# Modeling, Quantification and Visualization of Probabilistic Features in Fields with Uncertainties

**Dissertation**

zur Erlangung des akademischen Grades des Doktors der Naturwissenschaften (Dr. rer. nat.)

am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

vorgelegt von
**Kai Pöthkow**

Berlin, 2014

Erstgutachter: Prof. Dr. Christof Schütte, *Freie Universität Berlin*
Zweitgutachter: Prof. Dr. Holger Theisel, *Otto-von-Guericke-Universität Magdeburg*

Tag der Disputation: 27. Mai 2015

# Abstract

A fundamental property of scientific data is that the true value of a quantity can not be determined with arbitrary precision. It is only possible to enclose it using intervals or characterize the *uncertainty* using probability distributions. This property is shared by all types of real-valued data, both measurements and simulation results. Examples include measurements of basic physical quantities like velocity as well as long-term temperature forecasts that are computed using climate models. The uncertainty of quantities is an important information that is often indicated using confidence intervals in tables and graphical representations like 1D plots. However, for 2D and 3D data, the uncertainty can not be adequately represented using standard visualization methods in most cases.

This thesis proposes methods to facilitate analysis and visualization of uncertain scalar, vector and tensor fields. The focus is on the extraction of spatiotemporal geometric and topological features, e.g. isocontours and critical points, from the fields. The approaches are are well founded on probability theory. We employ parametric and nonparametric random fields as mathematical models for the uncertainty and spatial correlations. The probability distributions are estimated from ensemble data that combine results of multiple simulation runs which are based on, e.g., varying simulation parameters. Furthermore, we introduce condition analysis to feature-based visualization. Condition numbers quantify the sensitivity, i.e. the amplification or attenuation of uncertainty of features relative to the uncertainty of the input fields.

We propose a generic approach to probabilistic feature extraction that is the basis for the estimation of spatial distributions of various features in uncertain fields. In this framework, probabilities for the existence of features can be computed from local marginal distributions and formal feature definitions. Numerically, the probabilities can be estimated using Monte Carlo integration. To overcome the high computational cost of this approach, we propose fast approximate methods which employ surrogate functions and lookup tables for the estimate feature probabilities. The proposed methods are evaluated qualitatively and quantitatively using uncertain fields from climate and biofluid mechanics simulations as well as medical imaging.

# Zusammenfassung

Eine grundlegende Eigenschaft von naturwissenschaftlichen Daten ist, dass der wahre Wert einer Größe nicht beliebig genau bestimmbar ist. Es ist lediglich möglich, ihn durch Intervalle einzugrenzen oder die *Unsicherheit* durch eine Wahrscheinlichkeitsverteilung zu charakterisieren. Dies gilt für alle reellwertigen Daten, sowohl für Mess-, als auch für Simulationsergebnisse. Beispiele sind Messungen von grundlegenden physikalischen Größen wie Geschwindigkeit oder auch langfristige Temperaturvorhersagen, die durch Klimamodelle berechnet werden. Die Unsicherheit von Ergebnissen ist eine wichtige Information, die in Natur- und Ingenieurwissenschaften häufig durch Konfidenzintervalle in 1D-Plots und Tabellen angezeigt wird. Im Gegensatz dazu ist es bisher bei der Visualisierung von 2D- und 3D-Daten mithilfe von Standardmethoden meist unmöglich, die Datenunsicherheit zu repräsentieren.

Diese Arbeit stellt wahrscheinlichkeitstheoretisch fundierte Methoden vor, die die Analyse und Visualisierung von Skalar-, Vektor- und Tensorfeldern mit Unsicherheiten ermöglichen. Der Fokus liegt dabei auf der Extraktion von raumzeitlichen geometrischen und topologischen Merkmalen aus den Feldern (z.B. Isokonturen und kritische Punkte). Wir nutzen parametrische und nichtparametrische Zufallsfelder, um Variabilität und räumliche Korrelation mathematisch zu modellieren. Die Wahrscheinlichkeitsverteilungen werden aus Ensemble-Datensätzen geschätzt, die mehrere Simulationsergebnisse (z.B. basierend auf variierenden Simulationsparametern) zusammenfassen. Wir untersuchen die Konditionszahlen von Merkmalsextraktionsmethoden, um die Sensitivität, d.h. die Verstärkung oder Abschwächung der Unsicherheit der Ergebnisse relativ zu Unsicherheiten in den Eingangsdaten abzuschätzen.

Wir stellen einen allgemeiner Ansatz für die probabilistische Merkmalsextraktion vor, der die Basis für die Berechnung räumlicher Wahrscheinlichkeitsverteilungen von verschiedenen Merkmalen in Skalar-, Vektor- und Tensorfeldern bildet. In diesem Framework werden Wahrscheinlichkeiten für die Existenz von Merkmalen aus lokalen Randverteilungen und formalen Merkmalsdefinitionen berechnet. Numerisch können die Wahrscheinlichkeiten durch Monte-Carlo-Integration bestimmt werden. Um den hohen Rechenaufwand dieses Ansatzes zu vermeiden, schlagen wir schnelle Berechnungsmethoden vor, wobei Merkmalswahrscheinlichkeiten näherungsweise mit Hilfe von Surrogatfunktionen bzw. Lookup-Tabellen geschätzt werden. Die vorgeschlagenen Methoden werden anhand von Daten aus Klima- und Biofluidmechaniksimulationen sowie aus der medizinischen Bildgebung qualitativ und quantitativ evaluiert.

# Acknowledgements

First and foremost I would like to thank Hans-Christian Hege for introducing me to the field of uncertainty visualization and for offering me the possibility to be part of the Department of Visualization and Data Analysis at the Zuse Institute Berlin (ZIB). His guidance, creativity and tremendous support made this work possible. I am grateful to Prof. Dr. Christof Schütte, vice president of the ZIB, for his encouragement and support. The atmosphere at the Department of Visualization and Data Analysis was inspiring and motivating and the discussions I had with my colleagues sharpened my mind and advanced our research. In particular, I want to thank Uli Homberg, Norbert Lindow, Cornelia Auer, Olaf Paetsch, Stefan Zachow, Hans Lamecker, Johannes Schmidt-Ehrenberg, Steffen Prohaska, Dagmar Kainmüller, Martin Grewe, Daniel Baum, Vincent Dercksen and Alex Kuhn. My special thanks go to Christoph Petz and Britta Weber for their commitment and extensive support, especially while we worked on the papers. It was a pleasure to be part of the team. I am indebted to all Amira developers, past and present, who created a powerful, flexible software package that proved to be a great foundation for the contributions of this thesis.

The cooperation with Leonid Goubergrits, Jens Schaller (Charité Berlin) and Leonardo Agudo Jácome (Federal Institute for Materials Research and Testing) was fruitful and I want to thank them for interesting discussions throughout the years. Their applications underlined the significance of visualization for their domains and were a great source of inspiration and motivation. I am also grateful to the colleagues in the scientific community for their helpful reviews and the productive discussions.

Last but not least, I want to thank my family for everything.

# Contents

# 1 Introduction

Numerical quantities with continuous range, like scalar values $y \in \mathbb{R}$, can be measured only with finite precision. Therefore, their exact 'true' values are unknown, i.e., all measured values of such quantities are afflicted with some uncertainty. In favorable cases, this uncertainty is small, but it is always present. This is true both for deterministic and random variables. While deterministic variables take infinitely precise values that practically cannot be determined, random variables take intrinsically random values regardless of the measurement. Thus, the 'true' values of measurands are unknowable and observations are only interpretable if, additionally to the measured values, also their uncertainties are expressed. A measured result, therefore, should *always* include two entities: the measured value and some indication of its uncertainty [TK94, Joi08].

## 1.1 Uncertainty in Science and Engineering

Almost all numerical data is affected by uncertainty because it comes either from measurements or from numerical computations that are afflicted by model uncertainty, discretization and quantization errors and influenced by boundary and initial conditions that are often based on measured data. Error estimation and analysis of error propagation, therefore, is ubiquitous in science and engineering and drawing conclusions from uncertain data is the normal case, not an exception. The most common way to represent, analyze, and deal with uncertainty is to employ methods from probability theory and statistics, see, e.g., [Fel71, For08, Lir02]. Feynman et al. stated in their famous textbook [FLS63]: "Our most *precise* description of nature must be in terms of *probabilities*."

Complementary to the probabilistic approach, some other frameworks to represent uncertainties have been developed. For example, one alternative approach is the modelling of data uncertainty using *fuzzy sets* that were introduced by Zadeh [Zad65] as an extension of the classical concept

of crisp sets. Membership functions are used to describe degrees of membership of elements of fuzzy sets. Intervals can be considered as specific fuzzy sets with full degree of membership. Data in fuzzy representation can be processed using interval arithmetic. Fuzzy sets are used in many disciplines such as geography to model various kinds of uncertainty ranging from classifications to uncertain measurements [Lod08]. *Possibility theory* was developed as an extension to fuzzy sets and fuzzy logic and is an alternative to probability theory [DP01]. However, these approaches are less commonly used compared to probabilistic methods. Thus, this thesis does not employ fuzzy/possibility theory and focuses on uncertainties described using methods from probability theory and statistics.

## 1.2 Uncertainty Quantification and Visualization: Challenges and Objectives

In order to facilitate informed decision making based on scientific data, the uncertainties that are present in the data must be *quantified*. For example, consider predictions that are computed numerically using climate models (e.g. forecasts for average temperatures). In addition to the absolute values and trends in the results, the uncertainty and variability of the data are crucial aspects that need to be considered when interpreting the predictions. Quantitative estimates of uncertainty are also important for, e.g. comparing different climate models.

Theoretical foundations and methods for uncertainty quantification have been developed in disciplines such as metrology, statistics and numerical analysis and are used in many areas of application. The majority of tables and 1D-graphs in publications in science and engineering express uncertainty or provide error estimates. Evidently, uncertainty is an important part of the information that has to be represented and communicated in order to prevent erroneous conclusions about the data. However, the majority of 2D and 3D visualization and feature extraction methods still ignore errors and uncertainties of data as well as their propagation through the different stages of data analysis. Thus, to control the propagation of uncertainties through the visualization pipeline and to convey the resulting uncertainties to the user is a major challenge for visualization research [JS03].

In particular, one important area of visual data analysis is the visualization of scalar, vector and tensor fields, either by direct display or by extraction and depiction of topological or geometrical features. We are interested in *how uncertainty affects such features*.

A key objective of this thesis is to establish a mathematical basis for modelling uncertainty of scalar-, vector- and tensor fields. We use the term *field*

for all types of these fields. Mathematically, we consider the fields as functions $y : \mathbb{R}^N \to \mathbb{R}^\ell$, where $N$ is the number of space dimensions and $\ell$ is the dimensionality of the values. For example, a physical scalar quantity like temperature in a three-dimensional domain can be modelled by a function $\mathbb{R}^3 \to \mathbb{R}$. We aim at an uncertainty model for these fields based on probability theory, since this leads to quantifiable and easily interpretable results. An important category of data for representing uncertainty are ensemble data. They are commonly used, e.g., in climate research and weather forecasting. Mathematical models for uncertain fields must be able to represent the structure of ensemble datasets well – including the probability distributions and spatial correlations.

Based on such mathematical models, we want to address the computation and depiction of uncertain equivalents to topological and geometric features. Important features in scalar fields are *isocontours* (we use this term to denote isolines, isosurfaces, and higher dimensional counterparts) or, more general, level sets. The uncertainty related to position and shape of isocontours has been discussed earlier [GR04, KWTM03, RLBS03, JS03]. However, none of these approaches were interpretable in terms of probability theory or statistics.

Among the most important features in *vector fields* are critical points (sources, saddles and sinks) and cores of swirling motion (vortices). We aim at a general framework to compute probabilities for the existence of such features from mathematical representations of uncertain vector fields. An interesting question from the application point of view is the locality of features, i.e., the question how far regions, in which some feature is notably present, are extended. Another objective of this thesis is to evaluate the proposed methods and apply them to datasets from science and engineering.

## 1.3 Contributions

This thesis is based on the work presented in these peer-reviewed papers:

- K. Pöthkow and H.-C. Hege. **Positional Uncertainty of Isocontours: Condition Analysis and Probabilistic Measures.** *IEEE Transactions on Visualization and Computer Graphics*, 17(10):1393–1406, 2011.

- K. Pöthkow, B. Weber, and H.-C. Hege. **Probabilistic Marching Cubes.** *Computer Graphics Forum*, 30(3):931 – 940, 2011.

- K. Pöthkow and H.-C. Hege. **Uncertainty Propagation in DT-MRI Anisotropy Isosurface Extraction.** In D. Laidlaw and A. Vilanova, editors, *New Developments in the Visualization and Processing of Tensor Fields*, Mathematics and Visualization, pages 209 – 225. Springer, Berlin, 2012.

- C. Petz, K. Pöthkow, and H.-C. Hege. **Probabilistic Local Features in Uncertain Vector Fields with Spatial Correlation.** *Computer Graphics Forum*, 31(3):1045 – 1054, 2012.
- K. Pöthkow, C. Petz, and H.-C. Hege. **Approximate Level-Crossing Probabilities for Interactive Visualization of Uncertain Isocontours.** *International Journal for Uncertainty Quantification*, 3:2:101–117, 2013.
- K. Pöthkow and H.-C. Hege. **Nonparametric Models for Uncertainty Visualization.** *Computer Graphics Forum*, 32(3):131 – 140, 2013.
- K. Pöthkow and H.-C. Hege. **Accelerated Probabilistic Feature Extraction Using Surrogate Functions.** *Under Review*, 2014.

During the work on this thesis the author also contributed to the paper by Goubergrits et al. [GSK*12]. The main contributions presented in this thesis are listed below.

**Mathematical Models for Uncertain Fields.** We propose to employ discrete random fields as mathematical model for uncertain fields (Chap. 3). In the simplest case all random variables in a field conform to some type of parametric probability distribution, e.g., Gaussian and are assumed to be statistically independent (uncorrelated).

However, in most applications such a simplified model does not appropriately represent the structure of the data. Specifically, the correlation structure is one of the essential properties of a random field and it has to be considered in order to compute accurate results. We propose models that take *arbitrary spatial correlations* into account. As an extension to discrete random fields we introduce methods to employ three types of nonparametric models for uncertain fields: empirical distributions, histograms and kernel density estimates (KDE). These models represent different types of distributions in a flexible manner while still considering spatial correlations.

**Sensitivity Analysis and Estimation of Uncertainty Propagation.** To describe the propagation of errors from the input data to computed features, we introduce *condition analysis* to feature-based visualization (Chap. 4). Using condition numbers, we assess the sensitivity and quantify the amplification or attenuation of uncertainty relative to perturbations of the input data by different steps in a data processing and visualization pipeline. We derive the condition number of the isocontour problem and show how average condition numbers can aid the selection of thresholds that correspond to robust isocontours (Sect. 4.2).

The second area of application for condition analysis is diffusion tensor image (DTI) data and the calculation of related scalar indices. DTI is an

important data acquisition technique for the investigation of diffusion processes of molecules, most notably for the study of neurological disorders in the human brain. In Sect. 4.3 we investigate the propagation of errors and uncertainty from the initial diffusion tensor field through the anisotropy measures (fractional anisotropy, FA, and relative anisotropy, RA) to the isosurfaces of these measures. We quantify the amplification or attenuation of uncertainty using the condition numbers of the respective anisotropy indices. Using this approach, we show that – in a first order approximation – the propagation of uncertainty for isosurface extraction using FA and RA is *equal*. We present results for phantom and brain DTI data.

**Uncertain Equivalents to Features in Scalar and Vector Fields.** We propose several local probabilistic measures as uncertain equivalents of features in scalar, vector and tensor fields. In a first step, we investigate the positional uncertainty of isocontours (Chap. 5). We assume that the data have been sampled on nodes of some mesh. For the case of statistically independent random variables, our approach works on the level of PDFs that are interpolated between sample points. We define the *isocontour density* (ICD) and the *level-crossing probability field* (LCP) that quantify positional uncertainty of isocontours for all points in a continous domain.

As an extension to these approaches, we present a general computational framework for probabilistic feature extraction (Chap. 6). Relevant features are defined for given grid entities using indicator functions. Probabilities for the existence of these features are computed as integrals over local marginal probability density functions (PDFs) and computed using Monte Carlo (MC) integration. The proposed procedure can be applied to any type of mesh, both structured and unstructured.

Based on the generic framework, we propose methods to compute spatial distributions of local features from uncertain scalar and vector fields considering the local correlation structure. These distributions can be used to display important structures of the data. For scalar random fields, we focus on level-crossing probabilities (Sect. 6.2). In order to reveal the probability for the occurrence of an isocontour of a given isolevel at some spatial location, we compute probabilities that grid cells are crossed by an isocontour. In vector-valued fields we consider critical points (sources, sinks and saddles) and swirling motion vortex cores (Sect. 6.3). But the proposed general approach can be applied to other local features in scalar, vector and tensor fields as well.

**Fast Approximation Methods.** MC integration is a straightforward way to estimate the feature probabilities. However, a major disadvantage is the high computational cost that prevents interactive data analysis. To overcome

this drawback, several approximation methods have been developed. In addition to two specific approaches for cell-wise level-crossing probabilities in discretized Gaussian fields, we propose a flexible approximation method based on *surrogate functions*. The surrogate functions are constructed from attributes of example grid cells and their feature probabilities (the training set) and can predict probabilities for new grid cells and datasets. We provide a quantitative and qualitative evaluation of the generalization performance and show that the results computed by surrogate functions approach the ground truth for increasing sizes of the training sets.

The estimated feature probabilities can be used for visualizations that not only give an impression of the uncertainty, but also allow quantitative analysis of the data. We demonstrate the utility of these methods by applying them to data from biofluid mechanics and climate research simulations. We also discuss the impact that model selection has on the respective results and give recommendations for choosing adequate probabilistic models in various areas of application.

# 2 Related Work

This chapter provides an overview of existing methods that address the analysis and visualization of fields that are affected by uncertainty. The next section discusses related work on visual representations of uncertainty and theoretical frameworks for classification and evaluation of visualization methods. In Sect. 2.2 we concentrate on feature extraction methods that take uncertainties into account. We discuss the relation of these publications to this thesis in Sect. 2.3.

## 2.1 Uncertainty Visualization

An early introduction to uncertainty visualization describing various aspects of uncertainty propagation and several visualization methods was presented by Pang et al. [PWL97]. Johnson and Sanderson considered the representation of uncertainty to be a major challenge in visualization research [JS03]. Surveys of publications on the visualization of field data that is affected by uncertainty were presented by Brodlie et al. [BAOL12] as well as Potter et al. [PRJ12]. MacEachren et al. [MRH*05] presented a review of several approaches that were developed specifically for geography and cartography to represent uncertainty of data and for improving decision making when dealing with uncertainty. Torre Zuk [Zuk08] presented a theoretical framework to aid the development and qualitative evaluation of visualizations that support reasoning under uncertainty.

In several different areas of visualization research methods to represent data uncertainty have been proposed. The visualization of ensemble data was addressed by Potter et al. [PWB*09] and Sanyal et al. [SZD*10]. Both papers present visualization tools for weather forecasts and simulated climate data. Sanyal et al. also conducted an evaluation of their tool's efficiency. Kinkeldey et al. [KMKS13] investigated the effectiveness of noise lines used as annotations to represent attribute uncertainty. Potter et al. [PKXJ12] pro-

posed to compare ansatz PDFs to histograms and compute statistical distance measures for interactive visualization. Liu et al. [LLBP12] used Gaussian mixture models to approximate large ensemble datasets for volume visualization.

Love et al. [LPK05] described methods to manipulate ensemble data, assess uncertainty propagation and adapt well known visualization methods. Methods to display uncertain data using volume rendering include special transfer functions that take mean values and variances into account [DKLP01,DKLP02]. Pfaffelmoser and Westermann [PW12,PW13a] presented methods to visualize the local as well as the global correlation structure of random fields. Yang et al. [YXK13] developed an approach to estimate covariances and cross-covariances for stochastic simulation results in 2D and visualized them using glyphs. Günther et al. introduced mandatory critical points with corresponding merge and split graphs for random scalar fields where all variables have finite support [GST14]. For uncertain multivariate data Feng et al. [FKLT10] proposed methods to display variants of scatter plots and parallel coordinate plots that take the probability density functions of the data into account.

**Visualization of Uncertain Vector Fields.** Visualizations of uncertain vector fields can be created using texture mapping approaches as proposed by Botchen et al. [BWE05]. Hlawatsch et al. [HLNW11] introduced glyphs for the static visualization of unsteady flow with uncertainty indicated by angular confidence intervals. Specific visualization techniques for uncertain flow fields include uncertainty glyphs [WPL96], stream ribbons and envelopes showing streamline uncertainty [LPSW96] and the incorporation of uncertainty in reaction-diffusion visualizations [SJK04]. The uncertainty of particle positions and movements was estimated and visualized by Lodha et al. [LFC02]. Allendes Osorio and Brodlie [AOB09] adapted LIC (Line Integral Convolution) visualization methods to indicate directional uncertainties in vector fields. Zuk et al. [ZDG*08] used glyphs and an interactive tool to visualize and explore uncertain bidirectional vector fields. Bhatia et al. [BJB*12] visualized uncertainty introduced by streamline computation of crisp vector fields. Pfaffelmoser et al. proposed an approach to estimate and visualize the uncertainty of gradient vector directions and magnitudes in 2D scalar fields [PMW13]. They provide closed form solutions for mean values and covariances, derive confidence intervals and employ those quantities for glyph and colormapping visualization methods.

**Visualization of Probabilistic Image Segmentations.** In medical applications, uncertainties of segmentations are of interest. Kniss et al. [KUS*05] presented a volume rendering approach that allows the user to interactively

explore the class probabilities of segmentations and uncertainty of surface boundaries by deferring the classification decision to the rendering stage. Uncertainty of tissue classification in medical volume data was also visualized by animation using fuzzy time-dependent transfer functions for direct volume rendering by Lundström et al. [LLPY07]. Saad et al. [SHM10] used shape and appearance knowledge to evaluate and visualize segmentation uncertainty of medical data sets. Praßni et al. [PRH10] used the uncertainty of probabilistic segmentation algorithms as a cue for the improvement of the segmentation results in a semi-automatic work flow.

## 2.2 Feature Extraction Methods

An alternative to direct display of uncertain fields is the estimation of probabilities for the existence of meaningful features in the field. Features can be defined locally, e.g. such that probabilities can be estimated for all points in a domain, or globally where distributions of spatially extended structures (e.g. streamlines in flow fields) are of interest.

**Uncertainty of Lines and Surfaces.** The uncertainty of surface shapes and positions of isolines and isosurfaces was addressed in several publications. Pang et al. [PWL97] created fat surfaces by displaying two surfaces that enclose the volume in which the true (but unknown) surface is located. Grigoryan and Rheingans [GR04] used point primitives for rendering uncertain surfaces: A large number of points were randomly displaced along the isosurface normals in a distance proportional to the uncertainty, random numbers, and a user-defined scale factor. Pauly et al. [PMG04] presented a formulation of likelihood and confidence maps that describe the possible surface reconstructions from point cloud data for the whole domain. Instead of describing the uncertainty of a single surface, certainty measures are displayed using cut planes in a volume.

Isosurfaces in uncertain data were presented by Johnson and Sanderson [JS03] where a combined volume and surface rendering was applied to display surface uncertainty. Rhodes et al. [RLBS03] used color and texture mapping on isosurfaces to indicate areas of high data uncertainty. The uncertainties of the surface's position and shape are not visualized by this approach. Note that there was no specification of the mathematical model describing the uncertainty in references [JS03] and [RLBS03]. Kindlmann et al. [KWTM03] used the magnitude of the flowline curvature as a measure for isosurface uncertainty and mapped this to surface color. Djurcilov et al. [DP00] presented contour lines that are stippled in areas of high uncertainty and continuously drawn in regions where the data is reliable. Zehner et al. [ZWK10] proposed to combine isosurfaces with additional geometry

to indicate the positional uncertainty and show spatial confidence intervals. The distribution parameters were computed from ensemble data and colormapped to mean isosurfaces.

