1997. 1

[31] T. Schultz, L. Schlaffke, B. Schölkopf, and T. Schmidt-Wilcke. Hifive: A hilbert space embedding of fiber variability estimates for uncertainty modeling and visualization. *CGF*, 32(3):121–130, 2013. 2

[32] M. E. Tipping and C. M. Bishop. Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society Series B*, 61(3):611–622, 1999. doi: 10.1111/1467-9868.00196 2

[33] W. S. Torgerson. Multidimensional scaling: I. theory and method. *Psychometrika*, 17(4):401–419, Dec 1952. doi: 10.1007/BF02288916 1

[34] S. Tripathi and R. S. Govindaraju. Engaging uncertainty in hydrologic data sets using principal component analysis: Banpca algorithm. *Water Resources Research*, 44(10), 2008. doi: 10.1029/2007WR006692 1

[35] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008. 1

[36] L. Van Der Maaten, E. Postma, and J. Van den Herik. Dimensionality reduction: a comparative review. *J Mach Learn Res*, 10:66–71, 2009. 1

[37] N. Vaswani, Y. Chi, and T. Bouwmans. Rethinking pca for modern data sets: Theory, algorithms, and applications [scanning the issue]. *Proceedings of the IEEE*, 106(8):1274–1276, 2018. doi: 10.1109/JPROC.2018.2853498 1

[38] K. Q. Weinberger and L. K. Saul. An introduction to nonlinear dimensionality reduction by maximum variance unfolding. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2*, AAAI'06, 4 pages, pp. 1683–1686. AAAI Press, 2006. 1

[39] D. Weiskopf. Uncertainty visualization: Concepts, methods, and applications in biological data visualization. *Frontiers in Bioinformatics*, 2, 2022. doi: 10.3389/fbinf.2022.793819 2

[40] A. Wismüller and J. A. Lee. Recent advances in nonlinear dimensionality reduction, manifold and topological learning, 2010. 1

[41] S. Zabel, P. Hennig, and K. Nieselt. Vipurpca: Visualizing and propagating uncertainty in principal component analysis. *IEEE Transactions on Visualization and Computer Graphics*, 30(4):2011–2022, 2024. doi: 10.1109/TVCG.2023.3345532 1, 2, 4, 7, 9

[42] C. Zhang, M. Caan, T. Höllt, E. Eisemann, and A. Vilanova. Overview + detail visualization for ensembles of diffusion tensors. *CGF*, 36(3):121–132, 2017. 2

A Supplemental Material

A.1 Calculation of the Covariance of Covariances

$\mathbf{C_C}$ is a $r \times r$ matrix that can be computed as

$$\mathbf{C_C}[j,k] = \frac{\mathbf{d}[j]\,\mathbf{d}[k]}{N^2} s_1 \tag{70}$$

$$s_1 = \frac{1}{2}s_2 + \sum_{i=1}^{N}\left(s_3 + \frac{N-2}{2N}s_4\right)$$

$$\begin{aligned} s_2 &= \overline{\mathbf{C}}[\mathbf{T}[j,1],\mathbf{T}[k,1]]\cdot\overline{\mathbf{C}}[\mathbf{T}[j,2],\mathbf{T}[k,2]] \\ &+ \overline{\mathbf{C}}[\mathbf{T}[j,1],\mathbf{T}[k,2]]\cdot\overline{\mathbf{C}}[\mathbf{T}[j,2],\mathbf{T}[k,1]] \\ &+ \overline{\mathbf{C}}[\mathbf{T}[j,2],\mathbf{T}[k,2]]\cdot\overline{\mathbf{C}}[\mathbf{T}[j,1],\mathbf{T}[k,1]] \\ &+ \overline{\mathbf{C}}[\mathbf{T}[j,2],\mathbf{T}[k,1]]\cdot\overline{\mathbf{C}}[\mathbf{T}[j,1],\mathbf{T}[k,2]] \end{aligned}$$