Allendes Osorio and Brodlie [AB08] modeled the uncertainty of scalar fields using random fields. To display spatial distributions of uncertain isolines they computed the probability that the scalar value at a given position is contained in an interval between an isovalue and a second user-defined parameter. The positional uncertainty of isolines and isosurfaces can be quantified using level-crossing probabilities. A formulation of first-crossing probabilities that can be computed and visualized quickly using a ray casting approach was presented by Pfaffelmoser et al. [PRW11]. For 2D ensemble data Pfaffelmoser et al. [PW13b] proposed a different approach using nonparametric modelling to visualize the uncertainty of contour lines. Schlegel et al. [SKS12] proposed to use Gaussian process regression – also known as Kriging – for interpolation between sample points of discretely sampled and scalar valued Gaussian random fields and investigated the influence of varying parameters of correlation functions on level-crossing probabilities. Whittaker et al. [WMK13] presented a nonparametric approach for visualizing ensembles of isocontours that is based on a measure of data depth. The previous five papers present the approaches which are most closely related to the methods introduced in this thesis.

**Features in Uncertain Vector Fields.** A global approach to feature extraction is uncertain vector field topology which was presented by Otto et al. [OGHT10]. This includes the estimation of distributions for sources and sinks as well as the topological regions of the field. Subsequent work extended these methods to detect closed streamlines [OGT11a] and topological structures of 3D vector fields [OGT11b].

Probabilities for the existence of vortex cores – for which several criteria can be employed – in vector valued random fields can be computed locally using joint distributions representing the uncertainty in a given neighborhood [OT12] (published parallel to [PPH12]). Friman et al. presented methods to compute spatial distributions of path lines in blood flow measurements [FHH*11].

Most of the approaches for uncertain vector fields are based on local features of *crisp* vector fields. Particularly for flow fields there exists a large body of work, see e.g. [HH89, PVH*03, LHZP07, MLP*10]. A specific indicator of critical points in 2D and 3D vector fields is the Poincaré-Hopf index [GTS04, TG09]. Polthier and Preuß presented operators to classify vector field singularities in piecewise constant vector fields [PP02]. Centers of locally swirling flow can be detected using the approach presented by Sujudi and Haimes [SH95]. Szymczak [Szy11] proposed methods to compute

Morse decompositions given a user-specified error bound and with respect to perturbation of the piecewise constant input vector field.

**Feature Detection and Uncertainty Analysis for Diffusion Tensor MRI Fields.** There are some general approaches to the analysis of uncertainties in Diffusion Tensor Imaging (DTI). Pajevic and Basser [PB03] introduced a non-parametric statistical method, the DTI bootstrap, and used it to confirm that the tensor components are usually normally distributed due to thermal noise. They also estimated probability distributions for various other tensor-derived quantities. Koay et al. [KCPB07] presented a framework to analyze error propagation in DTI for different diffusion tensor representations. Considering objective functions for nonlinear least square optimizations they formulate error propagation equations that relate tensor-derived quantities to the diffusion-weighted MRI data. However, this method is restricted to the propagation of variances and does not directly yield the resulting probability distributions. Schultz et al. [SSSSW13] proposed to embed PDFs into a reproducing kernel Hilbert space and derived specific glyph based visualization methods.

Another area of research (indirectly related to Chap. 4 and 5 of this thesis) is the investigation of uncertainty of fiber tracks in the brain computed from DTI data. Jones [Jon03] used bootstrapping to determine confidence intervals for fiber orientations (cones of uncertainty). Anderson [And01] investigated the effects of noise in DTI data of human brains to fiber tracking, while Lazar et al. [LA03] focused on tractography in synthetic tensor fields. The sensitivity of fiber tracking results to parameter changes was investigated by Brecheisen et al. [BVPtHR09]. Friman et al. [FFW06] proposed a Bayesian approach to generate distributions of fiber tracks. The advantage of the latter approach is that prior knowledge about the fiber tracks can be incorporated using a fully probabilistic framework.

Several papers analyze the impact of noise and uncertainty on scalar DTI indices. Pierpaoli and Basser [PB96] statistically compared rotationally variant and invariant anisotropy indices. They show that for in vivo measurements the invariants are, in general, superior to the rotationally variant indices. Papadakis et al. [PXH*99] studied the signal to noise ratios (SNR) of different anisotropy measures using data from simulations and in vivo experiments. Chang et al. [CKPB07] used matrix perturbation theory to estimate the uncertainty of several DTI-derived parameters including FA, RA and the direction of the principal eigenvectors. Compared to bootstrap approaches this method requires significantly fewer diffusion weighted images. The work of Hasan et al. [HAN04] focused on the question whether FA is more robust to noise than RA. For that they derived an analytical expression that directly relates RA and FA and that can be evaluated us-

ing Monte Carlo simulation. References [PXH*99, HAN04, CKPB07] all state that, in general, *FA is superior to RA* regarding to noise immunity and uncertainty propagation. One of the aims of this thesis is to consider a further processing step and to assess the robustness of anisotropy isosurfaces, i.e. thresholding, of FA and RA.

Isosurfaces in anisotropy scalar fields generated from DTI data have been used by Zhukov et al. [ZMB*03] to create segmentations of the ventricles, the corpus callosum, and the internal capsule of the human brain. Large connected components of isosurfaces of FA have been used as segmentations of major brain structures by Schultz et al. [STS07]; they used additional information in the tensor field to automatically detect the specific brain region being represented by the isosurface segment. In a clinical study Snook et al. [SPB07] used anisotropy isosurfaces for the comparison of different stages of neurodevelopment.

## 2.3 Discussion

The term 'uncertainty' is used with different meanings and for addressing different problems, even in the narrow field of data visualization. For instance, Jänicke et al. [JWSK07] and Wang et al. [WYM08] attribute some 'average uncertainty' or 'local statistical complexity' to spatial or spatiotemporal domains to characterize *spatial variance* of data values and thereby to identify significant parts of datasets. For some applications the uncertainty related to categorical variables is important, e.g. for segmentation [LLPY07, PRH10] and the visualization geographical information [MRH*05, KMKS13].

In this thesis, however, statistical parameters or nonparametric distributions are attributed to each sample point or cell in a computational grid to express the uncertainty of data values due to *measurement errors* and other sources of uncertainty.

Some of the papers mentioned above employ ad-hoc-concepts of uncertainty which are not interpretable in terms of probability theory and statistics or fuzzy theory, e.g. [RLBS03, GR04], while the vast majority of the publications above which consider the uncertainty of data values in some spatial domain employ parametric probability distributions. However, most of the methods are restricted to Gaussian distributions and many assume the random variables to be statistically independent (uncorrelated), e.g. [OGHT10, AB08].

The approaches presented in this thesis are well-founded on concepts from probability theory. For the probabilistic modelling using discrete random fields both Gaussian and nonparametric distributions can be employed. The assumption of statistical independence between points in a domain is useful only if there is prior domain knowledge or evidence supporting a

white noise model. However, to accurately model uncertain data for many applications the consideration of spatial correlation is essential. Our methods are not custom-tailored for a specific task. The rather general approach and the flexibility related to probabilistic modelling make them useful in a broad range of application domains.

# 3

# Mathematical Models for Uncertain Fields

All data based on a measurement that is not just simple counting is uncertain. This chapter gives a brief overview of different manifestations and causes of errors or uncertainties in science and engineering, and establishes mathematical models that are well-founded in probability theory and statistics. New contributions to data visualization are the modeling with random fields employing parametric and nonparametric models and approaches to compute correct local marginal distributions from various types of fields. For general references in probability theory and statistics see e.g., Feller [Fel71], for random fields e.g., Adler et al. [ATW09], and for uncertainty in measurements e.g., Fornasini [For08]. This chapter is based on the papers [PH11, PWH11, PPH12, PH13].

## 3.1 Errors and Uncertainty

In general, data uncertainty consists of several parts. Systematic and random errors occur in all measurements. Rounding and discretization lead to additional uncertainty. It is important to distinguish between the terms *error* and *uncertainty*. In a measurement where significant random errors *can* occur, an observed value may, by chance, be very close to the true value. In this case the error is low, while the uncertainty (assessed by repeated measurements) is high [TK94].

**Systematic Errors.** An error that always occurs in the same way and extent if a measurement of some quantity is repeated, is called *systematic error*. It can be additive (constant shift) or a multiplicative (deviation of constant percentage). Reasons for systematic errors can be calibration deficits, envi-

ronmental conditions, or too simple models of the measured quantity. As the errors are constant over repeated measurements it is not possible to detect and eliminate them with only one measurement procedure. Sometimes it is possible to compare results of multiple measuring devices or methods, and thereby to minimize the systematic error. In general it is assumed that most values have an unknown systematic error that is impossible to eliminate completely.

**Random Errors.** At each repetition of a measurement *random errors* affect the results differently and randomly. Multiple observations can be statistically analyzed and the quantity can be described by statistical parameters. If the quantity has a deterministic value, i.e., is not a result of a stochastic process, several reasons can lead to random fluctuations; this includes interference of the environment with the measurement process, like background noise that occurs when e.g., temperature, humidity or vibrations influence the measurement.

**Other Reasons for Uncertainty.** Other sources of uncertainty lie in the measurement devices and the computers that process the data. The precision of a result is always limited, for instance by the number of bits of floating point variables. Quantization or discretization of continuous phenomena to sample point or grid based representations lead to uncertainty as well [PWL97].

## 3.2 Uncertainty Model

We do not distinguish between 'raw' measured data and data computed from measured data: conceptually we consider the possibly complex computation as part of the measurement process. For a brief overview of concepts from probability theory that are employed below, see Appendix A. In the following we assume occurrence of *additive* errors only, i.e. we assume that an observed value can be written as

$$\begin{aligned}\text{observed value} \;=\; &\text{true value} + \text{systematic error} \\ &+ \text{random error.}\end{aligned}$$

Thus, we do not consider *multiplicative* errors. Furthermore, as systematic errors have to be dealt with in a highly application-specific way, we assume that systematic errors have been minimized and are negligible. Hence we consider the simplified case

$$\text{observed value} \;=\; \text{true value} + \text{random error.}$$

Let $h \in \mathbb{R}$ be the true value of of the quantity of interest. We model the uncertainty, the observational errors as a random variable $Z$ such that the random variable

$$Y = h + Z \tag{3.1}$$

represents the observation. The $i$-th observation is regarded as the $i$-th realization of the random variable

$$v_i = h + Z_i \,. \tag{3.2}$$

We assume that the sample mean $\overline{Y}$ converges to the true value $h$ in a series of many observations and thus

$$\mathrm{E}(Z) = 0. \tag{3.3}$$

The random variable $Z$ is assumed to be described by a probability density function (PDF) $\varphi(y)$ such that

$$\mathrm{E}(Z) = \int_{-\infty}^{\infty} \varphi(y)\, y \,\mathrm{d}y = 0 \tag{3.4}$$

holds. From this it follows that

$$\mu = \mathrm{E}(Y) = h \tag{3.5}$$

and the random variable $Y$ has the PDF $f(y) := \varphi(y - h)$. The cumulative distribution function (CDF)

$$F(a) = \int_{-\infty}^{a} f(y)\, dy \tag{3.6}$$

provides the *probability* that a realization of $Y$ is less than or equal to $a$. The spread of the random values can be characterized by the standard deviation $\sigma = \sqrt{\mathrm{E}(Y - \mathrm{E}(Y))^2}$ or its square $\sigma^2$, the variance.

Possible distributions $f$ are, e.g., uniform or normal distributions. The normal (or Gaussian) distribution plays a fundamental role in applications, because it represents the distribution of random variables in many natural phenomena as well as the distribution of measured values of deterministic quantities. A theoretical explanation of this fact is provided by the central limit theorem, see, e.g., [Fel71].

The distributions $f_i$ can be acquired in various ways depending on the input data and the quantities one is interested in. The unknown expected value $\mathrm{E}(Y)$ can be estimated by the *arithmetic mean*; the standard deviation $\sigma$ by the *sample standard deviation* $s$. Sometimes, for example if the distributions are unknown or non-parametric, more advanced statistical methods such

as bootstrapping, jackknifing or other resampling techniques are needed to analyze the data [Fel71, MME05].

## 3.3 Uncertain Fields

As input for the analysis of *spatial data* we consider an uncertain field with values in $\mathbb{R}^\ell$, living in an $N$-dimensional spatial domain $\mathcal{I}^* \subseteq \mathbb{R}^N$. The true values at all positions $\mathbf{x} \in \mathcal{I}^*$ are assumed to be described by a continuous unknown $\ell$-valued function

$$h : \mathcal{I}^* \to \mathbb{R}^\ell, \mathbf{x} \mapsto h(\mathbf{x}).$$

The variable $\mathbf{x}$ may embody space, time and other parameters. In practice the function $h(\mathbf{x})$ can only be measured at a finite set of $M$ points $\mathbf{x}_j \in \mathcal{I}$, $j \in \{1, 2, ..., M\}$ where $\mathcal{I} \subset \mathcal{I}^*$. We assume that the points form a computational grid and are close enough to exhibit a sampling rate that exceeds the Nyquist rate. Depending on the application the data may be defined on the vertices (nodes) or other grid entities, see Sect. 3.3.2.

The uncertainties are modeled by considering the components of the $\ell$-dimensional vectors $Y_{\mathbf{x}_j}$ as random variables. These random vectors form a *parameter-discrete random field* [ATW09] written as

$$\{Y_{\mathbf{x}_j} : \mathbf{x}_j \in \mathcal{I}\}. \tag{3.7}$$

The space of values that the random variables can take on is called *state space* while the space of locations in the domain is called *parameter space* of the random field. The complete discretized random field with $\ell$-dimensional state space and $N$-dimensional parameter space can also be represented using a single random vector $\mathbf{Y} \in \mathbb{R}^{\ell M}$. For example, a 3D vector field with $\ell = 3$ can be written as

$$\mathbf{Y} = (Y_{\mathbf{x}_0}, Y_{\mathbf{x}_1}, \ldots Y_{\mathbf{x}_{M-1}})^T = (Y_{x,0}, Y_{y,0}, Y_{z,0}, Y_{x,1}, \ldots)^T. \tag{3.8}$$

### 3.3.1 Ensemble Data

In many cases the uncertainty of data sets is represented by storing $L$ realizations $v_i \in \mathbb{R}^{\ell M}$, comprising an *ensemble* $\{v_i | \, i \in \{1, 2, \ldots, L\}\}$. Each $v_i$ contains values for all $\ell$ variables at all $M$ vertices and thus represents a single obsevation (snapshot) of the field. In space $\mathbb{R}^{\ell M}$ the terms *data point*, *realization* and *ensemble member* have equivalent meaning. The realizations are acquired by repeated measurements or numerical simulations with varying input parameters. It is assumed that the space of possible realizations is sampled reasonably well by the ensemble. We assume that the distribution

of the random vector $\mathbf{Y}$ can be described by a probability density function (PDF) $f$.

As the PDF of the statistical population from which an ensemble is drawn is typically unknown, an important task is the *model selection*, which constrains the function $f$ such that known conditions, e.g., regarding smoothness are fulfilled and that $f$ can be estimated from the available data.

### 3.3.2 Probabilistic Models for Discretely Sampled Fields

In a *parametric* setting, assumptions about the type of probability distribution are made, based on knowledge of the application domain or on statistical tests. For example, assuming that $\mathbf{Y}$ conforms to a multivariate Gaussian distribution one needs to estimate its characterizing parameters, namely its mean vector $\mu \in \mathbb{R}^{\ell M}$ and covariance matrix $\Sigma \in \mathbb{R}^{\ell M \times \ell M}$. These parameters should be good estimates for the true expected values and covariances of the underlying distribution from which the ensemble was drawn.

*Nonparametric* models allow a more flexible representation of probability distributions. Nonparametric methods aim at an optimal fit for the *entire* PDF $f$, while parametric methods aim at good estimates for the *parameters* of a fixed type of PDF.

The estimated PDFs can be used to compute feature probabilities (e.g. for level crossings in scalar fields or critical points in vector fields) using Monte Carlo integration. In this thesis, we focus local features that can be identified by operators acting locally on data values in the neighborhood of a point in the field. To refer to grid entities like nodes, edges, faces and volume cells, we use the term *$\eta$-cell*: a 0-cell is a *vertex*, a 1-cell is an *edge*, a 2-cell is a *polygon*, a 3-cell is a *polyhedron*, and so on. The sampling grid is a $N$-dimensional grid composed of $N$-cells that discretize a $N$-dimensional geometric domain in $\mathbb{R}^N$. $C_\eta$ denotes the set of all $\eta$-cells of a grid. The size of the neighborhood thus depends on the feature definition and the local grid structure.

## 3.4 Gaussian Random Fields

In many applications it is assumed that the random variables are Gaussian distributed. Of course, not all uncertain scalar, vector and tensor fields are normally distributed, but for many of them this is approximately the case for fundamental reasons (central limit theorem). Whether or not a given field is Gaussian, can either be statistically tested or assured by empirical knowledge and statistical considerations. An example of the second case are measurements of blood flow and tissue velocity by phase contrast MRI; due

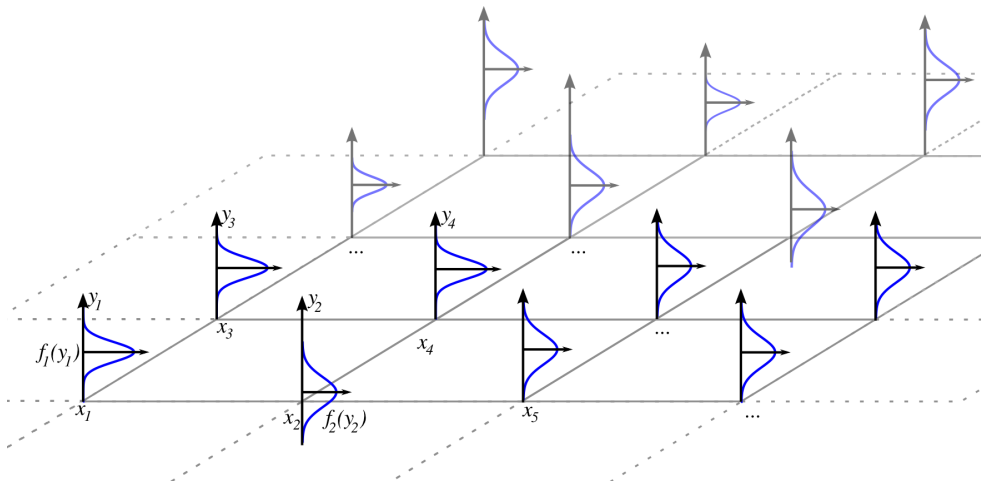



**Figure 3.1:** *Illustration of a parameter-discrete scalar random field with random variables $Y_{\mathbf{x}_j}$ associated with the nodes of a two-dimensional regular grid. Here, all local marginal distributions conform to Gaussian distributions.*

to the inherent noise in MRI the resulting vector fields are uncertain and can be shown to be correlated Gaussian random fields [FHH*11].

### 3.4.1 Joint Distributions and Correlation Structure

In general the random variables in a field are not statistically independent. For Gaussian fields the dependencies between the different points in the field can be quantified using *covariances* or *correlation coefficients*.

For a field $\mathbf{Y}$ given as a combined random vector conformable to Eq. (3.8) and consisting of $\ell M$ random variables $Y_i$ the variances and covariances can be represented by a covariance matrix

$$\Sigma = [\mathrm{Cov}(Y_i, Y_j)]_{i=1,2,\ldots,(\ell M);j=1,2,\ldots,(\ell M)}.$$

Then, $\mathbf{Y}$ conforms to a multivariate Gaussian distribution $\mathbf{Y} \sim \mathcal{N}(\mu, \Sigma)$ with $\mu = [\mathrm{E}(Y_1), \mathrm{E}(Y_2), \ldots, \mathrm{E}(Y_{(\ell M)})]$, it can be uniquely described by a joint probability density function

$$f(\mathbf{y}) = \frac{1}{(2\pi)^{(\ell M/2)} \det(\Sigma)^{1/2}} \exp\left(-\tfrac{1}{2}(\mathbf{y}-\mu)^T \Sigma^{-1}(\mathbf{y}-\mu)\right)$$

with $\mathbf{y} \in \mathbb{R}^{\ell M}$.

Note that we do not require the covariances to be defined by some analytic correlation function (e.g., exponential or linear). Both covariances

between values at different vertices in the field and, if $\ell > 1$, covariances between different values at each vertex are considered. Thus, statistical dependencies between any two values in a discretized multivariate random field independent of the distance between the vertices can be represented using this model. The full covariance matrix quadratically with the number of sample points $M$.

This model is very flexible as it allows arbitrary correlations. However, further assumptions about the structure of the correlations are made for many applications. For example, *correlation functions* model correlations that depend only on the distance $h$ between the respective points in the field. Exponential correlation functions of the form

$$R(h) = \exp(-\gamma\, h), \tag{3.9}$$

where $\gamma$ is the falloff rate of spatial correlation, are frequently used [Abr97, PRW11, SKS12].

In Sect. 6.4.2.5 we show how correlation functions can be used to model random fields such that the local marginal distributions have compact representations, i.e. few parameters and, thus, low memory complexity.

### 3.4.2 Local Marginal Distributions in Gaussian Fields

In order to compute local feature probabilities for a cell $c \in C_\eta$ the components of $\mathbf{Y}$ that do *not* correspond to that cell have to be marginalized out, yielding a local random vector $\mathbf{Y}_c$. The probability distribution of $\mathbf{Y}_c$ represents not only the point-wise $\ell$-valued uncertain data of that location, but captures also the spatial correlation of the data in its local neighborhood. Let $K_c$ be the number of degrees of freedom for cell $c$ then $\mathbf{Y}_c \in \mathbb{R}^{\ell K_c}$. To compute a marginal PDF $f_c$ from higher dimensional PDF $f$ we have to compute

$$f_c(\mathbf{y}_c) = \int f(\mathbf{y}_c, \mathbf{z})\, \mathrm{d}\mathbf{z}, \tag{3.10}$$

where we have reordered the components of row vector $\mathbf{y}$ such that vector $\mathbf{z}$ contains all dimensions of $\mathbf{Y}$ that are *not* in $\mathbf{Y}_c$, and $(\mathbf{y}_c, \mathbf{z}) = \mathbf{y}$. In general, this high-dimensional integration is difficult to perform. However, marginalization of parametric Gaussian distributions has the elegant property that the marginals are again Gaussian distributions for which the components of the means vector and covariance matrix that do not correspond to $c$ are simply deleted, see Sect. A.4 in the appendix for a proof.

The marginal distribution for $\mathbf{Y}_c$ consisting of $\ell K_c$ random variables has the reduced covariance matrix

$$\Sigma_c : C_\eta \to \mathbb{R}^{\ell K_c \times \ell K_c}. \tag{3.11}$$

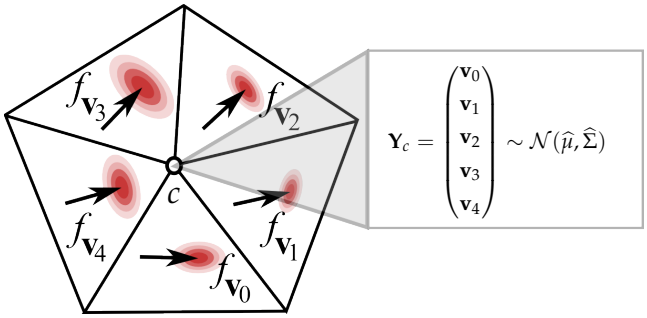



**Figure 3.2:** *Illustration for the star of a vertex $c$ in a triangulated domain with uncertain vectors defined per triangle. The marginal PDFs are indicated for each vector. The local correlated random vector $\mathbf{Y}_c$ consists of all vector components of the neighborhood of $c$.*

The total number of the marginalized covariance matrices is proportional to the number of cells. Similarly, mean values can be condensed to cell neighborhoods, yielding

$$\mu_c : C_\eta \to \mathbb{R}^{\ell K_c}. \tag{3.12}$$

The correlated random vector $\mathbf{Y}_c$ for each $\eta$-cell $c$ and the neighborhood of $c$ is defined by a multidimensional normal PDF $f_c$ that is described by $\mu_c$ and $\Sigma_c$. It is a specific property of Gaussian fields the $\Sigma_c$ and $\mu_c$ are independent of mean values and covariances that correspond to cells outside the neighborhood. Fig. 3.2 depicts the neighborhood of a 0-cell (node) of an uncertain vector field defined on the faces of a triangulated domain (2-cells).

**Example.** Consider a problem where the task is to compute probabilities for *classes of realizations of the random field* that are characterized by the fact that $m \leq (\ell M)$ random variables $Y_i$ are constrained to subsets $S_i$. Since the $\ell M$ random variables representing the random field are possibly correlated, we have to integrate the $\ell M$-dimensional density function $f(y_1, \ldots, y_{\ell M})$. Assuming that the $\ell M$ random variables have been ordered such that the constrained random variables are the first $m$ ones, we have to compute integrals of the form

$$\mathrm{P}(Y_1 \in S_1, \ldots, Y_m \in S_m) = \int_{S_1} \mathrm{d}y_1 \ldots \int_{S_m} \mathrm{d}y_m \int_{\mathbb{R}} \mathrm{d}y_{m+1} \ldots \int_{\mathbb{R}} \mathrm{d}y_{\ell M}\ f(y_1, \ldots, y_{\ell M}). \tag{3.13}$$

This means, we have to marginalize the variables $Y_{m+1}, \ldots, Y_{\ell M}$ and to compute the remaining $m$-dimensional integral. We can utilize the nice property given in Eq. (3.10) that marginalized distributions are again Gaussian distributions with the 'right' means and covariances:

$$\begin{aligned}
&\int_{-\infty}^{\infty} \mathrm{d}y_{m+1} \ldots \int_{-\infty}^{\infty} \mathrm{d}y_{\ell M} \, \frac{1}{(2\pi)^{(\ell M/2)} \det(\Sigma)^{1/2}} \exp\left(-\tfrac{1}{2}(\mathbf{y}-\mu)^T \Sigma^{-1} (\mathbf{y}-\mu)\right) \\
&\quad = \frac{1}{(2\pi)^{m/2} \det(\Sigma_c)^{1/2}} \exp\left(-\tfrac{1}{2}(\mathbf{y}_c-\mu_c)^T {\Sigma_c}^{-1} (\mathbf{y}_c-\mu_c)\right) \\
&\quad =: f_c(y_1, \ldots, y_m)
\end{aligned} \tag{3.14}$$

where $f_c$ is the density function for the reduced $m$-dimensional random vector $\mathbf{Y}_c$ and $\mathbf{y}_c$, $\mu_c$ and $\Sigma_c$ are the quantities $\mathbf{y}$, $\mu$ and $\Sigma$ with the $(\ell M) - m$ rows/columns *deleted* that correspond to the *marginalized variables* $Y_{m+1} \ldots Y_{\ell M}$. Plugging Eq. (3.14) into Eq. (3.13) yields

$$\mathrm{P}(Y_1 \in S_1, \ldots, Y_m \in S_m) = \int_{S_1} \mathrm{d}y_1 \ldots \int_{S_m} \mathrm{d}y_m \; f_c(y_1, \ldots, y_m). \tag{3.15}$$

### 3.4.3 Parameter Estimation

Given $L$ realizations $\{v_i | \, i \in \{1, 2, \ldots, L\}\}$ of the random field $\mathbf{Y}$ , i.e. a sample of observations with components $v_i^{(j)}$, $j \in \{1, \ldots, \ell M\}$, the sample means

$$\hat{\mu}^{(j)} = \frac{1}{L} \sum_{k=1}^{L} v_k^{(j)}$$

and the entries of the sample covariance matrix

$$\widehat{\mathrm{Cov}}^{(j,k)} = \frac{1}{L-1} \sum_{l=1}^{L} (v_l^{(j)} - \hat{\mu}^{(j)})(v_l^{(k)} - \hat{\mu}^{(k)})$$

for all $j, k \in \{1, \ldots, \ell M\}$ are unbiased estimates of the means $\mu^{(j)}$ and covariances $\Sigma^{(j,k)}$, respectively.

## 3.5 Nonparametric Probabilistic Models

We assume that the underlying probability distribution is sampled by $L$ realizations $\{v_{i \in \{1,\ldots,L\}}\}$ with $v_i \in \mathbb{R}^{\ell M}$. In contrast to parametric models that employ a specific type of probability distribution, the structure of a nonparametric model is not predefined but determined from empirical data. The term 'nonparametric' does not mean that these models are parameter-free. The number and type of parameters is flexible, in contrast to parametric

methods that work with a specific model that is fixed in advance. Apart from the following models there are several other possible approaches; see Scott's book [Sco92] for a comprehensive overview.

### 3.5.1 Empirical Distributions

Given a collection of $L$ data sets, an associated random vector is said to conform to an empirical distribution $Y \sim Emp(\nu_i)$ if it can only take values that are present in the collection. The corresponding PDF is parameter-free and consists of a combination of scaled $\delta$-functions

$$f(\mathbf{y}) = \sum_{i=1}^{L} \phi_i \delta(\mathbf{y} - \nu_i) \tag{3.16}$$

with weight factors $\phi_i$ and $\sum \phi_i = 1$. The CDF is a piecewise constant function with steps at the locations of the sample points [Sco92, section 2.1]. This model performs no inter- or extrapolation.

### 3.5.2 Histograms

The relative frequency of data points in bins defined by a (regular) discretization of the codomain $\mathbb{R}^{\ell M}$ locally estimates the density of samples. The PDF is piecewise constant while the CDF is piecewise linear. Random samples, e.g. for Monte Carlo methods, can be drawn from a histogram distribution by selecting a bin at random (with probabilities proportional to the numbers of points in the bins) and drawing from an uniform distribution with the extent of that bin.

### 3.5.3 Kernel Density Estimation

In many cases the population is known (or with good reasons assumed) to be smooth. Then empirical and histogram distributions are not smooth enough. Kernel density estimation (KDE) aims to approximate the true underlying distribution using a sum of basic kernel functions. The method is also called 'kernel smoothing' since it can be interpreted as a convolution of an empirical distribution with a kernel.

**Kernel Estimator.** A kernel estimator for the density of a sampled distribution is defined by

$$f(\mathbf{y}, \mathbf{H}) = \sum_{i=1}^{L} \phi_i \kappa(\mathbf{y};\, \nu_i,\, \mathbf{H}), \tag{3.17}$$

where $\kappa$ is a kernel, $L$ is the number of data points $\nu_i$ and $\mathbf{H}$ is the bandwidth matrix. The weight factors $\phi_i$ can be interpreted as prior probabilities for the

corresponding components of the estimate. For large datasets with many data points $\nu_i$ the density function is usually constructed differently, e.g. using the expectation maximization algorithm (EM). The properties of the PDF differ depending on the kernel. Random sampling of that PDF can be performed by choosing a kernel at random (for each sample) with probability $\phi_i$ and then drawing from that kernel distribution.

**Multidimensional Kernels.** The crucial parameter for a Gaussian kernel $\kappa_{\mathcal{N}}(\mathbf{y};\, \nu_i,\, \mathbf{H})$ in a multidimensional state space is the bandwidth matrix $\mathbf{H}$. There are three common choices for the type of bandwidth matrix $\mathbf{H}$: (i) scaled identity matrices $\mathbf{H} = H^2\,\mathbf{I}$, which means that each kernel is radially symmetric with constant variance in all directions, (ii) diagonal matrices

$$\mathbf{H} = \mathrm{diag}(H_1^2, H_2^2, ..., H_{\ell M}^2) \tag{3.18}$$

that contain individual bandwidths for all dimensions but do not represent any correlation, and (iii), symmetric positive definite matrices that can represent individual bandwidths and any linear dependencies between the dimensions. The third variant is the most general.

Other frequently used types of kernels include the rectangular, triangular and the Epanechnikov kernel. However, the choice of the kernel type is not as crucial for the smoothing quality as the bandwidth parameters [Sil92, p. 43].

**Automatic Bandwidth Selection.** The aim of bandwidth selection is to minimize the mean integrated squared error

$$\mathrm{MISE}(\mathbf{H}) = E\left(\int (f(\mathbf{x}, \mathbf{H}) - f^*(\mathbf{x}))^2\, d\mathbf{x}\right).$$

where $f$ is the kernel density estimate and $f^*$ is the true underlying PDF. Though $f^*$ is unknown in practice, using asymptotic analysis MISE can be approximated by *asymptotic MISE* (for $L \to \infty$), and useful information can be extracted from this quantity [Sco92, JMS96].

While for a single 1D distribution, "in the hands of an expert, interactive visual choice of the smoothing parameter is a very powerful way to analyze data" [JMS96], for more complex data this approach is not suited.

In case the sample standard deviations $\sigma_i$ are reasonable descriptions of the distribution dispersion in dimension $i$, automatic bandwidth selection methods can lead to good results. A simple bandwidth selection method

that we used in our implementation is Silverman's rule of thumb

$$H_i = \left(\frac{4}{d+2}\right)^{1/(d+4)} L^{-1/(d+4)} \sigma_i, \tag{3.19}$$

where $d$ is the dimensionality of the distribution and $L$ is the number of data points. Alternative methods include Scott's rule of thumb and computationally more expensive approaches like cross-validation [Sco92].

## 3.6 Nonparametric Discrete Random Fields

### 3.6.1 A Toy Example

We consider a discretized scalar field with 2 grid points and an ensemble consisting of $L = 50$ realizations. The joint distributions of the random variables $Y_1, Y_2$ are visualized in Fig. 3.3. In subfigure (a) the realizations are indicated by linear interpolants, and in (b) by a scatterplot with the red points depicting the positions of the $\delta$-peaks of the corresponding empirical distribution. In (c) a 2D histogram is shown and in (d) and (e) two different kernel density estimates are indicated, see Sect. 3.6.3. In (f) a *parametric* Gaussian distribution, created using maximum likelihood estimation, is shown for comparison.

### 3.6.2 Marginalization in Nonparametric Fields

Like in the case of Gaussian fields we have to solve Eq. (3.10) to obtain a local marginal random vector $\mathbf{Y}_c$ from $\mathbf{Y}$. This is necessary for the computation of local feature probabilities. While we can utilize the marginalization property given in Eq. (3.14) for Gaussian fields, in other cases of non-Gaussian fields this high-dimensional integration can be difficult to perform or even intractable. However, in the following we show that marginals of nonparametric distributions can be computed by employing model-specific approaches.

**Empirical Distributions.** Marginalizing out the spare dimensions from $\mathbf{Y} \sim Emp(\nu_i)$ is performed by projecting the points $\nu_i \in \mathbb{R}^{\ell M}$ orthogonally to the cell's subspace yielding $\nu_{i,c} \in \mathbb{R}^{\ell K_c}$. That means the spare dimensions are discarded and the marginal random vector is then $\mathbf{Y}_c \sim Emp(\nu_{i,c})$.

**Histograms.** The solution for Eq. (3.10) in the case of histogram PDFs can be obtained indirectly by first projecting the data points $\nu_i$ to get $\nu_{i,c}$ and then computing the histogram for the subspace of $c$. If the bin sizes are fixed, this

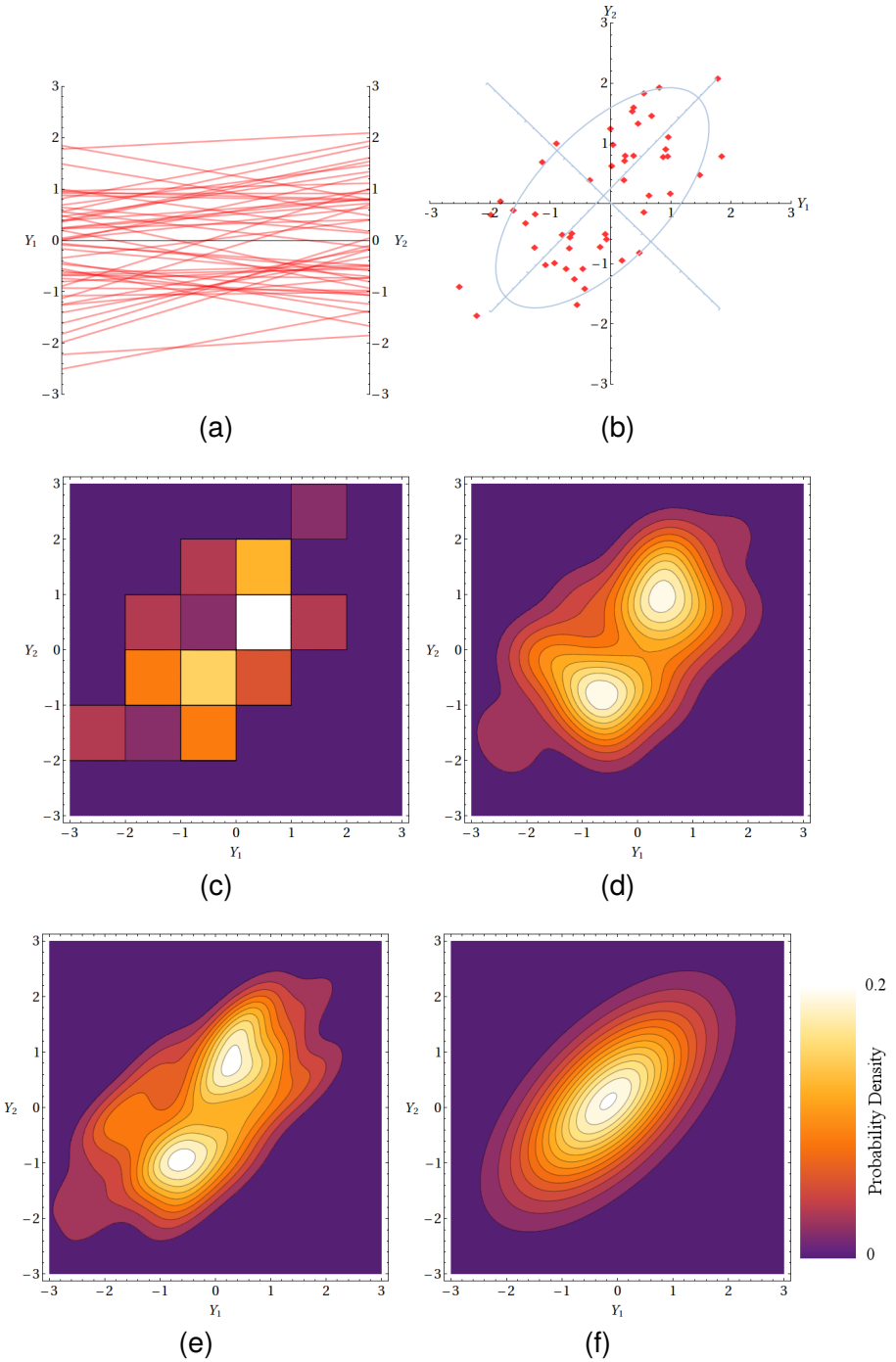



**Figure 3.3:** *A toy example of a discretized scalar field with just* 2 *grid points and an ensemble consisting of* $L = 50$ *realizations. Depicted are joint distributions of the random variables* $Y_1, Y_2$*. The subfigures (a-e) visualize* nonparametric *distributions, while subfigure (f) shows a* parametric *Gaussian distribution for comparison: (a) empirical distribution, depicted by linear interpolants (each line shows a sample); (b) empirical distribution, depicted as scatterplot; eigenvectors of the covariance matrix displayed in light blue; (c) 2D histogram; (d) KDE using a Gaussian kernel; (e) KDE using a Gaussian kernel and principal components transformation. Note that the PDF in (e) represents the correlation of the data better than that in (d).*

is equivalent to solving a discrete marginalization problem by summing up the counts of all bins over the dimensions that are marginalized out.

**Kernel Density Estimates.** To obtain a local random vector we marginalize out the other dimensions from the kernel estimates. For that we compute Eq. (3.10) for the PDF in Eq. (3.17):

$$f_c(\mathbf{y}_c) = \int \sum_{i=1}^{L} \phi_i \kappa(\mathbf{y}_c, \mathbf{z};\, \nu_i,\, \mathbf{H})\, \mathrm{d}\mathbf{z}. \tag{3.20}$$

For arbitrary kernels $\kappa$ this is, again, a difficult problem, but for Gaussian kernels we can utilize the marginalization property of parametric Gaussian distributions (see Sect. A.4 in the appendix) and by interchanging summation and integration. The PDF is then

$$f_c(\mathbf{y}_c) \quad = \quad \sum_{i=1}^{L} \phi_i \kappa(\mathbf{y}_c;\, \nu_{i,c},\, \mathbf{H}_c), \tag{3.21}$$

where $\nu_{i,c}$ are the projected data points (see above) and $\mathbf{H}_c$ is the marginal bandwidth matrix, which contains the entries of $\mathbf{H}$ that correspond to the marginal distribution. Interchanging summation and integration is possible for kernels that are valid PDFs. This is a special case of Fubini's theorem [Kal02, p. 14].

For all possible marginals $\mathbf{Y}_c$ of $\mathbf{Y}$ we have to make sure that the local PDF $f_c$ is a correct marginal of the random field. Specifically, the marginal distributions for each cell must be *consistent* over multiple neighborhoods that contain it. Due to the large number of cells in a field, manual bandwidth selection is not feasible and we have to employ automatic bandwidth selection. To make the bandwidth estimates consistent for methods like Silverman's rule we define

$$\bar{d} = E(\ell K_c) \tag{3.22}$$

to have a fixed value for the number of dimensions for all cells $\{c\}$ of interest and substitute $\bar{d}$ for $d$ when we apply Eq. (3.19). The number of data points $L$ will usually be constant for the field. This works analogously for Scott's rule.

These marginalization properties have important implications. For any distribution from which we can draw samples, an approximate solution for marginal distributions can be easily computed. Of course, this approach does not reduce the theoretical complexity of marginalization in the general case, but it is a way to obtain accurate approximations. The results are given in terms of kernel estimators and not as parameters of the distributions that were approximated.

### 3.6.3 Principal Components (PC) Transformation

In Sect. 3.5.3 the most commonly used types of bandwidth matrices were described. The most general type, which only restricts the matrix to be symmetric and positive definite, can accurately represent correlations in the distribution. A disadvantage is that for high dimensional state spaces a large number of bandwidth parameters have to be specified or estimated.

An approach that combines the simplicity of diagonal matrices with the possibility to represent correlation in KDE is to perform a *principal component transformation* [SS04]. Before estimating the density, a principal component analysis (PCA) of the data points $\nu_i$ is computed. Rewriting the data points as column vectors (with the empirical mean $\mu$ subtracted) of a matrix

$$\mathbf{A} = \begin{bmatrix} (\nu_{1,c} - \mu) & (\nu_{2,c} - \mu) & \cdots & (\nu_{L,c} - \mu) \end{bmatrix}$$

we can compute the Karhunen-Loève transform

$$\mathbf{B} = \mathrm{KLT}(\mathbf{A}) = \begin{bmatrix} \nu'_{1,c} & \nu'_{2,c} & \cdots & \nu'_{L,c} \end{bmatrix}.$$

Now the transformed points $\nu'_{i,c}$ are centered and given with respect to the PCA modes, i.e. the basis given by the eigenvectors of the covariance matrix of the data is used. The matrix $\mathbf{m}$ resulting of the KLT describes the transformation between the original coordinate system and the PCA modes. The data points can be mapped using

$$\nu'_{i,c} = \mathbf{m}\,(\nu_{i,c} - \mu) \quad \text{and} \quad \nu_{i,c} = (\mathbf{m}^{-1}\,\nu'_{i,c}) + \mu.$$

We perform KDE for the transformed points $\nu'_i$ and with respect to the PCA modes. The principal components are uncorrelated. Thus, we can employ diagonal bandwidth matrices given by Eq. (3.18) without any unwanted loss of correlation.

Fig. 3.3 (d) shows a kernel density estimate using a Gaussian kernel with diagonal bandwidth matrix for the original basis, while (e) shows a density estimate computed with respect to the eigenvector basis (using PC transformation). Density estimation on the transformed data points results in a PDF that represents the correlation of the data much better than the PDF estimated directly for the original basis.

## 3.7 Discussion

In this chapter we established models for uncertain fields that are well-founded on stochastic methods. Our approach is suitable for all applications which acquire data using standard uncertainty estimation and for which the

assumptions stated in Sect. 3.1 and 3.3 are appropriate. Parametric models allow a compact representation of uncertainty and can be analysed using a vast amount of statistical tools. However, in contrast to previous methods that were restricted to Gaussian fields we also presented a more flexible non-parametric approach that is able to work with various types of distributions. For KDE we proposed an approach to compute correct (consistent) marginal distributions, perform a principal component transformation in order to efficiently capture correlations and use automatic bandwidth selection to obtain a model for local feature extraction.

The task of model selection where the aim is to find the optimal probabilistic model for a given dataset and considering other application specific constraints is discussed in the context of actual results from probabilistic feature extraction methods in Sect. 6.5.

# 4

# Condition Numbers and Sensitivity Analysis

Before we investigate uncertain equivalents of features we address the question of how uncertainty is propagated from the input data to the solution of a numerical computation. Let us consider a problem where we have to compute some quantity ("feature") $\rho \in \mathbb{R}^m$ from input $\alpha \in \mathbb{R}^n$. Since $\alpha$ is not exactly known, we should instead consider an input set $D$ that contains all perturbed inputs $\tilde{\alpha}$, i.e. instead of a pointwise mapping $\alpha \mapsto \rho(\alpha)$ a set-valued mapping $\rho : D \to E = \rho(D)$. The effect of perturbations of input data on the output quantities – also called the *condition* of a problem $(\rho, \alpha)$ – can be expressed by some measure of a ratio of output versus input sets [DH03]. Parts of this chapter are based on the papers [PH11] and [PH12].

## 4.1 Condition Numbers and the Propagation of Uncertainty

A *condition number* describes the sensitivity of a solution for a given problem to perturbations of the input data, *independently* of the algorithm and the character of the perturbations. Let $\|\cdot\|$ be norms in $\mathbb{R}^m$ and $\mathbb{R}^n$. If a perturbation $\varepsilon$ distorts the input $\alpha$ to $\alpha + \varepsilon$ then the *absolute normwise condition* of a problem $(\rho, \alpha)$ is defined (in a first order approximation) as the smallest number $\kappa_{abs} \geq 0$ with the property that there is a real number $\delta > 0$ so that for all $0 < \|\varepsilon\| < \delta$ the inequality

$$\|\rho(\alpha) - \rho(\alpha + \varepsilon)\| \leq \kappa_{abs} \|\varepsilon\| \tag{4.1}$$

holds. A problem is said to be *well-conditioned* if $\kappa_{abs}$ is low and it is said to be *ill-conditioned* if $\kappa_{abs}$ is high. The exact meaning of low and high depends on the problem at hand.

If we assume that $\rho(\alpha)$ is (totally) differentiable, because of the mean value theorem the condition number can be calculated through the derivative:

$$\kappa_{abs} = \|\nabla_\alpha \rho(\alpha)\|, \tag{4.2}$$

where $\nabla_\alpha$ is the gradient in parameter space $\mathbb{R}^n$. The relative condition $\kappa_{rel} = \frac{||\alpha||}{||\rho(\alpha)||}\kappa_{abs}$ describes the propagation of relative errors.

The condition number $\kappa_{abs}$ is a dimensionful quantity. Its unit of measurement is [unit of output data] per [unit of input data]. The quantity relates the errors of (in general) different dimensions; loosely speaking, $\kappa_{abs}$ can be seen as the *amplification factor* of input errors. Its absolute size therefore depends on the units used. Condition numbers in different datasets therefore can be compared only if the same units of measurements are used in each dataset.

## 4.2 Condition Analysis of the Isocontour Problem

Before we investigate uncertainty propagation of isocontour extraction, crisp (or "certain") level sets and isocontours are revisited. A *scalar field* is a function $y : \mathbb{R}^N \to \mathbb{R}$ with $\mathbf{x} \mapsto y(\mathbf{x})$ and its *gradient* is the vector field $\nabla y(\mathbf{x}) = \left(\frac{\partial y(\mathbf{x})}{\partial x_1}, \ldots, \frac{\partial y(\mathbf{x})}{\partial x_N}\right)$, where $x_1 \ldots x_N$ are the Cartesian coordinates. The points $\mathbf{x}_c \in \mathbb{R}^N$ where $\|\nabla y(\mathbf{x}_c)\| = 0$ are called *critical points* of $y$. A *level set* of $y$ for constant level $\vartheta \in \mathbb{R}$ is defined as

$$\Omega = \{\mathbf{x}_s \in \mathbb{R}^d : y(\mathbf{x}_s) = \vartheta\}. \tag{4.3}$$

or $\Omega = y^{-1}(\vartheta)$.

In computer graphics and visualization *isocontours* and *isosurfaces* are extracted from scalar fields using e.g. marching squares or marching cubes [LC87] algorithms. These methods only find $(N-1)$-dimensional intersections of $y$ with a given isovalue $\vartheta$. The resulting isocontours do not necessarily contain all points of the true level set. In particular they do not contain the complete set of critical points $\mathbf{x}_c$ with $y(\mathbf{x}_c) = \vartheta$ if these points form a $N$-dimensional plateau.