$$\begin{aligned} s_3 &= \mathbf{C}_i[\mathbf{T}[j,1],\mathbf{T}[k,1]]\cdot\widehat{\mathbf{m}}_i[\mathbf{T}[j,2]]\cdot\widehat{\mathbf{m}}_i[\mathbf{T}[k,2]] \\ &+ \mathbf{C}_i[\mathbf{T}[j,1],\mathbf{T}[k,2]]\cdot\widehat{\mathbf{m}}_i[\mathbf{T}[j,2]]\cdot\widehat{\mathbf{m}}_i[\mathbf{T}[k,1]] \\ &+ \mathbf{C}_i[\mathbf{T}[j,2],\mathbf{T}[k,2]]\cdot\widehat{\mathbf{m}}_i[\mathbf{T}[j,1]]\cdot\widehat{\mathbf{m}}_i[\mathbf{T}[k,1]] \\ &+ \mathbf{C}_i[\mathbf{T}[j,2],\mathbf{T}[k,1]]\cdot\widehat{\mathbf{m}}_i[\mathbf{T}[j,1]]\cdot\widehat{\mathbf{m}}_i[\mathbf{T}[k,2]] \\ s_4 &= \mathbf{C}_i[\mathbf{T}[j,1],\mathbf{T}[k,1]]\cdot\mathbf{C}_i[\mathbf{T}[j,2],\mathbf{T}[k,2]] \\ &+ \mathbf{C}_i[\mathbf{T}[j,1],\mathbf{T}[k,2]]\cdot\mathbf{C}_i[\mathbf{T}[j,2],\mathbf{T}[k,1]] \\ &+ \mathbf{C}_i[\mathbf{T}[j,2],\mathbf{T}[k,2]]\cdot\mathbf{C}_i[\mathbf{T}[j,1],\mathbf{T}[k,1]] \\ &+ \mathbf{C}_i[\mathbf{T}[j,2],\mathbf{T}[k,1]]\cdot\mathbf{C}_i[\mathbf{T}[j,1],\mathbf{T}[k,2]] \end{aligned}$$

for $j,k \in \{1,\ldots,r\}$ and $\widehat{\mathbf{m}}_i = \mathbf{m}_i - \overline{\mathbf{m}}$ for $i \in \{1,\ldots,n\}$.

A.2 Solutions of Unbounded Multivariate Integrals

Most derivations require finding an unbounded multivariate integral in a closed-form. We give a general formulation of this and start with univariate case

$$p(x) = \mathcal{N}_1(m,c)(x) = \frac{1}{\sqrt{2\pi c}}\, e^{-\frac{(x-m)^2}{2c}} \tag{71}$$

with mean m and variance c. We are interested in the unbounded integral of the product of p with a 4-th order polynomial in x written as Taylor expansion around m given by

$$r(x) = \sum_{i=0}^{4} \frac{1}{i!}(x-m)^i\, r_i \tag{72}$$

where r_i are the coefficients of the polynomial $r(x)$. Then we get

$$\int_{-\infty}^{\infty} p(x)\, r(x)\, dx = r_0 + \frac{1}{2}c\, r_2 + \frac{1}{8}c^2\, r_4. \tag{73}$$

The derivation (73) is straightforward, we provide a Maple sheet `maple01.txt` in the additional material. Note that (73) shows that only the even coefficients of $r(x)$ contribute to the integral. This translates to the multivariate case as

$$p(\mathbf{x}) = \mathcal{N}_n(\mathbf{m},\mathbf{C})(\mathbf{x}) \tag{74}$$

and

$$r(\mathbf{x}) = \sum_{i_1=0}^{4}\sum_{i_2=0}^{4-i_1}\cdots\sum_{i_n=0}^{4-i_1-\ldots-i_{n-1}}\left(\prod_{k=1}^{n}\frac{(\mathbf{x}[k]-\mathbf{m}[k])^{i_k}}{i_k!}\right) r_{i_1,\ldots,i_n} \tag{75}$$

with the polynomial coefficients $r_{i_1,\ldots,i_n}$ with $i_1,\ldots,i_n \in \{0,\ldots,4\}$ and $i_1+\ldots+i_n \le 4$. Then

$$\int_{\mathbb{R}^n} p(\mathbf{x})\, r(\mathbf{x})\, d\mathbf{x} = r_{0,\ldots,0} + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{C}[i,j]\, r_{o_1(i,j),\ldots,o_n(i,j)} \tag{76}$$

$$+\frac{1}{8}\sum_{i_1=1}^{n}\sum_{j_1=1}^{n}\sum_{i_2=1}^{n}\sum_{j_2=1}^{n}\mathbf{C}[i_1,j_1]\mathbf{C}[i_2,j_2]\, r_{o_1(i_1,j_1,i_2,j_2),\ldots,o_n(i_1,j_1,i_2,j_2)}$$

where $o_k(i_1,j_1,i_2,j_2)$ is the number of appearances of k in the arguments (i_1,j_1,i_2,j_2), e.g., $o_1(2,1,3,1) = 2, o_2(2,1,3,1) = 1, o_3(2,1,3,1) = 1$. The derivation of (76) is provided as Maple sheet `maple01.txt` in the additional material.