To determine which parts of an isocontour are the result of well or ill-conditioned computation we determine the absolute normwise condition for the isocontour problem. By means of the condition number we can assess where potential errors and uncertainty are attenuated or amplified even if *no* information about the uncertainty of the data is available. Additionally the isovalues leading to contours that, on average, are well or ill-conditioned are of interest. This information can be used to aid the selection of thresholds for isocontour extraction or segmentation.

To extract isocontours from a scalar field we have to solve $y(\mathbf{x}) = \vartheta$ to find the level-crossing points. If the gradient $\nabla y(\mathbf{x})$ is invertible, then, according to the inverse function theorem, we can write $y^{-1}(\vartheta) = \mathbf{x}$. The derivative satisfies

$$\|(y^{-1})'(\vartheta)\| = \|\nabla y(\mathbf{x})\|^{-1}. \tag{4.4}$$

Thus, the absolute normwise condition of the problem $(y^{-1}, \vartheta)$ is

$$\kappa_{abs}(\mathbf{x}) = \|\nabla y(\mathbf{x})\|^{-1}. \tag{4.5}$$

Fig. 4.1 shows an example of a one-dimensional function. The calculation of $x_1$ is well-conditioned while the calculation of $x_2$ is ill-conditioned. In other words, the position of the level crossing at $x_1$ is less prone to perturbations of $y$ than the position of $x_2$. With Eq. (4.5) we can see that in case of a plateau in $y$, where the norm of the gradient is zero, the extraction of an isocontour is an ill-posed problem. The condition number $\kappa_{abs}$ provides an estimate for the propagation of input error to positional error of the isocontour.

### 4.2.1 Average Condition Numbers

To investigate which of the isocontours are well- or ill-conditioned we compute average condition numbers with respect to the possible isovalues. For that we define the total condition of the isocontour $y^{-1}(\vartheta)$

$$\hat{\kappa}_{abs}(\vartheta) = \int_{y^{-1}(\vartheta)} \kappa_{abs}(\mathbf{x}) \, \mathrm{d}\omega$$

and use the area (or length)

$$a(\vartheta) = \int_{y^{-1}(\vartheta)} 1 \, \mathrm{d}\omega$$

of that isocontour to calculate the average condition number. We are interested in the condition of isocontours regardless of their size. So we choose to weight the average condition by the average area because $\hat{\kappa}_{abs}(\vartheta)$ scales with the size of the isocontour. For the isosurfaces in the interval $[\vartheta_1, \vartheta_2] \subset \mathbb{R}$ the *average condition number* is

$$\bar{\kappa}_{abs}([\vartheta_1, \vartheta_2]) = \frac{\int_{\vartheta_1}^{\vartheta_2} \hat{\kappa}_{abs}(t) \, \mathrm{d}t}{\int_{\vartheta_1}^{\vartheta_2} a(t) \, \mathrm{d}t}. \tag{4.6}$$

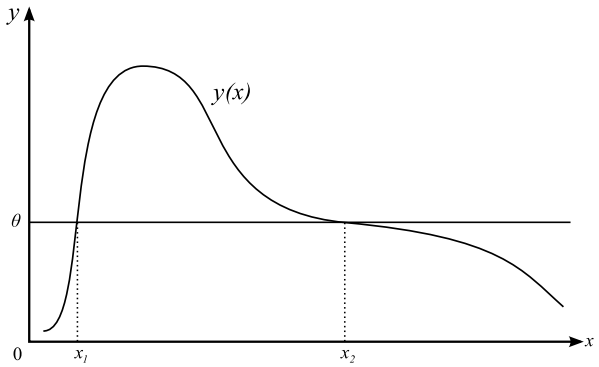



**Figure 4.1:** *Calculation of the intersection point $x_1$ of $y(x)$ with threshold $\vartheta$ is well-conditioned, while calculation of $x_2$ is ill-conditioned, i.e., the position of $x_1$ is less sensitive to perturbations of $y$ than the position of $x_2$.*

Federer's Coarea formula [Fed69], c.f. also [SSD*08]

$$\int_{\mathbb{R}} \int_{y^{-1}(t) \cap I^*} q(\mathbf{x}) \, \mathrm{d}\omega \, \mathrm{d}t = \int_{\mathcal{I}^*} q(\mathbf{x}) \, \|\nabla y(\mathbf{x})\| \, \mathrm{d}V, \tag{4.7}$$

where $q : \mathcal{I}^* \to \mathbb{R}$ is any function defined on the same domain $\mathcal{I}^*$ as $y$, allows us to express integrals over level sets as integrals over the domain.

Restricting the thresholds to the interval $[\vartheta_1, \vartheta_2]$ we can rewrite Eq. (4.6) using integrals over $V_t = \{\mathbf{x} \in R^d \mid \vartheta_1 \leq y(\mathbf{x}) \leq \vartheta_2\}$:

$$\bar{\kappa}_{abs}([\vartheta_1, \vartheta_2]) = \frac{\int_{V_t} 1 \, \mathrm{d}V}{\int_{V_t} \|\nabla y(\mathbf{x})\| \, \mathrm{d}V}. \tag{4.8}$$

This formulation is numerically more convenient than integrals over surfaces, since calculating the sum over potentially diverging values of $\kappa_{abs}(\mathbf{x})$ can lead to overflows or accuracy issues. It also saves us from having to generate many isosurfaces. Instead, we approximate the result of (4.8) by computing sums over the discretized scalar field and its gradient magnitude field.

### 4.2.2 Examples

Fig. 4.2 shows plots with average condition numbers $\bar{\kappa}_{abs}$ for small ranges of possible isovalues and isosurfaces with the condition mapped to surface color for two scalar fields given on uniform grids. Arrows indicate the used isovalues in the plots. On (spatial) average, the isocontours for isovalues at the minima of $\bar{\kappa}_{abs}$ have the best condition numbers while those at the maxima have the worst average condition numbers in each dataset. The

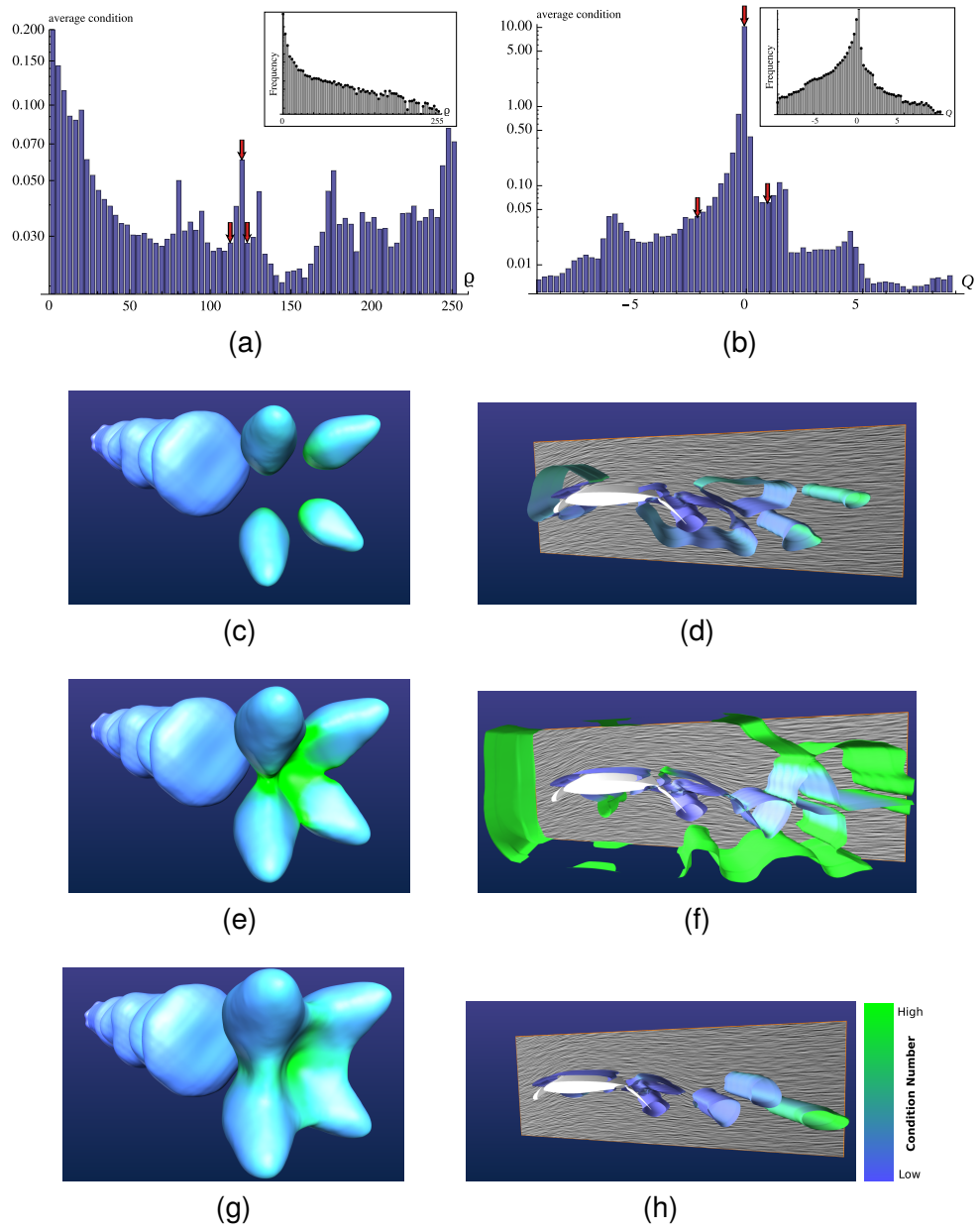



**Figure 4.2:** *The average condition number is shown for isosurfaces in the fuel dataset (a) and the Q field of a flow dataset (b). For comparison the histograms are included. Figures (c), (e) and (g) show isosurfaces for $\vartheta = 113$, $\vartheta = 119$ and $\vartheta = 125$ in the fuel dataset; (d), (f) and (h) show isosurfaces for $\vartheta = -2$, $\vartheta = 0$ and $\vartheta = 1$ in the $Q$ field. The isosurfaces in (e) and (f) have relatively high average condition numbers and should be considered less reliable than the other surfaces.*

selection of thresholds is often based on histograms so, for comparison, we also depict the histograms in the respective subfigures.

Our first example is the fuel dataset (freely available at http://www.volvis.org), describing the density of fuel during an injection process. The plot showing average condition numbers in Fig. 4.2a was computed using 70 disjoint equally sized intervals on the range $(0, 255)$. It indicates that isovalues close to zero yield diverging condition numbers. Additionally there are several less significant peaks. These peaks are not present in the histogram which means that the average condition plot is better suited for the selection of reliable thresholds. In Fig. 4.2c, 4.2e and 4.2g isosurfaces for $\vartheta = 113$, $\vartheta = 119$ and $\vartheta = 125$ in the fuel dataset are shown. The isosurface close to a saddle point (Fig. 4.2e) is relatively ill-conditioned.

Fig. 4.2b represents average condition numbers of isosurfaces in the $Q$ field (Okubo-Weiss parameter [Hun87]) of a smoothed vector field. We used a single timestep from a simulation of flow around an airfoil. Again the average condition numbers were computed using 70 disjoint equally sized intervals on the range $(-9, 9)$. The plot exhibits several peaks, some of which are not visible in the histogram. In Fig. 4.2d, 4.2f and 4.2h isosurfaces for $\vartheta = -2$, $\vartheta = 1$ in the $Q$ field enclose areas of high strain and high vorticity respectively.

### 4.2.3 Discussion

Isosurfaces for $\vartheta \approx 0$ in the $Q$ field (similar to Fig. 4.2f) have previously been used to show the separate regions of dominant strain and vorticity [SWTH07, RB09]. However, the isovalue $\vartheta = 0$ is a possibly problematic choice because it yields the most ill-conditioned isosurface in the whole dataset because parts of the level set lie in areas with very low (or zero) gradient magnitude. Generally speaking, all isosurfaces close to critical points, plateaus in particular, or other areas of low gradient magnitude ("near-plateaus") are on average relatively ill-conditioned.

These results can be related to those of Bajaj et al. [BPS97] and Pekar et al. [PWH01] who identified the significant isosurfaces, specifically those that separate homogenous regions in a dataset. They show that a high average (or total) gradient magnitude is a good criterion for isosurfaces to be considered to be the most "meaningful" ones in a dataset. Empirically, we found out that isovalues corresponding to surfaces with maximal average gradient magnitude also correspond to those with the best average condition numbers.

## 4.3 Condition Analysis of Anisotropy Isosurface Extraction from DTI

Diffusion Tensor MRI provides estimates for the major orientations of water diffusion within tissue. From these, conclusions about the microstructure of the tissue can be drawn. For example, the dominant direction of anisotropic diffusion in white matter of the brain corresponds to the orientation of neural axons. Several data acquisition and processing techniques for DTI have been established and used to assess the development or pathology of white matter for a variety of diseases, see, e.g., [LBMP*01, Fil09] and, for a current overview, [LSMC11].

Diffusion tensor fields are computed from diffusion weighed MR images and usually defined on some regular grid. In Euclidean space each tensor $D(\mathbf{x}_j)$ associated with a point $\mathbf{x}_j \in \mathbb{R}^3$ can be represented by a symmetric matrix. Note that we will drop the argument $\mathbf{x}_j$ where applicable to simplify notation. In order to extend the discretely sampled tensor field to a field in a continuous domain, various reconstruction schemes have been proposed, see, e.g., reference [HSNHH10]. Each tensor is uniquely described by its eigenvalues $\lambda_1, \lambda_2, \lambda_3$ and its eigenvectors $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ that satisfy $D\mathbf{e}_i = \lambda_i \mathbf{e}_i$ for $i \in \{1, 2, 3\}$. Diffusion tensors are positive definite, i.e. all eigenvalues are positive.

The large amount of information makes analysis and visualization of tensor fields difficult. A simplified representation of a tensor field can be achieved by mapping the tensor values to scalar quantities. There is a variety of such quantities that are all based on the tensors' eigenvalues, e.g., total and mean diffusivity, relative (RA) and fractional anisotropy (FA) among others [EK06]. All these measures are invariant under rotation and scaling of the coordinate system, as well as sorting of the eigenvalues. This chapter focuses on anisotropy measures and their isosurfaces. The FA and RA are given by

$$FA = \sqrt{\frac{1}{2}} \sqrt{\frac{(\lambda_1 - \lambda_2)^2 + (\lambda_1 - \lambda_3)^2 + (\lambda_2 - \lambda_3)^2}{\lambda_1^2 + \lambda_2^2 + \lambda_3^2}} \tag{4.9}$$

and

$$RA = \sqrt{\frac{1}{2}} \frac{\sqrt{(\lambda_1 - \lambda_2)^2 + (\lambda_1 - \lambda_3)^2 + (\lambda_2 - \lambda_3)^2}}{\lambda_1 + \lambda_2 + \lambda_3}. \tag{4.10}$$

We use the term *anisotropy index* ($AI$) to denote both $FA$ and $RA$. The level sets of the $AI$ with respect to a threshold $\vartheta$ are the sets of all locations $\mathbf{x}$ where $AI = \vartheta$, also written as $AI^{-1}(\vartheta)$. We assume that regularity conditions are fulfilled that guarantee that these level sets are surfaces. For display, such level sets are usually approximated by triangulated *isosurfaces* and then displayed, or they are raycasted directly. Surfaces of this kind have

been used for the segmentation of important anatomic structures of the human brain [ZMB*03, STS07].

DTI data, like all measured data, is affected by errors and uncertainty. This means that the *true values* of measured and derived quantities are unknown and the data can be safely interpreted only if the uncertainties are considered. The impact of noise and uncertainty on the results of several data acquisition schemes and processing methods in DT-MRI has been thoroughly studied (see Sect. 2.2). However, until now authors have only investigated the uncertainty of the resulting values and have not considered uncertainty propagation during thresholding and isosurface extraction.

In this section we study the propagation of errors and uncertainty from the initial tensor field through the anisotropy measures to the isosurfaces of these measures. For this we estimate the amplification or attenuation of uncertainty by the condition numbers of the numerical problem (Sect. 4.3.3). We also address the question whether one measure is more immune to uncertainty than the other.

### 4.3.1 Uncertainty Model for DTI

We model the uncertainty of the tensor field's eigenvalues using a discrete random field $\{\lambda_i(\mathbf{x}_j)\}$, where $\mathbf{x}_j$ runs through the vertices of the sampling grid. The values are distorted by additive measurement errors, i.e. an observation $\lambda_i$ is given by

$$\lambda_i = \lambda_i^0 + \tilde{\lambda}_i \tag{4.11}$$

where $\lambda_i^0$ is the true but unknown quantity. We assume that each $\tilde{\lambda}_i$ is a zero-mean random variable. This means, we assume that the systematic errors have been minimized and can be neglected. A measure for the uncertainty of $\lambda_i$ is the standard deviation $\sigma_{\lambda i}$ or its square, the variance $\sigma_{\lambda i}^2$ of $\tilde{\lambda}_i$. The variance can be estimated analytically from the variances of the diffusion weighted MR images [KCPB07]. The specific probability distributions can be estimated using parametric and non-parametric statistical methods. For example, it has been shown that the DT eigenvalues are affected by additive Gaussian noise [PB03]. Assuming that noise is the result of a combination of many sources of measurement errors, e.g. thermal noise, vibrations and background radiation, the presence of Gaussian noise is explained by the Central Limit Theorem which states that the distribution of a sum (or mean) of $n$ random variables converges to a normal distribution for sufficiently large $n$.

### 4.3.2 Signal to Noise Ratio (SNR)

In previous work the signal to noise ratio (SNR) was used to compare the noise immunity of $RA$ and $FA$ [HAN04, PXH*99]. For this chapter two different definitions of SNR are relevant. Let $y$ be the function of interest (e.g. an image or a signal). When referring to complete datasets or images we use the average intensity $\mu_y$ and the standard deviation $\sigma_y$ of the noise to define the *global*

$$\widehat{SNR}_y = \frac{\mu_y}{\sigma_y}, \tag{4.12}$$

assuming $\sigma_y$ is constant for the whole image. A *local* SNR can be defined for all points $\mathbf{x}$ in a dataset as

$$SNR(\mathbf{x})_y = \frac{y(\mathbf{x})}{\sigma_y(\mathbf{x})}, \tag{4.13}$$

where $\sigma_y(\mathbf{x})$ is the specific standard deviation at location $\mathbf{x}$, see [Mur01, pp. 299-300].

### 4.3.3 Condition Numbers of Anisotropy Index Computation

The absolute normwise condition for FA is given by

$$\kappa_{FA}^{abs} = \|\nabla_\lambda FA\| = \left|\left| \left( \frac{\partial FA}{\partial \lambda_1}, \frac{\partial FA}{\partial \lambda_2}, \frac{\partial FA}{\partial \lambda_3} \right)^T \right|\right| = \sqrt{\frac{1}{2}} \frac{|\lambda_1 + \lambda_2 + \lambda_3|}{(\lambda_1^2 + \lambda_2^2 + \lambda_3^2)} \tag{4.14}$$

and describes the propagation of absolute errors. The relative normwise condition is given by $\kappa_{FA}^{rel} = \frac{||(\lambda_1,\lambda_2,\lambda_3)||}{FA}\kappa_{FA}^{abs}$ and describes the propagation of relative errors. Similiarly, the absolute normwise condition for RA is given by

$$\kappa_{RA}^{abs} = \|\nabla_\lambda RA\| = \left|\left| \left( \frac{\partial RA}{\partial \lambda_1}, \frac{\partial RA}{\partial \lambda_2}, \frac{\partial RA}{\partial \lambda_3} \right)^T \right|\right| = \frac{3\sqrt{\lambda_1^2 + \lambda_2^2 + \lambda_3^2}}{\sqrt{2}(\lambda_1 + \lambda_2 + \lambda_3)^2} \tag{4.15}$$

and the relative normwise condition is given by $\kappa_{RA}^{rel} = \frac{||(\lambda_1,\lambda_2,\lambda_3)||}{RA}\kappa_{RA}^{abs}$. Fig. 4.3a-4.3c show the $FA$, $RA$ and their condition numbers for a 1D tensor field varying between isotropy and linear anisotropy, i.e. $\lambda_1$ increasing (linearly) in $x$ direction while $\lambda_2 = \lambda_3 = 1$ are kept constant.

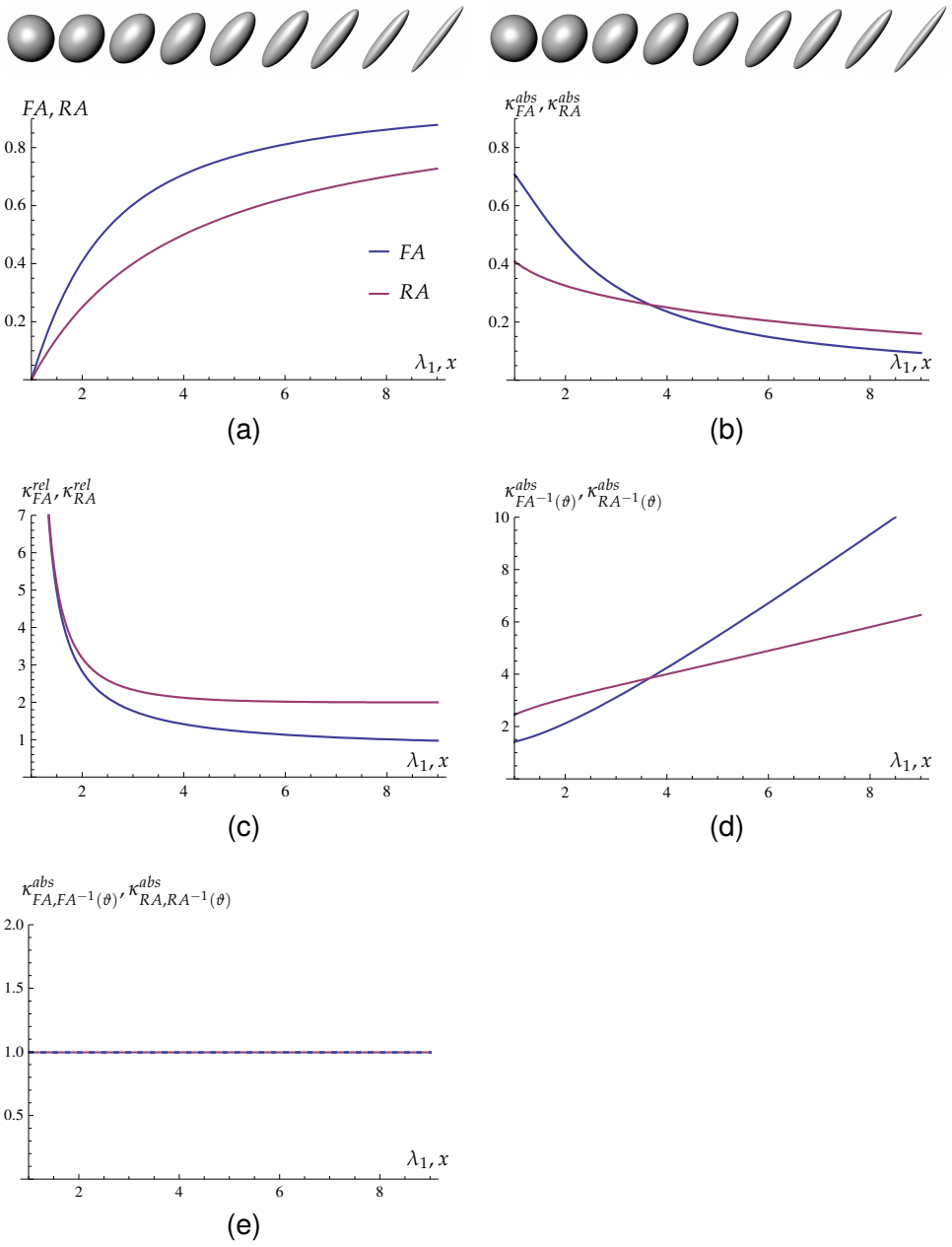



**Figure 4.3:** *In (a)-(c) the FA, RA and their condition numbers are shown for a 1D tensor field varying between isotropy and linear anisotropy, i.e. $\lambda_1$ increases linearly in x direction while $\lambda_2 = \lambda_3 = 1$ are constant. The condition numbers for isosurface extraction are shown in (d). In (e) the equality of the combined condition numbers for FA and RA is indicated. Throughout this example field the combined condition numbers are constant.*

### 4.3.4 Uncertainty Propagation

**Combined Condition Numbers.** Recall that the absolute normwise condition of the isocontour problem $(y^{-1}, \vartheta)$ is

$$\kappa^{abs}_{y^{-1}(\vartheta)}(\mathbf{x}) = \|\nabla y(\mathbf{x})\|^{-1}. \tag{4.16}$$

We denote the condition numbers for isosurface extraction from anisotropy fields by $\kappa^{abs}_{FA^{-1}(\vartheta)}$ and $\kappa^{abs}_{RA^{-1}(\vartheta)}$. They are shown in Fig. 4.3(d) for the 1D tensor field.

Using the condition numbers we can estimate the propagation of uncertainty. Let $\tilde{\lambda} = \|(\tilde{\lambda}_1, \tilde{\lambda}_2, \tilde{\lambda}_3)\|$ be a (random) perturbation of the eigenvalues. Then first order estimations of the perturbation of the results for $FA$ and $RA$ are given by

$$\widetilde{FA} = \kappa^{abs}_{FA}\, \tilde{\lambda} \quad \text{and} \quad \widetilde{RA} = \kappa^{abs}_{RA}\, \tilde{\lambda}. \tag{4.17}$$

Note that these are rough estimates because the Taylor series is truncated after the first term, i.e., covariances between the eigenvalues and higher derivatives are not considered. Analogously, the error propagation for isosurface extraction is estimated by

$$\widetilde{FA^{-1}(\vartheta)} = \widetilde{FA}\, \kappa^{abs}_{FA^{-1}(\vartheta)} \quad \text{and} \quad \widetilde{RA^{-1}(\vartheta)} = \widetilde{RA}\, \kappa^{abs}_{RA^{-1}(\vartheta)}, \tag{4.18}$$

where $\widetilde{FA^{-1}(\vartheta)}$ and $\widetilde{RA^{-1}(\vartheta)}$ are perturbations of the isosurface point positions.

Obviously these two steps can be integrated into one, resulting in a single measure for error propagation that we refer to as the *combined condition numbers*

$$\kappa^{abs}_{FA,FA^{-1}(\vartheta)} = \kappa^{abs}_{FA}\, \kappa^{abs}_{FA^{-1}(\vartheta)} \quad \text{and} \quad \kappa^{abs}_{RA,RA^{-1}(\vartheta)} = \kappa^{abs}_{RA}\, \kappa^{abs}_{RA^{-1}(\vartheta)}, \tag{4.19}$$

which relate the perturbations of the eigenvalues to the perturbations of the isosurfaces, i.e.

$$\widetilde{FA^{-1}(\vartheta)} = \widetilde{\lambda}\kappa^{abs}_{FA,FA^{-1}(\vartheta)} \quad \text{and} \quad \widetilde{RA^{-1}(\vartheta)} = \widetilde{\lambda}\kappa^{abs}_{RA,RA^{-1}(\vartheta)}. \tag{4.20}$$

We also use the condition numbers to estimate the standard deviation (or standard error) of the $FA$ and $RA$:

$$\sigma_{FA} = \sigma_\lambda \kappa^{abs}_{FA} \quad \text{and} \quad \sigma_{RA} = \sigma_\lambda \kappa^{abs}_{RA}. \tag{4.21}$$

**Comparison of *FA* and *RA*.** The graphs in Fig. 4.3a, computed for a simple 1D tensor field, show the nonlinear nature of $FA$ and $RA$. For small values

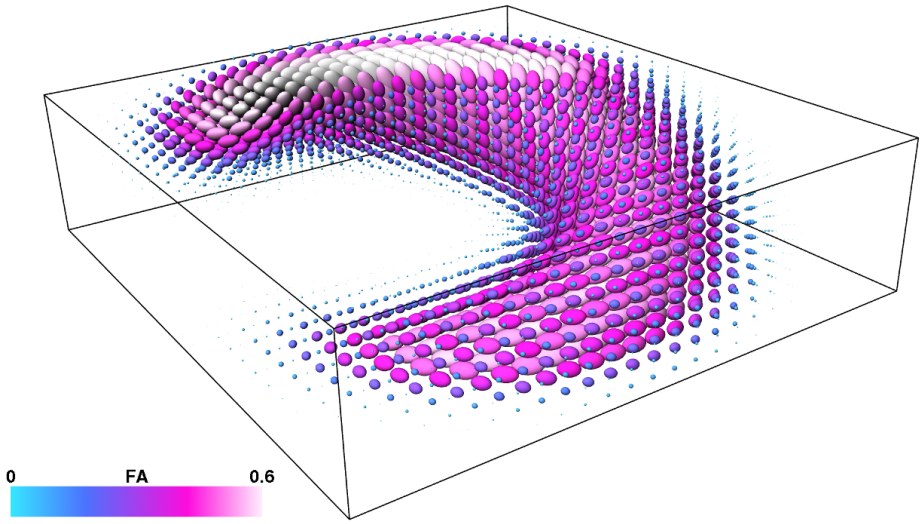



**Figure 4.4:** *The synthetic spiral DT dataset is visualized by ellipsoid glyphs with FA mapped to glyph color.*

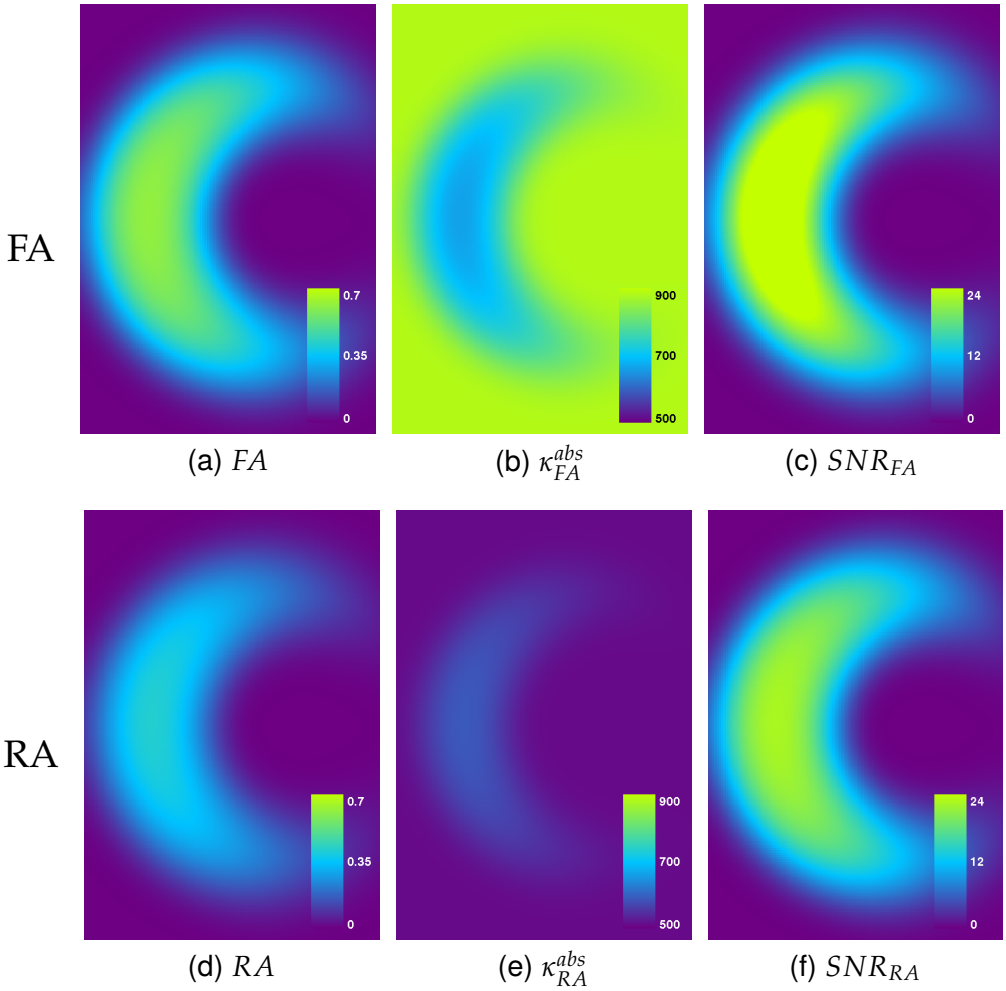
of $\lambda_1$ the curves have a steep slope, while for increasing values the slope gets more flat. This means that the sensitivity of the functions depends on the actual values of all eigenvalues. On the left side of the plot in Fig. 4.3(a) small changes of $\lambda_1$ lead to large changes of $FA$ and $RA$, i.e. perturbations are *amplified*. On the right side of the plot perturbations are *attenuated*. Both effects are stronger for $FA$ than for $RA$.

We can observe these properties in the plots of the condition numbers in Fig. 4.3b. On the left side of the plot $\kappa^{abs}_{FA}$ is larger than $\kappa^{abs}_{RA}$ and vice versa on the right side. On the other hand the condition numbers in Fig. 4.3b for isosurface extraction show a different behavior. On the left side of the plot $\kappa^{abs}_{FA^{-1}(\vartheta)}$ is smaller than $\kappa^{abs}_{FA^{-1}(\vartheta)}$ and vice versa on the right side. This corresponds to right side of the graph for $FA$ in Fig. 4.3a which is closer to a plateau than that of $RA$, i.e. the isosurface extraction is more ill-conditioned.

If we compare Fig. 4.3b with Fig. 4.3d we see that where $\kappa^{abs}_{FA} < \kappa^{abs}_{RA}$ holds also $\kappa^{abs}_{FA^{-1}(\vartheta)} > \kappa^{abs}_{RA^{-1}(\vartheta)}$ holds, and vice versa. Indeed the combined condition numbers are equal:

$$\kappa^{abs}_{FA,FA^{-1}(\vartheta)} = \kappa^{abs}_{RA,RA^{-1}(\vartheta)}. \tag{4.22}$$

This means that – in a first order approximation – the *propagation of uncertainty for isosurface extraction* in FA and RA fields is *equal*. This equality is indicated in Fig. 4.3e.




(a) $FA$ (b) $\kappa_{FA}^{abs}$ (c) $SNR_{FA}$

(d) $RA$ (e) $\kappa_{RA}^{abs}$ (f) $SNR_{RA}$

**Figure 4.5:** *Textured slices in the spiral dataset showing from left to right: anisotropy measure, absolute condition and SNR (assuming constant* $\widehat{SNR}_\lambda$*=20 for all eigenvalues). First row: FA, second row RA.*

**Examples.** We employ a synthetic spiral DTI dataset generated as described by Bergmann et al. [BLS05] that is visualized using ellipsoid glyphs in Fig. 4.4 with $FA$ mapped to glyph color. The $FA$, $RA$, $\kappa_{FA}^{abs}$, $\kappa_{RA}^{abs}$ and $SNR$ are shown in Fig. 4.5. Note that the $FA$ values are higher than $RA$ as well as the $\kappa_{FA}^{abs}$ values are higer than $\kappa_{RA}^{abs}$, i.e. absolute errors are amplified more for $FA$ than for $RA$. Nevertheless the higher $FA$ values lead to a higher SNR relative to $RA$. We exemplarily assume a constant $\widehat{SNR} = 20$ for all eigenvalues.

To apply our methods to real world data we consider a brain dataset consisting of 148 x 190 x 160 DTs. The eigenvalue fields are smoothed using a

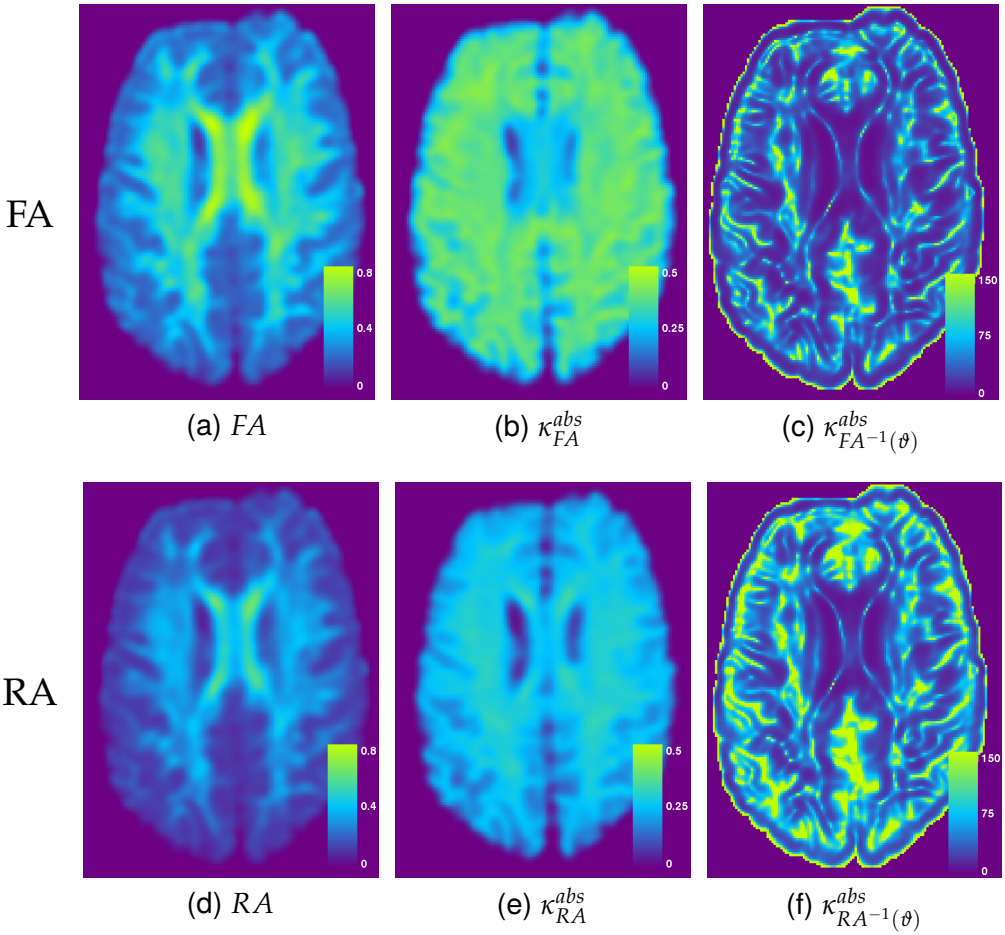



**Figure 4.6:** *Textured slices in the brain dataset showing from left to right: anisotropy measures, condition numbers for the anisotropy measures and condition numbers for isosurface extraction. First row: FA, second row RA.*

3D Gaussian kernel with a standard deviation of 1.2 voxel widths to estimate the mean values from the noisy data. The $FA$ and $RA$ as well as $\kappa^{abs}_{FA}$, $\kappa^{abs}_{FA^{-1}(\vartheta)}$, $\kappa^{abs}_{RA}$ and $\kappa^{abs}_{RA^{-1}(\vartheta)}$ are shown in Fig. 4.6.

In Appendix C we show the equality in Eq. (4.22) for the 2D case where analytical treatment is still readily comprehensible. The condition numbers of the $AI$ computations and the isosurface extraction problem from phantom and brain DTI data are shown in Sect. 5.5. There, the equality given in Eq. (4.22) can also be visually verified for the uncertain isosurfaces. For the 3D datasets the numerical gradient estimation introduces additional errors, but the differences between the combined condition numbers for FA and RA are below 1%.

### 4.3.5 Discussion

Our outcomes reproduce the previous result [PXH*99,HAN04,CKPB07] that $FA$ yields higher $SNR$ than $RA$, see Fig 4.5. This finding has led to the conclusion that $FA$ is more immune to noise and uncertainty. This statement is true, but applicable only if the intended final result is an immediate depiction of diffusion anisotropy.

However, in some applications isosurface extraction or thresholding are subsequent steps in visual analysis of DTI anisotropy. If this step is included too, the uncertainty propagation from the eigenvalues to the spatial position of the isosurface has also to be considered. For sensitivity analysis of isosurface extraction not only the scalar field but also the gradient magnitude has to be taken into account. Our results show that in this context there is no clear superiority of $FA$ compared to $RA$.

Our results also have implications for fiber tracking. Many fiber tracking algorithms use thresholds of $FA$ to restrict the resulting tracks to anisotropic areas of the brain. The uncertainty of the shape of these areas leads to uncertainties in the resulting fiber tracks. This is related to the sensitivity of fiber tracking results to variations of the anisotropy threshold that was investigated by Brecheisen et al. [BVPtHR09].

# 5

# Isocontours of Random Fields in Continuous Domains

A standard technique for visualizing crisp scalar fields, i.e. fields that are not afflicted with uncertainty, is to depict level sets, which under certain regularity conditions are $(N-1)$–dimensional isocontours.

We are interested in equivalents to isocontours in uncertain data. Obviously the position and shape of an isocontour is not precisely defined in this case. It is our aim to quantify the positional uncertainty of isocontours with respect to a given isovalue from a random field defined according to the models given in Chap. 3. This chapter is based on the publications [PH11, PH12].

## 5.1 Isolines and Isosurfaces

We recall a few basic definitions and facts about crisp scalar fields and isocontours, c.f. [Mil63], [PT88], [Mat02]. Assume that the data values are interpolated in $\mathcal{I}^*$ by a smooth function $y$ ($C^1$ or higher). A point $\mathbf{p}$ in $\mathcal{I}^*$ is called a *critical point* of $y$ if $\nabla y_{\mathbf{p}} = 0$. Other points of $\mathcal{I}^*$ are called *regular points* of $y$. Given a real number $\vartheta$ we call $y^{-1}(\vartheta)$ the *$\vartheta$-level* of $y$, and we say it is a *critical level* (and that $\vartheta$ is a *critical value* of $y$) if it contains at least one critical point of $y$. Other real numbers $\vartheta$ are called *regular values* of $y$ and the corresponding levels $y^{-1}(\vartheta)$ are called *regular levels*. From the inverse function theorem it follows that for a regular value $\vartheta$, $y^{-1}(\vartheta)$ is a smooth, codimension one submanifold of $\mathcal{I}^*$, which we then call *isocontour* (if it is non-empty). For a critical value $\vartheta$ the corresponding critical level $y^{-1}(\vartheta)$ is not manifold. At a saddle point with value $\vartheta$, connected components of the critical level touch. Maxima and minima with value $\vartheta$ are isolated points in the $\vartheta$-level.

A critical point $\mathbf{p}$ is called *non-degenerate* if the Hessian $H_{\mathbf{p}}\, y$ is non-singular, i.e. $det\, H_{\mathbf{p}}\, y \neq 0$. From the Morse Lemma it follows that non-degenerate critical points are *isolated*. If *all* critical points of function $y$ are non-degenerate (and thus are isolated), $y$ is called *Morse function*. If a function contains *degenerate* critical points, the level set $y^{-1}(\vartheta)$ can be 0- to $N$-dimensional, since points with $\nabla y_{\mathbf{p}} = 0$ can form arbitrary regions in the domain. We will refer to regions of degenerate points as *plateaus*.

### 5.1.1 Computational Problems of Isocontour Extraction

The foregoing might look like mathematical sophistry, but it is algorithmically relevant. Depending on the interpolation and reconstruction method, the resulting isosurface can vary. In case trilinear interpolation is used, marching cubes type algorithms have to deal with ambiguities. An explanation of these cases and literature on dealing with these ambiguities is given, for example, in [NY06]. Computing level sets for non-Morse functions at a value $\vartheta$ containing plateaus is even more complicated. Marching cubes type algorithms that assume $y^{-1}(\vartheta)$ to be a $N-1$-dimensional surface, inevitably fail. An extension of the marching cubes algorithm dealing with degenerate critical levels has been suggested by Weber et al. [WSH03]. Even if the level set contains no critical points, problems occur if it is close to a plateau: the condition number $\|\nabla y(\mathbf{x})\|^{-1}$ is then large, i.e. the computation of the isocontour is ill-conditioned, *independently* of the algorithm, see Chap. 4. The computed results therefore are not reliable and the visual impression of the computed contour can be misleading.

### 5.1.2 The Probabilistic Ansatz

Due to the fact that scientific data is affected by uncertainty and also due to the computational problems described above we aim for a probabilistic formulation. We are interested in quantities that describe *how likely* it is that an isocontour exists at each location of a domain, given a scalar random field. Different approaches are presented in this chapter and Sect. 6.2.

The probabilistic procedure does not have to deal with degenerate or ambiguous cases separately as it is the case for marching cubes and related algorithms. Probabilities for the occurrence of level crossings for critical isovalues or non-Morse functions (with respect to the mean values, for example) are computed correctly without treating any special cases. Fig. 6.4 illustrates the difference between the deterministic and the probabilistic approach by comparing a crisp isoline with results computed using methods introduced in Sect. 6.2.

## 5.2 Continuous Extension of Discrete Fields

Many visualization techniques and feature extraction methods require functions as input that are defined in a continuous domain. In order to apply standard visualization techniques one needs to specify for uncertain data what happens in regions between sample points. One possibility is to extend the parameter-discrete random field $\{Y_{\mathbf{x}_j} : \mathbf{x}_j \in \mathcal{I}\}$ to a *parameter-continuous* random field $\{Y_{\mathbf{x}} : \mathbf{x} \in \mathbb{R}^N\}$ with properties like, e.g., covariances between arbitrary locations that smoothly interpolate the covariances between discrete locations.

### 5.2.1 Level Crossings in Continuous Random Fields

The analysis of level sets in parameter-continuous random fields is an area of active research in mathematics, see, e.g., [AT07]. This research is triggered by applications in natural sciences, in which randomness is required to describe certain phenomena, see, e.g., [ATW09]. Depending on its covariance, a random field might be rather non-smooth. Its smoothness is directly determined by the differentiability of the covariance $\bar{C}$ at distance 0.

Given a random field and a level set $\Omega \subset \mathbb{R}^N$ that corresponds to some threshold $\vartheta$, the following quantities are of interest: for random fields from $\mathbb{R}^N$ to $\mathbb{R}^N$, the number of level crossings $v_B$ in some subset $B \subset \mathbb{R}^N$, and for random fields from $\mathbb{R}^N$ to $\mathbb{R}^{N'}$ with $N > N'$, the geometric measure (length, area, volume, ...) $A_B$ of the intersection $\Omega \cap B$. Then $v_B$ and $A_B$ itself are random variables, but currently there is no known way to compute their distribution for non-trivial situations. A tool to understand the distributions are Rice formulae; they allow to express the expectation values and higher moments of $v_B$ and $A_B$ as integrals over a function that depends on the joint distribution of the random field and its derivative [AW09].

For Gaussian random fields it can be shown (i) that the level sets are continuous, but in general are non-differentiable, and (ii) that their Hausdorff dimension is larger than $N - 1$ [Adl81]. Obviously, these objects do not represent what we are interested in: Instead of approximating level sets of the underlying field, level sets are considered whose properties are largely determined by covariances of measurement errors. A more adequate mathematical model is described below.

### 5.2.2 Interpolation of PDFs

As described earlier, we consider as input uncertain data represented by a parameter-discrete random field $\{Y_{\mathbf{x}i} : \mathbf{x}_i \in \mathcal{I}\}$ with parameter set $\mathcal{I} \subset \mathbb{R}^N$, state space $S = \mathbb{R}$, and random variables $Y_{\mathbf{x}i}$ that can be described by PDFs $f_i(y)$. In the following we additionally assume that the probability

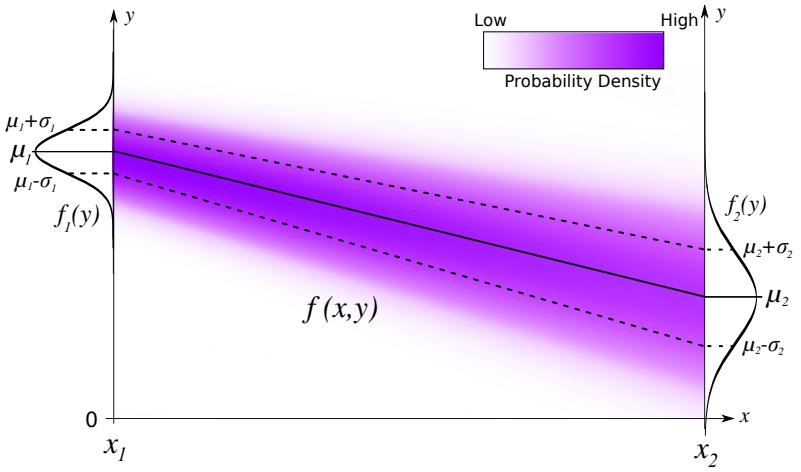



**Figure 5.1:** *Linear interpolation between two normal distributions given at two sample points $x_1$ and $x_2$.*

distributions for all $Y_{\mathbf{x}i}$ are of the same type, e.g., are *all* uniform or are *all* normal. We further assume that the PDFs can be functionally represented by their expected values $\mu_i = \mathrm{E}(Y_{\mathbf{x}i})$ and a finite number of $\hat{m} - 1$ central moments $\mathrm{E}((Y_{\mathbf{x}i} - \mu_i)^m)$, where $m \in \{2, \ldots, \hat{m}\}$.