A.3 Glossary of Symbols

Symbol	Explanation
$\mathbf{T}$	auxiliary matrix for transforming a symmetric matrix into its Mandel form
$\mathbf{v}(\mathbf{C})$	Mandel form of a matrix $\mathbf{C}$, e.g. its vector representation
$\mathbf{x}_i$	n-dimensional vector/data point
$p(\mathbf{x})$	n-dimensional probability density function (pdf) with $\mathbf{x} \sim \mathcal{N}_n(\mathbf{m},\mathbf{C})$
$\mathcal{N}_n(\mathbf{m},\mathbf{C})$	n-dimensional normal distribution with mean $\mathbf{m}$ and covariance $\mathbf{C}$
$\mathbf{m}$	center/mean fo some data
$\mathbf{C}$	covariance matrix for some data obtained by calculating $\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} - \mathbf{m}\mathbf{m}^{\mathrm{T}}$
$\mathbf{U},\mathbf{U}(\mathbf{C})$	orthogonal matrix obtained by using the first m eigenvectors of $\mathbf{C}$. Used to calculate the projection of the $\mathbf{x}_i \in \mathbb{R}^n$ to $\mathbf{y}_i \in \mathbb{R}^m$
$\mathbf{y}_i$	result of applying a PCA to $\mathbf{x}_i$
$\mathbf{m}_i$	local means
$\overline{\mathbf{m}}$	the mean of local means $\mathbf{m}_i$
$\mathbf{C}_i$	local covariance matrices
$\overline{\mathbf{C}}$	average of all local covariance matrices $\mathbf{C}_i$
$\mathbf{C}_m$	covariance of the local means
$\mathbf{m}_i'$	projection of the local means by an uncertainty-aware PCA acc. to Görtler et al.
$\mathbf{C}_i'$	projection of the local covariance matrices by an uncertainty-aware PCA acc. to Görtler et al.
$\mathbb{X}$	set of n individual realizations $\mathbf{x}_i$ each drawn from independent distributions $p_i(\mathbf{x}) \in \mathbb{P}$
$\mathbb{P}$	set of n individual probability density function $p_i(\mathbf{x})$
$\widetilde{\mathbf{m}}$	mean of all realizations in $\mathbb{X}$
$\widetilde{\mathbf{C}}$	covariance of all realizations in $\mathbb{X}$
$\widetilde{\mathbf{c}}$	mandel form of $\widetilde{\mathbf{C}}$
$\widetilde{\mathbf{M}}$	squared deviation of the mean $\mathbf{m}$ from from $\widetilde{\mathbf{m}}$
$\widetilde{\mathbf{C}_\mathbf{C}}$	squared deviation of the mean of covariances $\mathbf{m}_\mathbf{C}$ from from $\widetilde{\mathbf{c}}$
$\widetilde{\mathbf{M}}$	squared deviation of the mean $\mathbf{m}$ from from $\widetilde{\mathbf{m}}$
$\widetilde{\mathbf{C}_\mathbf{C}}$	squared deviation of the mean of covariances $\mathbf{m}_\mathbf{C}$ from from $\widetilde{\mathbf{c}}$
$\mathbf{M}$	covariance of means
$\mathbf{C}$	covariance of covariance matrices
$\widetilde{\mathbf{m}_k}$	mean $\widetilde{\mathbf{m}}$ for the kth sample of $\mathbb{X}$
$\widetilde{\mathbf{U}_k},\mathbf{U}(\widetilde{\mathbf{C}_k})$	The orthogonal matrix obtained by using the first m eigenvectors of $\widetilde{\mathbf{C}}$ for the kth iteration of Monte-Carlo sampling
$v(\mathbf{x})$	probability that $\mathbf{x}$ is an eigenvector
$e(\mathbf{x},\lambda)$	probability that $\mathbf{x}$ is an eigenvector with associated eigenvector λ
$E(\mathbf{x})$	set of all matrices that have $\mathbf{x}$ as eigenvector
$E(\mathbf{x},\lambda)$	set of all matrices that have $\mathbf{x}$ as eigenvector with associated eigenvector λ