To build a continuous extension we consider a continuous parameter set $\mathcal{I}^* \subseteq \mathbb{R}^N$ that contains all sample points $\mathbf{x}_i$. Given that the PDFs at the sample points are all of the same type, the most natural assumption is that the PDFs between the sample points, and thus in all $\mathbf{x}$ of $\mathcal{I}^*$, are also of that type. Therefore we extend the discrete model defined at the sample points $\mathbf{x}_i$, $i \in \{1, 2, ..., n\}$ to a continuous model in the whole domain $\mathcal{I}^*$ by (i) *interpolating* the expected values $\mu_i$ and the $m$-th roots of the central moments $\varsigma_{m,i} = \mathrm{E}((X - \mu)^m)^{\frac{1}{m}}$ in parameter space and (ii) inserting these interpolated values in the PDFs. This is a rather general method; it can be applied to all kinds of PDFs that can be parametrized by their moments. (Note that simple blending of PDFs between grid points would in general not preserve the type of the distribution; for instance blending two normal distributions would yield a bimodal distribution.)

For the functions $\mu(\mathbf{x})$ and $\varsigma_m(\mathbf{x})$ with $\mathbf{x} \in \mathcal{I}^*$ all sorts of interpolants can be used. For simplicity we use linear tensor product interpolation in the following.

As an example, assume that the random variables $Y_{\mathbf{x}i}$ are normally distributed $f_i(y) = \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{y-\mu_i}{\sigma_i}\right)^2\right)$ with $y \in S$. The interpolated PDF is then obtained by substituting the expected values $\mu_i$ and standard devia-

tions $\sigma_i$ by interpolants $\mu(\mathbf{x})$ and $\sigma(\mathbf{x}) \equiv \varsigma_2(\mathbf{x})$:

$$f(\mathbf{x}, y) = \frac{1}{\sigma(\mathbf{x})\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{y - \mu(\mathbf{x})}{\sigma(\mathbf{x})}\right)^2\right). \tag{5.1}$$

Fig. 5.1 shows an example in one dimension with linear interpolation between two sample points $x_1$ and $x_2$.

## 5.3 Local Measures for the Positional Uncertainty of Isocontours

In this section the aim is to quantify the positional uncertainty of isocontours based on the continuous extension of the discretely sampled input data. Because of the practical relevance of the Gaussian distribution all examples and explicit formulae in the following are provided for this case. However, the formulae are valid for arbitrary probability distributions, unless stated differently, and in many cases also explicit formulae can be easily derived.

### 5.3.1 Isocontour Density

We consider isocontours in $\mathcal{I}^* \subset \mathbb{R}^N$. Using an interpolated PDF $f(\mathbf{x}, y)$ we compute a spatial density of an isocontour by simply evaluating this function at a given *isovalue* $\vartheta \in \mathbb{R}$ for all points $\mathbf{x} \in \mathcal{I}^*$. For that we introduce a function

$$g_\vartheta(\mathbf{x}) := f(\mathbf{x}, y = \vartheta), \tag{5.2}$$

which we call *isocontour density* (ICD). It provides the probability density with respect to $\vartheta$ at position $\mathbf{x}$ which is a measure for the spatial distribution of the uncertain isocontour. The quantity $g_\vartheta(\mathbf{x})\,\mathrm{d}y$ is the probability that the true but unknown function takes a value in the interval $[\vartheta, \vartheta + \mathrm{d}y]$ at position $\mathbf{x}$. Note that $f$ is a normalized PDF for a specific $\mathbf{x}$ with respect to the *state space* $S$ (i.e. in $y$-direction). Thus, the values of $g_\vartheta(\mathbf{x})$ are probability densities with respect to $S$, not $\mathcal{I}^*$. For normally distributed data and given interpolants $\mu(\mathbf{x})$ and $\sigma(\mathbf{x})$ the ICD follows directly from Eq. (5.1):

$$g_\vartheta(\mathbf{x}) = \frac{1}{\sigma(\mathbf{x})\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\vartheta - \mu(\mathbf{x})}{\sigma(\mathbf{x})}\right)^2\right). \tag{5.3}$$

An example graph of $g_\vartheta(\mathbf{x})$ for the 1D case between two sample points $x_1$, $x_2$ is given in Fig. 5.2.

Due to the continuous interpolation the values of $g_\vartheta(\mathbf{x})$ at the grid points only depend on the probability distributions that are given there: $g_\vartheta(\mathbf{x}_i) =$

$f_i(\vartheta)$. That means $g_\vartheta(\mathbf{x})$ is a *continuous* function. As $\mu(\mathbf{x})$ and $\sigma(\mathbf{x})$ are piecewise linear functions, $g_\vartheta(\mathbf{x})$ is non-differentiable at the grid boundaries and differentiable within the grid cells.

For the limit $\sigma(\mathbf{x}) \to 0$ we expect $g_\vartheta(\mathbf{x})$ to return a *crisp* isocontour. If we consider $\lim_{\sigma(\mathbf{x})\to 0} g_\vartheta(\mathbf{x})$ and use the following definition of the Dirac delta distribution for $x \in \mathbb{R}$ [Kan98]

$$\delta(x) = \lim_{m\to 0} \frac{1}{m\sqrt{\pi}} \exp\left(-\left(\frac{x}{m}\right)^2\right), \tag{5.4}$$

with support $\{0\}$, then

$$\lim_{\sigma(\mathbf{x})\to 0} g_\vartheta(\mathbf{x}) = \delta(\vartheta - \mu(\mathbf{x})). \tag{5.5}$$

So, if $\sigma(\mathbf{x})$ vanishes (i.e. no uncertainty is considered) then the support of $g_\vartheta(\mathbf{x})$ is identical the set of $\vartheta$-level-crossings of $\mu(\mathbf{x})$, i.e. the level set

$$\Omega = \{\mathbf{x}_s \in \mathcal{I}^* : \mu(\mathbf{x}_s) = \vartheta\}. \tag{5.6}$$

For a symmetric distribution like the Gaussian distribution, at a fixed spatial position $x_c \in \mathcal{I}^* \subseteq \mathbb{R}$ (1D case) the interpolated PDF $f(x_c, y)$ takes its maximum at $y = \mu(x_c)$. From this it follows that, if $\sigma(x)$ is constant, $g_\vartheta(x)$ takes its *local maxima* at the points $x_s \in \Omega$.

The function $g_\vartheta(x)$ can be related to the condition numbers of isocontour extraction. Recall that Eq. (4.1) describes the relation of perturbations of the input data to the perturbation of the result. How does $g_\vartheta(x)$ propagate the uncertainty represented by the standard deviation of the input data?

Assuming that an interval $[x_j, x_k] \subset \mathcal{I}^*$ contains a level-crossing $\mu(x) = \vartheta$ and that $\sigma(x)$ is constant ($\sigma_j = \sigma_k$), the *inflection points* $x_a, x_b$ of $g_\vartheta(x)$ can be calculated using elementary calculus. If $[x_a, x_b] \subseteq [x_j, x_k]$, we can consider the distance between $x_a$ and $x_b$ to be a measure of the spread of $g_\vartheta(x)$. With regard to the fact that the inflection points of a normal distribution are located one standard deviation from the mean we define the spread of $g_\vartheta(x)$ as

$$s_{g_\vartheta} = \frac{1}{2}|x_a - x_b|. \tag{5.7}$$

Simple calculation shows that

$$\frac{1}{2}|x_a - x_b| = \sigma_{j,k}\frac{1}{|\mu_j - \mu_k|}. \tag{5.8}$$

As we assume linear interpolation, $|\mu_j - \mu_k|$ is the derivative of $\mu(x)$ with

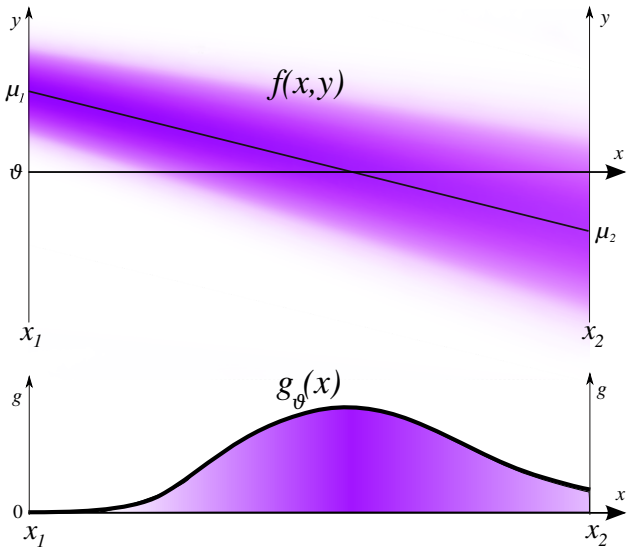



**Figure 5.2:** *Example graph for the ICD in a 1D grid cell. The values $g_\vartheta(x)$ are computed from the interpolated PDF $f(x,y)$ with respect to threshold $\vartheta$.*

respect to $x$ in the interval $[x_j, x_k]$. From this it follows that

$$s_{g_\vartheta} = \kappa_{abs}\, \sigma_{j,k}, \tag{5.9}$$

where $\kappa_{abs}$ is the condition number for the calculation of level-crossings of $\mu(x)$ (cf. Eq. (4.5)). This shows that the ICD propagates the uncertainty of the input data proportionally to the condition number.

### 5.3.2 Point-Wise Level-Crossing Probabilities

An alternative approach to calculate spatial distributions of isocontours is described in this section. Again, we consider a parameter-discrete random field $\{Y_{\mathbf{x}i} : \mathbf{x}_i \in I\}$ as a model for the uncertain input data (described in Sect. 3.3). For simplicity we consider a one dimensional parameter set first.

Let $\mathbf{x}_j, \mathbf{x}_k \in \mathcal{I}$ be adjacent sample points with associated random variables $Y_{\mathbf{x}j}, Y_{\mathbf{x}k}$ and PDFs $f_j(y)$, $f_k(y)$. If we assume monotonic interpolants between the realizations then we have at most one crossing of the constant level line $y = \vartheta$. The probability for a level-crossing along the line segment $[\mathbf{x}_j, \mathbf{x}_k]$ then is

$$\begin{aligned} \mathrm{P}_{[\mathbf{x}_j,\mathbf{x}_k]}(\vartheta \text{ crossed}) \;&=\; \mathrm{P}(Y_{\mathbf{x}j} \le \vartheta)\,\mathrm{P}(Y_{\mathbf{x}k} \ge \vartheta) \\ &\quad + \mathrm{P}(Y_{\mathbf{x}j} \ge \vartheta)\,\mathrm{P}(Y_{\mathbf{x}k} \le \vartheta). \end{aligned} \tag{5.10}$$

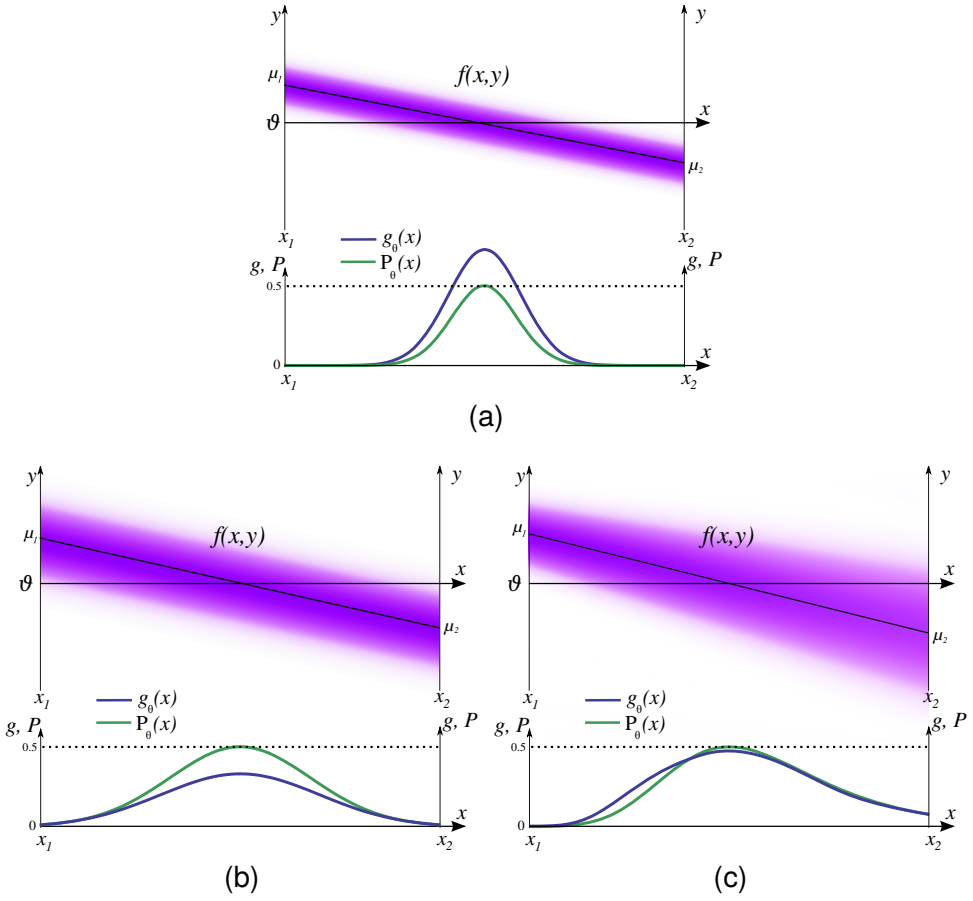



**Figure 5.3:** *Comparison between the ICD $g_\vartheta(x)$ and the LCP $P_\vartheta(x)$ for different Gaussian input distributions in 1D. The probability density $f(x,y)$ is indicated by a colormap.*

The probabilities on the right side of this equation can easily be calculated by the cumulative distribution functions $F_{(\cdot)}(\vartheta)$ that provide the probability that $Y_{(\cdot)}$ is less or equal to $\vartheta$. So, the level-crossing probability can be expressed as

$$\begin{aligned} \mathrm{P}_{[\mathbf{x}_j,\mathbf{x}_k]}(\vartheta \text{ crossed}) &= F_j(\vartheta)\,(1-F_k(\vartheta)) \\ &\quad +(1-F_j(\vartheta))\,F_k(\vartheta). \end{aligned} \tag{5.11}$$

With that it is possible to calculate a level-crossing probability between each two adjacent vertices of the grid. However, it is desirable to get a probability for each point in a continuous domain.

Using the interpolation scheme introduced in Sect. 5.2 it is possible to compute a CDF from interpolated PDF $f(\mathbf{x},y)$ and thus subdivide the grid

cells in order to compute probabilities for subsegments i.e.

$$\mathrm{P}_{[\mathbf{x},\mathbf{x}+\Delta\mathbf{x}]}(\vartheta \text{ crossed}).$$

We are then able to find the limit

$$\mathrm{P}_{\vartheta}(\mathbf{x}) = \lim_{\Delta\mathbf{x}\to 0} \mathrm{P}_{[\mathbf{x},\mathbf{x}+\Delta\mathbf{x}]}(\vartheta \text{ crossed}) \tag{5.12}$$

which evaluates to *level-crossing probability* (LCP)

$$\mathrm{P}_{\vartheta}(\mathbf{x}) = 2\,F_{\mathbf{x}}(\vartheta)\,(1 - F_{\mathbf{x}}(\vartheta)) \tag{5.13}$$

where $F_{\mathbf{x}}(\vartheta) = \int_{-\infty}^{\vartheta} f(\mathbf{x}, y)\mathrm{d}y$ is the CDF of the interpolated PDF. Equation (5.13) gives the probability that for two independent realizations $y_a$ and $y_b$ of a random variable distributed according to the interpolated PDF $f(\mathbf{x}, y)$ one of them is greater or equal to $\vartheta$ while the other is less or equal to $\vartheta$.

To show the range of function values we consider $P_{\vartheta}$ as a function of $F_{\mathbf{x}}$ in the interval $[0, 1]$. Obviously $P_{\vartheta}$ is not negative. Because $P_{\vartheta} = 2(F_{\mathbf{x}} - F_{\mathbf{x}}^2)$ we can write

$$\frac{\mathrm{d}P_{\vartheta}}{\mathrm{d}F_{\mathbf{x}}} = 2(1 - 2F_{\mathbf{x}}).$$

From $\mathrm{d}P_{\vartheta}/\mathrm{d}F_{\mathbf{x}} = 0$ and $F_{\mathbf{x}} \in [0, 1]$ it follows, that the function $P_{\vartheta}(\mathbf{x})$ takes a maximum at $F_{\mathbf{x}}^{(max)} = \frac{1}{2}$. Thus, the maximum value of $P_{\vartheta}$ is $P_{\vartheta}(\frac{1}{2}) = \frac{1}{2}$ and therefore $P_{\vartheta}(\mathbf{x}) \in [0, \frac{1}{2}]$.

If $P_{\vartheta}(\mathbf{x})$ is maximal then

$$F_{\mathbf{x}}^{(max)}(\vartheta) = \frac{1}{2} = 1 - F_{\mathbf{x}}^{(max)}(\vartheta)$$

holds. This means that $prob(Y_{\mathbf{x}} \leq y_{\vartheta}) = prob(Y_{\mathbf{x}} \geq y_{\vartheta}) = \frac{1}{2} = prob(Y_{\mathbf{x}} \leq y_{\frac{1}{2}}) = prob(Y_{\mathbf{x}} \geq y_{\frac{1}{2}})$ if $y_{\frac{1}{2}}$ is the *median* of the PDF $f_{\mathbf{x}}(y)$. Thus, $P_{\vartheta}(\mathbf{x})$ takes its maximum if the isovalue $\vartheta$ is equal to the *median* of the interpolated PDF $f(\mathbf{x}, y)$.

For normally distributed input data the LCP evaluates to

$$\mathrm{P}_{\vartheta}(\mathbf{x}) \quad = \quad \frac{1}{2}\left(1 - \mathrm{Erf}\left(\frac{\mu(\mathbf{x}) - \vartheta}{\sqrt{2}\,\sigma(\mathbf{x})}\right)\right), \tag{5.14}$$

where Erf is the *error function*.

### 5.3.3 Comparison

While the ICD maps points $\mathbf{x} \in \mathcal{I}^*$ to probability densities, LCP maps points to probabilities. The cardinality of the range of $g_{\vartheta}(\mathbf{x})$ depends on $f(\mathbf{x}, y)$. The

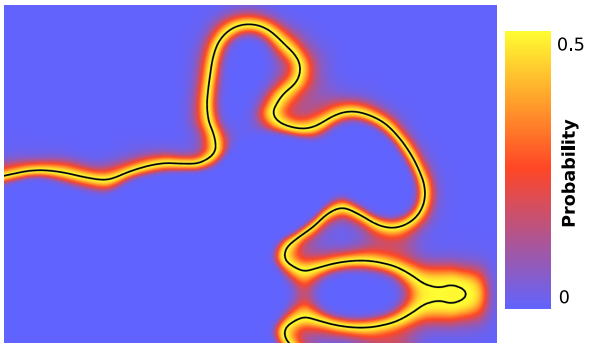



**Figure 5.4:** *The probability $P_\vartheta(\mathbf{x})$ (LCP) for an uncertain isocontour in a slice of the Fuel dataset is displayed using a colormap in combination with the crisp isoline $\mu^{-1}(\vartheta)$ of the mean field (black).*

density values $g_\vartheta(\mathbf{x})$ are relatively low in case of large values of $\sigma(\mathbf{x})$ and high in case of small values of $\sigma(\mathbf{x})$ because $f$ is normalized with respect to the state space. As we have shown in Sect. 5.3.2 the range of $P_\vartheta(\mathbf{x})$ is always $[0, \frac{1}{2}]$ and does not depend on the spread of the input distributions. The positions of the maxima of $g_\vartheta(\mathbf{x})$ and $P_\vartheta(\mathbf{x})$ are not necessarily identical. Comparisons for three different pairs of input distributions are shown in Fig. 5.3.

The ICD can be computed directly for data with analytically defined PDFs. For LCP we need CDFs which for the major practical cases are also available, either as closed analytical expressions or as fast numerical approximations. For instance for normally distributed input data the Erf function has to be computed; an efficient and accurate approximation was proposed by Winitzki [Win08]. If such approximations are used, both approaches are suitable for interactive real-time visualization of uncertain fields.

We will use the LCP exclusively in the remainder of this chapter, because a probability field is more easily interpretable than probability densities with respect to the state space. The fact that the range of $P_\vartheta(\mathbf{x})$ does not depend on the input data also simplifies the visual mapping.

## 5.4 Visualization Methods

The definitions in the previous sections are used as a basis for the design of interactive visualization methods that depict uncertain isolines and isosurfaces, and do not only give an impression of the uncertainty, but also depict quantitative measures.

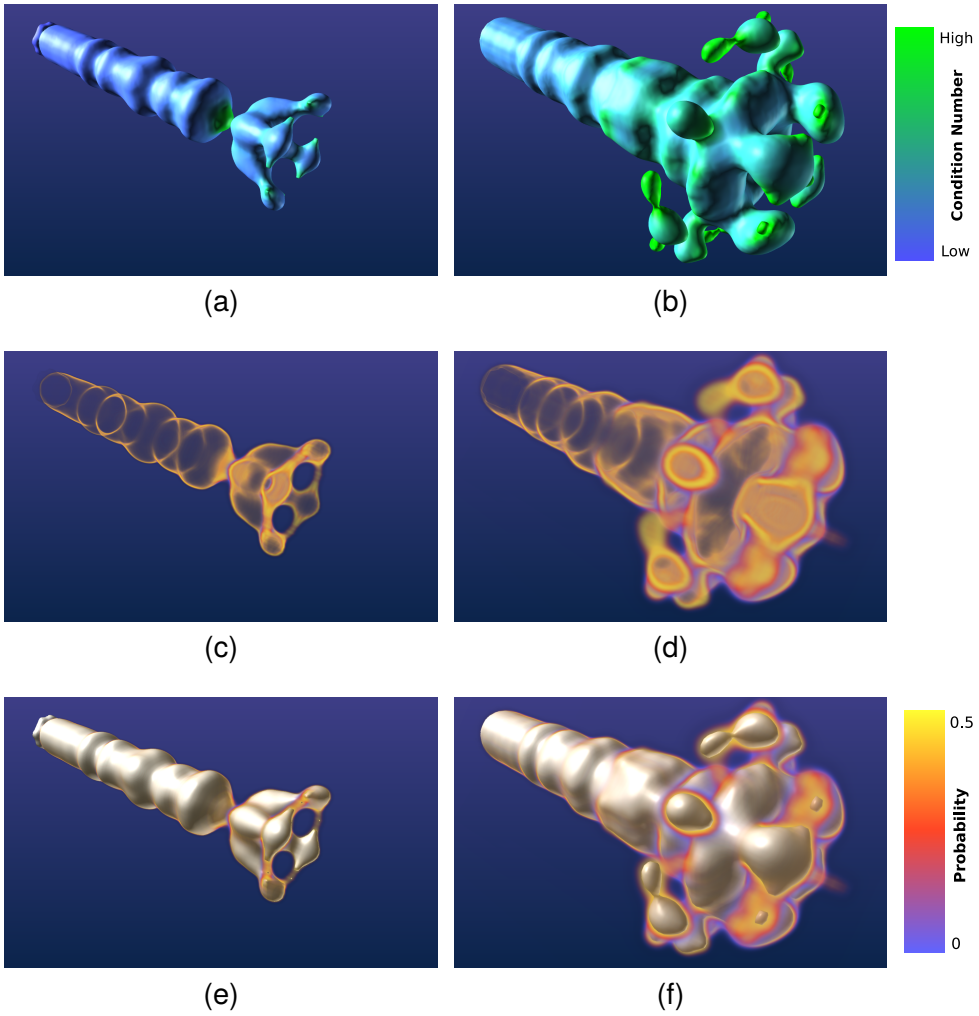



**Figure 5.5:** *Isosurfaces of the fuel dataset with a fixed (artificial) standard deviation $\sigma = 5$. The condition number for $\mu^{-1}(\vartheta)$ is color-mapped in the top row. The LCP is depicted by volume rendering in the middle row. The bottom row displays combined renderings of the LCP and the mean surface. The left column of images use isovalue $\vartheta = 90$ and reveal low positional uncertainty; the right column ($\vartheta = 22$) uncovers higher positional uncertainty due to higher condition numbers.*

The pure display of the distribution in 3D data can be difficult to interpret (cf. Fig. 5.5). In order to improve 3D impression, we can depict both the crisp isocontours $\mu^{-1}(\vartheta)$ of the mean field and the spatial distribution via the LCP.

To uncover the *reason* for the spatial distribution of an uncertain isocontour, both condition number $\kappa_{abs}$ and the standard deviation $\sigma(\mathbf{x})$ of the input distributions can be depicted side by side in two images. The con-

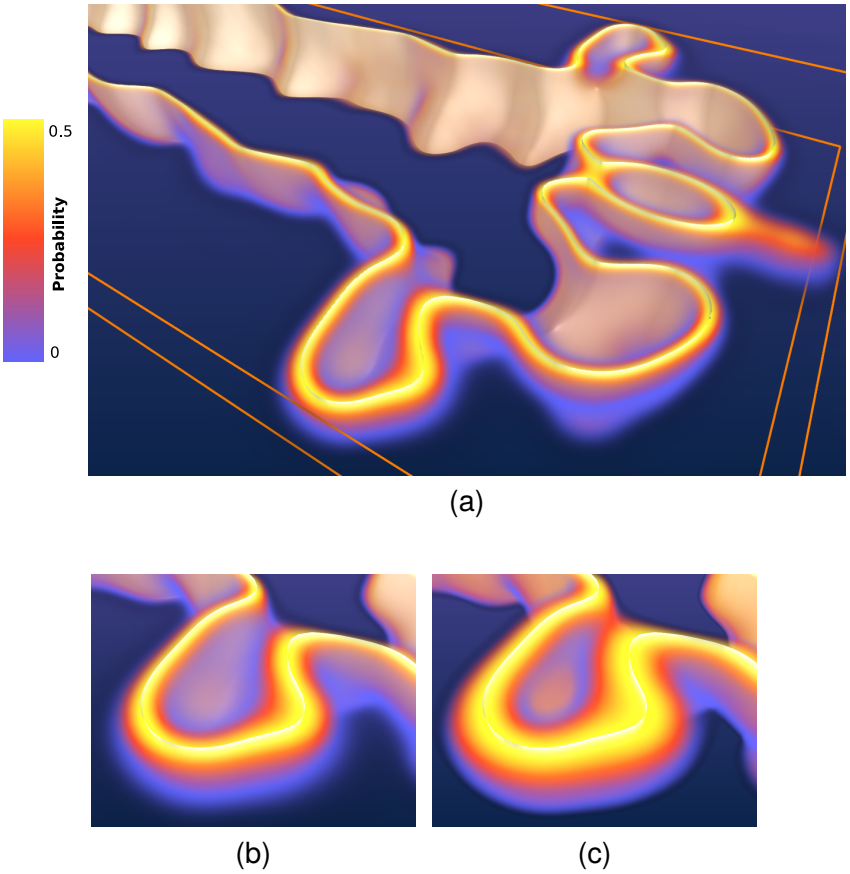



**Figure 5.6:** *Uncertain isosurface bounded by two clipping planes to display $P_\vartheta(\mathbf{x})$ inside the mean surface (a). Normal distributions are considered as input data in (a) and the closeup (b). When considering uniform distributions we obtain the spatial distribution depicted in (c).*

dition number on crisp isosurfaces is shown by a colormap. In addition to that, the values of the standard deviation $\sigma(\mathbf{x})$ can be indicated by the length of line glyphs, see Fig. 5.9d.

For two-dimensional input data, the LCP can easily be displayed as a 2D distribution using a colormap. For each texel the LCP has to be evaluated with respect to a given isovalue $\vartheta$. The visualization of uncertain isosurfaces in real time is realized using GPU-based ray casting [KW03]. Instead of employing user-defined transfer functions, the LCP is evaluated and the values are mapped to *transparency* and *color* which then are used in a emission-absorption model for ray integration. For combined volume and surface rendering the intersection of the ray and $\mu^{-1}(\vartheta)$ is tested in each step of the ray integration.

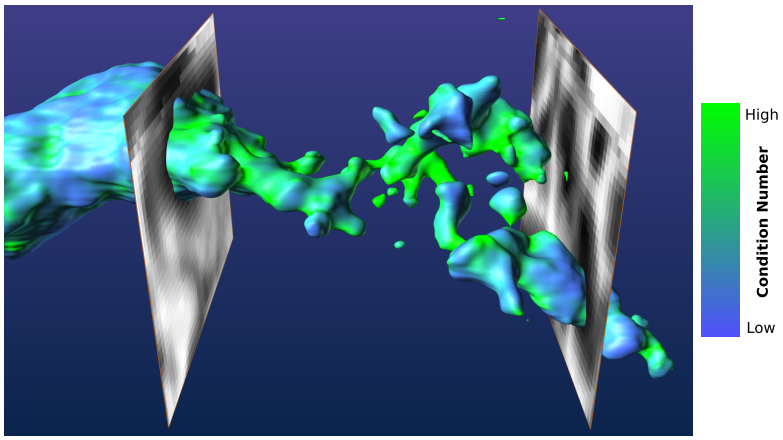
## 5.5 Results

In the following we apply the visualization methods to various datasets. Using a non-optimized implementation we achieve frame rates between 5 and 25 fps on an Intel Xeon X5550 2.66 GHz system with a GeForce GTX 285 GPU.

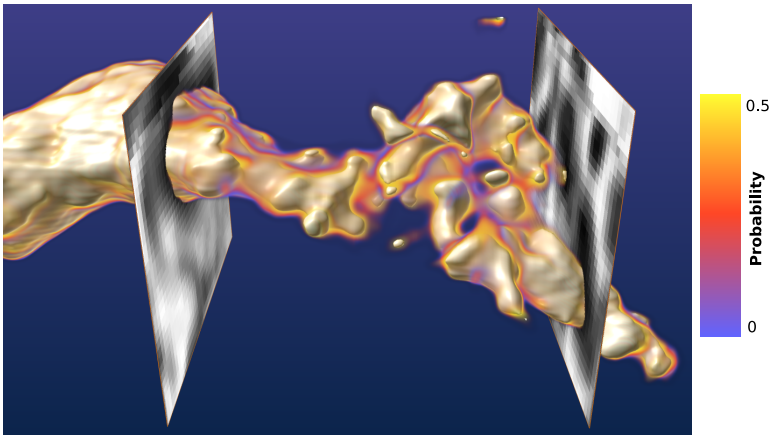
**Fuel Dataset.** As an example, an uncertain isoline in a slice of the fuel dataset is depicted in Fig. 5.4. We assume normal distributions as input data and artificially set the standard deviation to be constant $\sigma = 5$ (about 2% of the range of values in the dataset) because no information about the uncertainty was available. The LCP is displayed using a colormap and in combination with the crisp isoline $\mu^{-1}(\vartheta)$. In Fig. 5.5 isosurfaces in the same dataset are shown. The condition number mapped to surface color (top row), a volume rendering of the LCP (middle row) and a combined rendering of the LCP and the mean surface are displayed for isovalues $\vartheta = 90$ and $\vartheta = 22$. The combined rendering (bottom row) improves 3D impression of the visualization compared to the volume rendering alone. The uncertain isosurfaces in the second row reveal higher position uncertainty compared to the first row due to higher condition numbers in the respective areas. The fact that position uncertainty inside a closed mean surface is occluded when surface and volume rendering are combined can be met by the user by placing clipping planes to make the interior of the mean surface visible, cf. Fig. 5.6 (a). A close-up is shown in Fig. 5.6 (b). For comparison, the LCP considering uniform instead of normal distributions as input is displayed in Fig. 5.6 (c).

**Medical Volume Data.** Medical volume data from CT scanners usually does not contain explicit uncertainty information, but the amount of noise in the scans can be used as an estimate. An approach that was also used by Firbank et al. [FCHW99] to compute the signal to noise ratio is based on analysis of homogenous subsets of images (i.e. areas with constant signal). In case no area with constant signal is available in a dataset alternative methods such as single image SNR estimation [TSP01] should be applied.

As an example we considered a dataset used for planning an implant in the middle ear of a patient. Here the size and number of connected air pockets is important. We estimated the standard deviation of the noise from areas that contain air only. The noise was approximately normally distributed. We denoised the CT scan using a median filter and considered this as $\mu(\mathbf{x})$. The estimated standard deviation ($\sigma \approx 13.4$) is used to display an uncertain isosurface that depicts air pockets in the middle ear (shown in Fig. 5.7). This shows that the topology of the isosurface is not clearly




(a)




(b)

**Figure 5.7:** *Isosurface with condition numbers mapped to color (a) and an uncertain isosurface in a CT scan of the middle ear region (b).*

defined. The mean isosurface has several distinct parts that are connected by relatively large LCP values. Thus, the number, size and shape of air pockets are highly uncertain. This also implies that a segmentation of these areas by thresholding can lead to erroneous results.

**Simulated Climate Data.** Uncertainty in climate simulations is often represented by *ensembles* which contain multiple results for the simulated quantities. As an example we use daily average hindcast data from the DEMETER project [Pal04] where the results of 7 different climate models and 9 different sets of simulation parameters each constitute ensembles with 63 members.

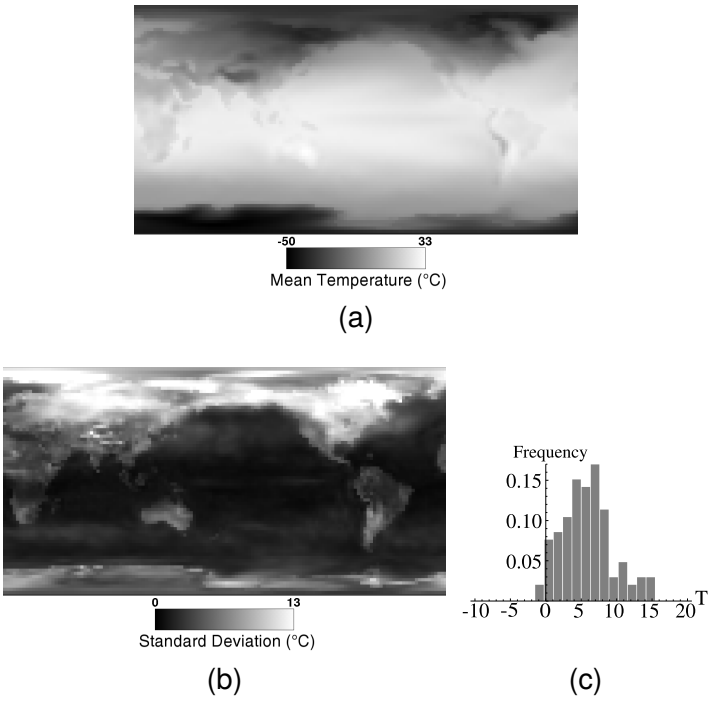



**Figure 5.8:** *Simulated climate data: The ensemble mean $\mu(\mathbf{x})$ for the temperature field at 2 m altitude is shown in (a), the standard deviation $\sigma(\mathbf{x})$ in (b), a histogram for the ensemble values at a single location in (c).*

This way, the simulation accounts for both model uncertainty and uncertainty of the input data.

To prepare for the extraction of uncertain isocontours, the data in a *temperature field* ensemble for February 20th, 2000 is analyzed statistically. For each location $\mathbf{x}$ we compute the ensemble mean $\mu(\mathbf{x})$ and standard deviation $\sigma(\mathbf{x})$ (see Fig. 5.8a–5.8b). Again, we model the uncertainty by normal distributions. In Fig. 5.8c the histogram of temperatures at a single location, but considering all ensemble members, is shown.

From these fields we are able to extract uncertain isolines (isotherms) using the LCP. Fig. 5.9a–5.9c depict uncertain isolines for -25°C, 0°C and 25°C, respectively, which reveal highly varying position uncertainty around the crisp ensemble mean isoline (black).

From DEMETER not only temperatures at the two-meter level but also temperatures at the pressure levels 850 hPa, 500 hPa and 200 hPa above the earth's surface are available. We analyze the ensembles for these levels in the same manner as the two-meter temperature field and use all results to construct a volume dataset with pressure mapped to the third coordinate. This volume is used for the extraction of uncertain isotherm surfaces. In

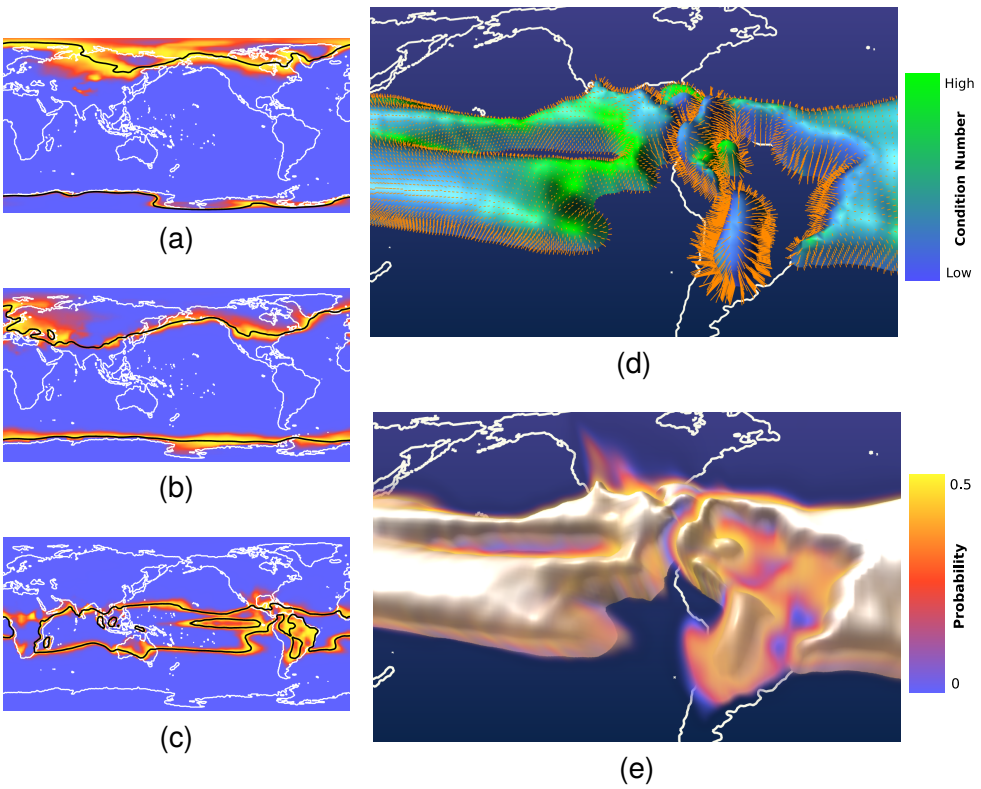



**Figure 5.9:** *Uncertain isotherms lines for -25°C, 0°C and 25°C are shown in (a)-(c), respectively. The crisp mean isoline is drawn in black. On the mean isotherm surface ($\vartheta = 25°C$) the condition numbers are mapped to color and the values of $\sigma(\mathbf{x})$ are depicted by the length of line glyphs in (d). The resulting uncertain isosurface is displayed in (e).*

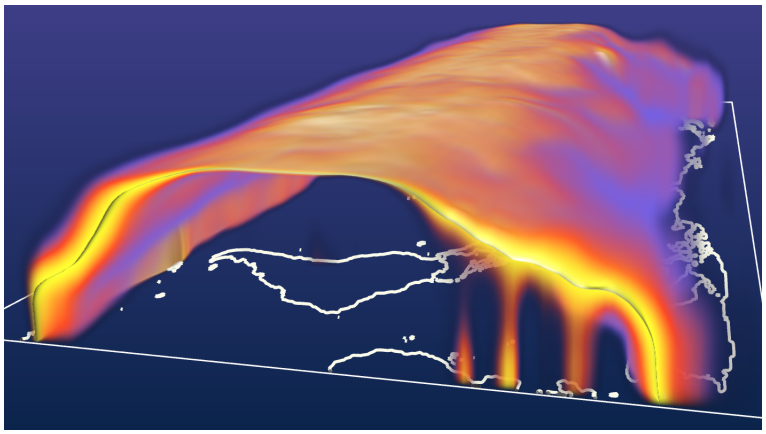


**Figure 5.10:** *Uncertain isotherm surface for 0°C. Note the ridges extending downwards to the earth's surface that are not indicated by the crisp isosurface.*

Fig. 5.9d the condition numbers for the crisp mean surface is displayed by a colormap for isovalue $\vartheta = 25°\text{C}$ and the values of $\sigma(\mathbf{x})$ are indicated by the length of line glyphs at that surface. The corresponding uncertain isotherm surface is shown in Fig. 5.9e. Comparing Fig. 5.9d to Fig. 5.9e we see that low values of $\sigma(\mathbf{x})$ with high values of $\kappa_{abs}$ and high values of $\sigma(\mathbf{x})$ with low values of $\kappa_{abs}$ both can result in similar amount of position uncertainty. Fig. 5.10 shows an uncertain isotherm surface for 0°C. This example shows that volume rendering of the LCP reveals also structural information that crisp isosurfaces do not display: possible topological changes of crisp isosurfaces are indicated as well as ridges of $P_\vartheta$.

**DTI Data.** We computed the condition numbers and estimate the propagation of uncertainty to the anisotropy indices and the related isosurfaces for a synthetic spiral and a brain DTI dataset.

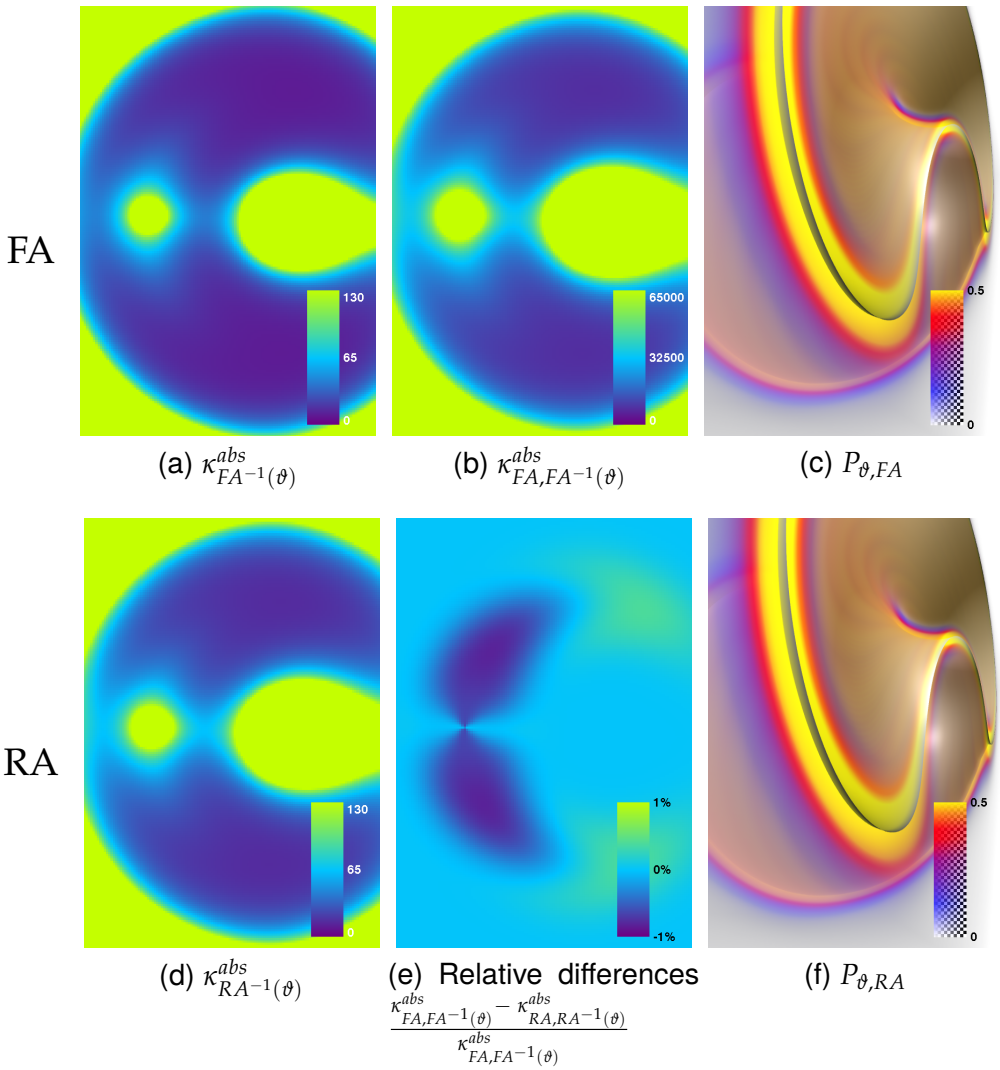
In Fig. 5.11 the condition numbers $\kappa^{abs}_{FA^{-1}(\vartheta)}$, $\kappa^{abs}_{RA^{-1}(\vartheta)}$, $\kappa^{abs}_{FA,FA^{-1}(\vartheta)}$ and the relative differences between $\kappa^{abs}_{FA,FA^{-1}(\vartheta)}$ and $\kappa^{abs}_{RA,RA^{-1}(\vartheta)}$ are shown along with two corresponding uncertain isosurfaces. The values of $\kappa^{abs}_{FA^{-1}(\vartheta)}$ are lower than $\kappa^{abs}_{RA^{-1}(\vartheta)}$, while the relative differences between $\kappa^{abs}_{FA,FA^{-1}(\vartheta)}$ and $\kappa^{abs}_{RA,RA^{-1}(\vartheta)}$ are smaller than 1%. The uncertain isosurfaces in Fig. 5.11(c) and (f) are depicted by volume renderings of $P_\vartheta$ combined with crisp isosurfaces $FA^{-1}(\vartheta)$ and $RA^{-1}(\vartheta)$.

For the brain dataset, we computed the uncertainty propagation from the DTI eigenvalues to the scalar anisotropy indices, cf. Sect. 4.3.4. From the $FA$ and $RA$ fields and the corresponding uncertainty estimations for $\widehat{SNR}_\lambda = 10$ and $\widehat{SNR}_\lambda = 20$ we generated the uncertain isosurfaces shown in Fig. 5.12 (see also Fig. 4.6). We chose the threshold $\vartheta = 0.5$ for $FA$ that was used previously for the segmentation of brain structures using isosurfaces [STS07]. The threshold for the corresponding $RA$ isosurface is $\vartheta \approx 0.32$. Again, the spatial distributions of the isosurfaces are indicated by volume renderings of $P_\vartheta$ that surround the mean (crisp) surfaces.

## 5.6 Discussion

Compared to previous approaches to isosurface uncertainty the methods presented in this chapter differ in regard to mathematical modelling and visualization methods.

**Modeling and Computation.** Rhodes et al. [RLBS03] do not use an error model, but assume that uncertainty is somehow quantified ("error value") and provided in the data set. Grigoryan and Rheingans [GR04] computed the position of point primitives (depicting a probabilistic surface) by multiplying an uncertainty value, a random number and a user-defined scale




(a) $\kappa^{abs}_{FA^{-1}(\vartheta)}$ (b) $\kappa^{abs}_{FA,FA^{-1}(\vartheta)}$ (c) $P_{\vartheta,FA}$

(d) $\kappa^{abs}_{RA^{-1}(\vartheta)}$ (e) Relative differences $\frac{\kappa^{abs}_{FA,FA^{-1}(\vartheta)} - \kappa^{abs}_{RA,RA^{-1}(\vartheta)}}{\kappa^{abs}_{FA,FA^{-1}(\vartheta)}}$ (f) $P_{\vartheta,RA}$

**Figure 5.11:** *Textured slices in the spiral dataset: condition number for isosurface extraction for FA (a) and RA (d), combined condition for FA (b) and relative differences between the combined condition numbers (e). In the right column uncertain isosurfaces (assuming constant SNR=20 for all eigenvalues) are depicted.*

factor. When applying the method to data from tumor-growth simulations the authors use the error estimation from the simulation as input, while for clinical data they extract crisp isosurfaces and use the inverse density gradient magnitude at the surface points as an uncertainty measure. Their model does not consider probability distributions.

As we have shown in Sect. 4.2 the inverse gradient magnitude at the points of an isocontour is the absolute normwise condition. The position

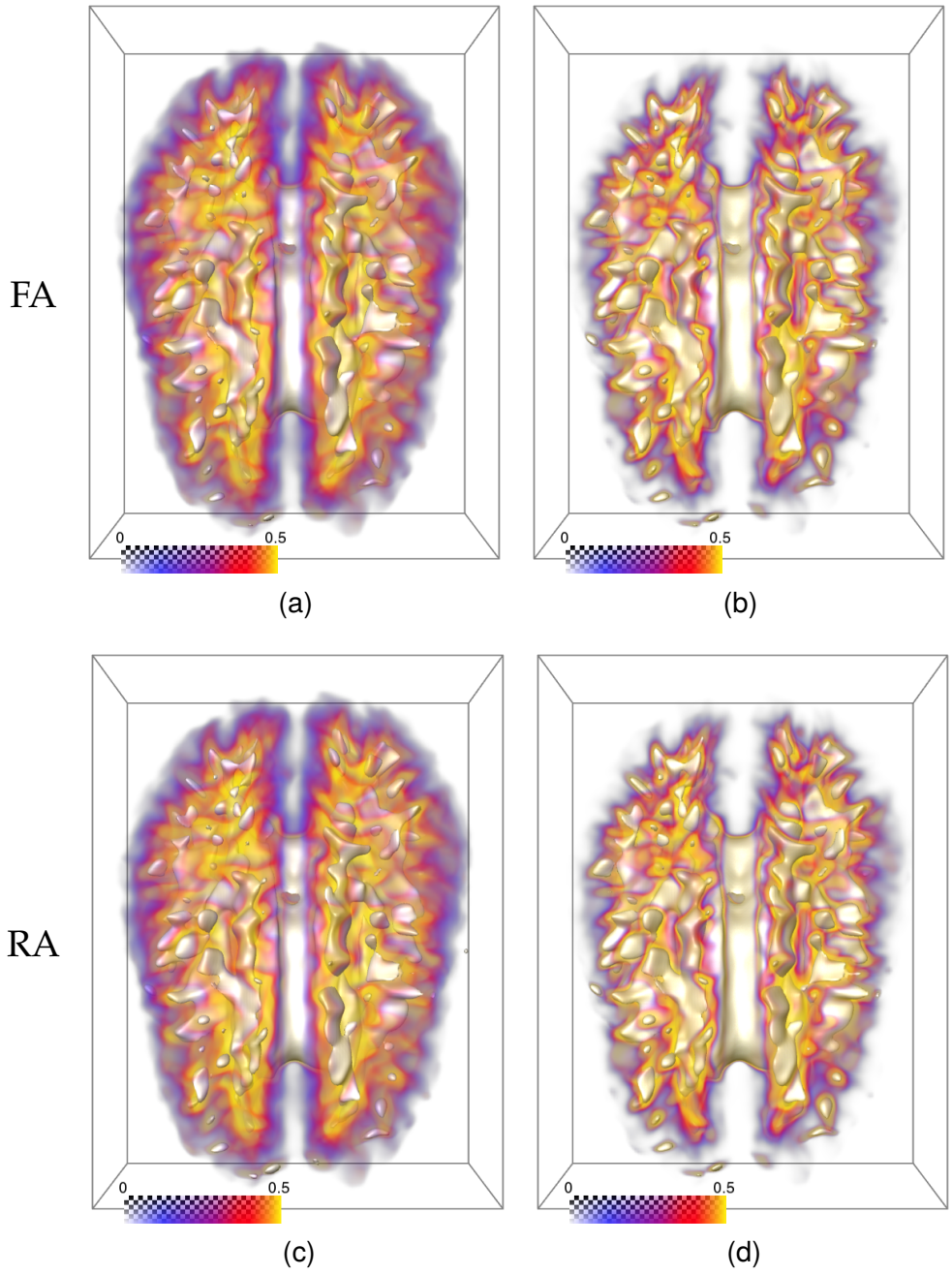



**Figure 5.12:** *Uncertain isosurfaces for* $\widehat{SNR}_\lambda = 10$ *(left) and* $\widehat{SNR}_\lambda = 20$ *(right) using FA (top) and RA (bottom). The mean crisp isosurfaces are shown in white while the level-crossing probabilities are mapped to color for volume rendering. The threshold is* $\vartheta = 0.5$ *for FA and* $\vartheta = 0.32$ *for RA.*

uncertainty of an isocontour, however, depends on the product of the condition number and the uncertainty of the input data. The ICD and LCP can be computed from datasets with uncertainty given by arbitrary probability distributions (with a finite number of moments) and in arbitrary-dimensional parameter spaces. The results of these functions are probabilities and probability densities, which means, that our methods quantify uncertainty in a statistically well-founded way.

**Visual Mapping.** The "error value" used by Rhodes et al. [RLBS03] is mapped to crisp isosurfaces by color or texture for visualization but the authors do not address positional uncertainty. Grigoryan and Rheingans [GR04] propose a heuristic to show the position uncertainty of surfaces. Clouds of point primitives are placed along the normals of crisp surfaces. An evaluation shows that this approach is more suitable to clarify position uncertainty compared to color coding. Zuk [Zuk08] points out that the user-defined scale factor in reference [GR04] may lead to an arbitrary *Lie Factor* in the visualizations.

The visualization methods presented in Sect. 5.4 are based on $P_\theta(\mathbf{x})$ and thus the output is a function of the input data only, i.e. there are no user-defined parameters. Difficulties in interpreting visualizations arise in areas of an image where the mean surface is normal to the camera vector. There, the amount of positional uncertainty is difficult to recognize. This can usually be worked around by the user by interactively changing the viewpoint. The separate display of the condition number $\kappa_{abs}$ and the standard deviation $\sigma(\mathbf{x})$ of the data at the surface points helps the user to understand *why* a given uncertain isosurface has a specific spatial distribution.

# 6

# Feature Probabilities in Discrete Random Fields

The methods to compute and visualize the positional uncertainty of isocontours presented in the previous chapter are based on discretely sampled uncertain scalar fields that are modelled by discrete random fields given on a computational grid, assuming that the random variables are spatially *un*correlated. A continuous extension of the discrete field is obtained by interpolating the PDFs such that the point-wise measures given in Eq. (5.2) and (5.11) can be evaluated.

In the following, we propose a framework that differs in two respects. We formulate a generic approach for computing grid cell-wise probabilities for the occurrence features where

1. any desired type of local feature can be defined using specific indicator functions, and

2. the computation of probabilities takes spatial correlations of the random field into account by integrating over multidimensional PDFs that are associated with the respective grid cells and their neighborhood.

This chapter is based on the publications [PWH11, PPH13, PPH12, PH13, PH14].

## 6.1 A Generic Framework

We assume that the data have been sampled on nodes of some mesh and that the uncertainty is modelled using an $\ell$-valued discrete random field as described in Chap. 3. In order to reveal the probability for the occurrence of a feature at some spatial location, we compute probabilities that features exist at each cell $c \in C_\eta$ where $C_\eta$ denotes the set of all cells in the grid for

which the feature is defined. The proposed procedure can be applied to any type of mesh, both structured and unstructured.

### 6.1.1 Feature Indicator Functions

Let $K_c$ be the number of degrees of freedom for cell $c$, e.g. the number of adjacent vertices for a polyhedral cell or the number of adjacent triangles for a vertex in a triangular mesh, then $\mathbf{Y}_c \in \mathbb{R}^{\ell K_c}$. A *feature indicator* $I$ is a boolean function defined for an $\eta$-cell $c$ and a realization of $\mathbf{Y}_c$ that determines if a feature occurs or not:

$$I : C_\eta \times \mathbb{R}^{\ell K_c} \to \{0, 1\} \tag{6.1}$$

Note that the neighborhood size of a cell is in general not the same for every cell in the grid, and depends on the combinatorial grid topology.

### 6.1.2 Feature Probabilities

Let $I$ be a feature indicator defined on $\eta$-cells of a parameter-discrete random field (parametric or nonparametric) defined on a computational grid. The cell-wise probability for the occurrence of the feature is then

$$\mathrm{P}_c = \int_D f_c(\mathbf{y}_c)\,\mathrm{d}\mathbf{y}_c = \int_{\mathbb{R}^{\ell K_c}} f_c(\mathbf{y}_c)\, I(c, \mathbf{y}_c)\,\mathrm{d}\mathbf{y}_c = \mathrm{E}(I(c, \cdot)), \tag{6.2}$$

where $D = \{\mathbf{y}_c \in \mathbb{R}^{\ell K_c} \mid I(c, \mathbf{y}_c) = 1\}$ and $f_c$ is probability density function of the $(\ell K_c)$-variate distribution associated with cell $c$ and $\mathbf{y}_c$ is a realization of $\mathbf{Y}_c$. The probability $\mathrm{P}_c$ can also be considered as the expected value of the feature indicator $I$ in cell $c$ and with respect to $f_c$.

### 6.1.3 Numerical Integration

The integral in Eq. (6.2) can be approximated using Monte Carlo sampling. For each type of PDF a specific sampling method has to be employed. To draw samples from a Gaussian distribution, e.g. to sample from a parametric normal distribution or a Gaussian kernel for KDE, we apply a 2 step algorithm. Uncorrelated samples conforming to a uniform distribution are generated and converted to normally distributed values using the Box-Muller transform [BM58]. These samples are adjusted to the multivariate normal distribution by applying a Cholesky decomposition to the covariance matrix and multiplying the samples with the lower triangle matrix [Gen04, p. 197]. Refer to Chap. 3 for methods to sample from nonparametric distributions. Gentle provides a comprehensive overview of random sampling algorithms in his book [Gen04]. In case KDE is used with the PC transformation method each Monte Carlo sample has to be transformed back to the original basis

for the evaluation of $I$ because indicator functions are defined with respect to this basis and not in terms of the PCA modes.

The number of vector components $\ell K_c$ can vary depending on the grid topology. For example, there can by vertices with different numbers of adjacent triangles in a triangulated 2D domain. The realizations evaluated by the indicator function $I$. From the ratio of samples that agree to the respective feature to those that don't we compute the feature probability for each grid cell. The number of samples can be manually set to a sufficient value, such that no significant Monte Carlo noise is observable anymore. Using pseudocode the algorithm can be summarized as follows:

**for each** cell $c$ {
  estimate density $f_c$
  #features $\leftarrow 0$
  **for** $1 \ldots$ #samples {
    $\mathbf{y} \leftarrow$ random sample $(y_1, \ldots, y_{(\ell K_c)})^T \sim f_c(\mathbf{y})$
    **if**$(I(c, \mathbf{y}) == 1)$ #features $\leftarrow$ #features $+ 1$
  }
  $P_c \leftarrow$ #features/#samples
}

The computational complexity for calculating the feature probability of one grid cell using MC integration is $O(\epsilon^{-2})$, where $\epsilon$ is the integration error. This integration method can be computationally very expensive, depending on the size of the grid, the probability distributions and the number of samples (or predefined $\epsilon$). Thus, it is desirable to find simplifications and approximations in order to speed up the computations, cf. Sect. 6.4.

## 6.2 Cell-Based Level-Crossing Probabilities

Based on the general approach presented in the previous chapter we aim to formulate a method to quantify the spatial distribution of isocontours in a discrete random field $\mathbf{Y}$ taking the spatial correlations into account.

Any realization $\mathbf{y}$ of $\mathbf{Y}$ defines a grid function. For any grid function imagine an extension to a $C^0$ function $\mathbf{y}^*$ that is defined in the continuous domain and that interpolates between the sample points $\mathbf{x}_j$ such that in each $\eta$–cell $c$ ($\eta \leq N$) the extremal values are taken at the vertices of $c$. Examples for such interpolations are linear interpolation for simplicial cells and $\eta-$linear interpolation for $\eta-$dimensional polytope cells.

Let $J$ be the set of indices of the vertex points of cell $c$. Then cell $c$ crosses the $\vartheta$-level of $\mathbf{y}^*$ if and only if in the set of differences $(y_j - \vartheta)_{j \in J}$

both signs occur. Equivalently, cell $c$ does not cross the $\vartheta$-level of $\mathbf{y}^*$, if and only if all differences $(y_j - \vartheta)_{j\in J}$ have the same sign. We want to compute the probability that a $\eta$-cell $c$ of the $N$-dimensional sample grid crosses the $\vartheta$-level of interpolated realizations of the random variables $Y_{j\in J}$. We call this the *$\vartheta$-level-crossing probability* of cell $c$ and denote it by $\mathrm{P}_c(\vartheta\text{-crossing})$.

The main differences of this approach compared to the methods presented in Chap. 5 are that (i) the results are computed per grid cell and not per point in the continuous domain and (ii) the correlation structure of the field is considered.

### 6.2.1 Indicators Functions for Level Crossings

To compute *level-crossing probabilities* in uncertain scalar fields ($\ell = 1$) with cells $c \in C_\eta$ and $\eta \geq 1$, i.e. probabilities that intersections between the field and a predefined level (or isovalue) $\vartheta$ exist, we use the indicator function

$$I_{\text{cross}}(c, \mathbf{y}_c) = \begin{cases} 0 & \forall\, y_{i,c}\,(y_{i,c} \leq \vartheta) \;\vee\; \forall\, y_{i,c}\,(y_{i,c} > \vartheta) \\ 1 & \text{otherwise,} \end{cases} \tag{6.3}$$

where $y_{i,c}$ are the components of $\mathbf{y}_c$.

In order to compute the probability we have to integrate the joint density function $f_c$ of the random vector $\mathbf{Y}_c$ over regions where the indicator function $I_{\text{cross}}$ has value 1. Adapting Eq. (6.2) we obtain a special case of the general feature probability formula

$$\mathrm{P}_c(\vartheta\text{-crossing}) = \int_D f_c(\mathbf{y}_c)\,\mathrm{d}\mathbf{y}_c = \int_{\mathbb{R}^{\ell K_c}} f_c(\mathbf{y}_c)\; I_{\text{cross}}(c, \mathbf{y}_c)\,\mathrm{d}\mathbf{y}_c,$$

where $D = \{\mathbf{y}_c \in \mathbb{R}^{\ell K_c} \mid I_{\text{cross}}(c, \mathbf{y}_c) = 1\}$.

The general procedure to compute such probabilities can be applied to any type of mesh entity, for example to arbitrary polyhedral cells in grids of arbitrary dimension $d$. In the following we consider exemplarily edges, rectangles and cuboids (duals of voxels) in such grids. Obviously the procedure can be extended to $\eta$-simplices or arbitrary $\eta$-polyhedra with $\eta \leq d$.

### 6.2.2 Level-Crossing Probabilities for Different Cell-Types

An equivalent and intuitive formulation to compute the level-crossing probability is to integrate the joint density function $f_c$ over sets $\{y_j \in \mathbb{R} \mid y_j \leq \vartheta\}$ and $\{y_k \in \mathbb{R} \mid y_k \geq \vartheta\}$ using Eq. (3.15).
Alternatively we can compute the probability

$$\mathrm{P}_c(\vartheta\text{-crossing}) = 1 - \mathrm{P}_c(\vartheta\text{-non-crossing})\,, \tag{6.4}$$

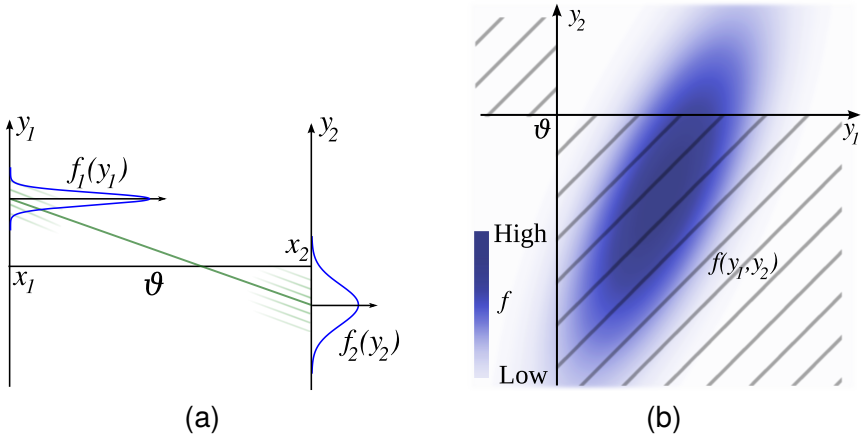



**Figure 6.1:** *Example for the 1D case (edge): (a) The marginal distributions at the grid points are shown in blue. Exemplarily, one realization of the linear interpolant is shown (green solid line); other realizations with lower probability are indicated (green transparent lines). In the depicted case the $\vartheta$-level-crossing probability is relatively high. In (b) a density plot of the joint distribution with a correlation coefficient of 0.75 is displayed. The quadrants constituting the integration domain for the computation of the level-crossing probability are indicated by the hatched grey area.*

which in cells of dimension greater then one is less expensive to calculate. In the following we show more specific equations for crossing probabilities for several different grid cell types.

**Edges (1-cells).** For a scalar field in one or more dimensions we consider two random variables $Y_1$, $Y_2$ that are associated with adjacent grid points $x_1$, $x_2$. Consider the random vector $\mathbf{Y}_c = [Y_1, Y_2]$ where the joint probability distribution is described by a bivariate PDF $f_c(y_1, y_2)$ with $y_1, y_2 \in \mathbb{R}$, see Fig. 6.1.

The $\vartheta$-level-crossing probability is given by

$$\begin{aligned} \mathrm{P}_c(\vartheta\text{-crossing}) &= \mathrm{P}(Y_1 \leq \vartheta, Y_2 > \vartheta) + \mathrm{P}(Y_1 > \vartheta, Y_2 \leq \vartheta) \\ &= \int\limits_{y_1 \leq \vartheta} \mathrm{d}y_1 \int\limits_{y_2 > \vartheta} \mathrm{d}y_2 \, f_c(y_1, y_2) + \int\limits_{y_1 > \vartheta} \mathrm{d}y_1 \int\limits_{y_2 \leq \vartheta} \mathrm{d}y_2 \, f_c(y_1, y_2) \end{aligned} \tag{6.5}$$

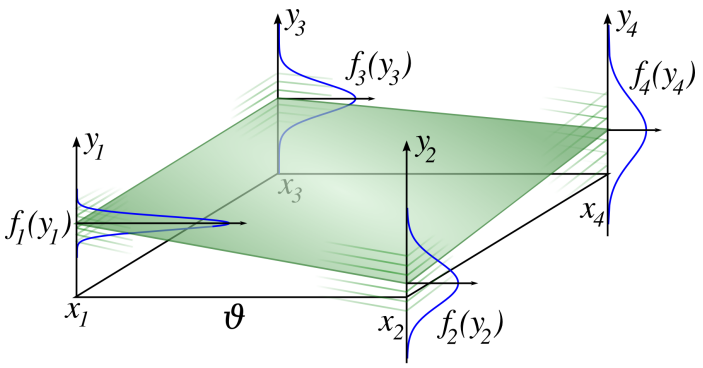



**Figure 6.2:** *Example for the computation of a level-crossing probability in 2D: The marginal distributions at the grid points are shown in blue. Exemplarily, one realization of the bilinear interpolant is shown in green (the particular one, where all random variables take the value of their means). The $\vartheta$-level-crossing probability of the interpolant is relatively low in this specific case.*

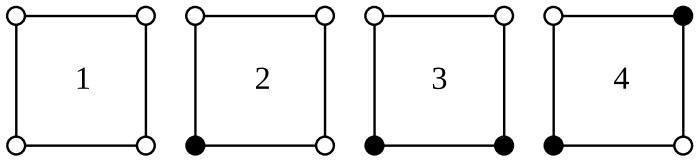



**Figure 6.3:** *Four distinct configurations for the marching squares algorithm. The other configurations can be constructed by inverting, rotating and mirroring the grid points. The integrals of the probabilistic formulation correspond to these cases.*

Alternatively:

$$\begin{aligned} \mathrm{P}_c(\vartheta\text{-non-crossing}) = &\quad \mathrm{P}(Y_1 \leq \vartheta, Y_2 \leq \vartheta) + \mathrm{P}(Y_1 > \vartheta, Y_2 > \vartheta) \\ = &\int\limits_{y_1 \leq \vartheta} \mathrm{d}y_1 \int\limits_{y_2 \leq \vartheta} \mathrm{d}y_2 \; f_c(y_1, y_2) + \int\limits_{y_1 > \vartheta} \mathrm{d}y_1 \int\limits_{y_2 > \vartheta} \mathrm{d}y_2 \; f_c(y_1, y_2). \end{aligned} \tag{6.6}$$

Since the four quadrants

$$\begin{aligned} &\{(y_1, y_2) | y_1 \leq \vartheta \text{ and } y_2 \leq \vartheta\}, \\ &\{(y_1, y_2) | y_1 \leq \vartheta \text{ and } y_2 > \vartheta\}, \\ &\{(y_1, y_2) | y_1 > \vartheta \text{ and } y_2 \leq \vartheta\} \text{ and} \\ &\{(y_1, y_2) | y_1 > \vartheta \text{ and } y_2 > \vartheta\} \end{aligned}$$

are disjoint and their union is $\mathbb{R}^2$ we can read off Eq. (6.4).

**Rectangles (2-cells).** For a scalar field in two or more dimensions we consider $\mathbf{Y}_c = [Y_1, Y_2, Y_3, Y_4]$ at the grid points $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, $\mathbf{x}_4$ that are the corners of a pixel, see Fig. 6.2. The joint probability distribution is described by a four-dimensional Gaussian PDF $f_c(y_1, y_2, y_3, y_4)$ with $y_1, y_2, y_3, y_4 \in \mathbb{R}$.

Using integrals over $f_c$ we can compute probabilities for the different cases of the *marching squares* algorithm, see Fig. 6.3. Probabilities for the four distinctive cases are:

$$\mathrm{P}_{\vartheta,1} = \int_{(y_1 > \vartheta \wedge y_2 > \vartheta \wedge y_3 > \vartheta \wedge y_4 > \vartheta)} \mathrm{d}y_1 \int \mathrm{d}y_2 \int \mathrm{d}y_3 \int \mathrm{d}y_4 \; f_c(y_1, y_2, y_3, y_4) \tag{6.7}$$

$$\mathrm{P}_{\vartheta,2} = \int_{(y_1 \leq \vartheta \wedge y_2 > \vartheta \wedge y_3 > \vartheta \wedge y_4 > \vartheta)} \mathrm{d}y_1 \int \mathrm{d}y_2 \int \mathrm{d}y_3 \int \mathrm{d}y_4 \; f_c(y_1, y_2, y_3, y_4) \tag{6.8}$$

$$\mathrm{P}_{\vartheta,3} = \int_{(y_1 \leq \vartheta \wedge y_2 \leq \vartheta \wedge y_3 > \vartheta \wedge y_4 > \vartheta)} \mathrm{d}y_1 \int \mathrm{d}y_2 \int \mathrm{d}y_3 \int \mathrm{d}y_4 \; f_c(y_1, y_2, y_3, y_4) \tag{6.9}$$

$$\mathrm{P}_{\vartheta,4} = \int_{(y_1 \leq \vartheta \wedge y_2 > \vartheta \wedge y_3 \leq \vartheta \wedge y_4 > \vartheta)} \mathrm{d}y_1 \int \mathrm{d}y_2 \int \mathrm{d}y_3 \int \mathrm{d}y_4 \; f_c(y_1, y_2, y_3, y_4) \tag{6.10}$$

The remaining 12 cases can be constructed by rotating and mirroring the grid points.

The level-crossing probability for a pixel can be computed by considering the *complement* of the cases where *no* level crossing occurs:

$$\mathrm{P}_c(\vartheta\text{-crossing}) = 1 - \int_{\substack{(y_1 \leq \vartheta \wedge y_2 \leq \vartheta \wedge y_3 \leq \vartheta \wedge y_4 \leq \vartheta) \\ \vee (y_1 > \vartheta \wedge y_2 > \vartheta \wedge y_3 > \vartheta \wedge y_4 > \vartheta)}} \mathrm{d}y_1 \int \mathrm{d}y_2 \int \mathrm{d}y_3 \int \mathrm{d}y_4 \, f_{\mathbf{Y}}(y_1, y_2, y_3, y_4) \tag{6.11}$$

**Cuboids (3-cells).** For a scalar field in three or more dimensions we consider 8 random variables located at the corners of a cuboid, whose joint probability function is an 8-dimensional Gaussian PDF. Of the $2^8 = 256$ cases, we have 254 cases with crossing (comprised of 14 distinct marching cubes cases) and 2 cases without crossing. The simplest way to compute level-crossing probabilities again is to use Eq. (6.4) and compute probabilities that *no* level crossing occurs, analogously to Eq. (6.11).

### 6.2.3 Visual Mapping

We employ a visual mapping for discrete fields that is similar to the methods described in Sect. 5.4. The main difference is that the smallest scale of features that can be resolved for visualization is determined by the computational grid while the continuous methods can reveal finer structures

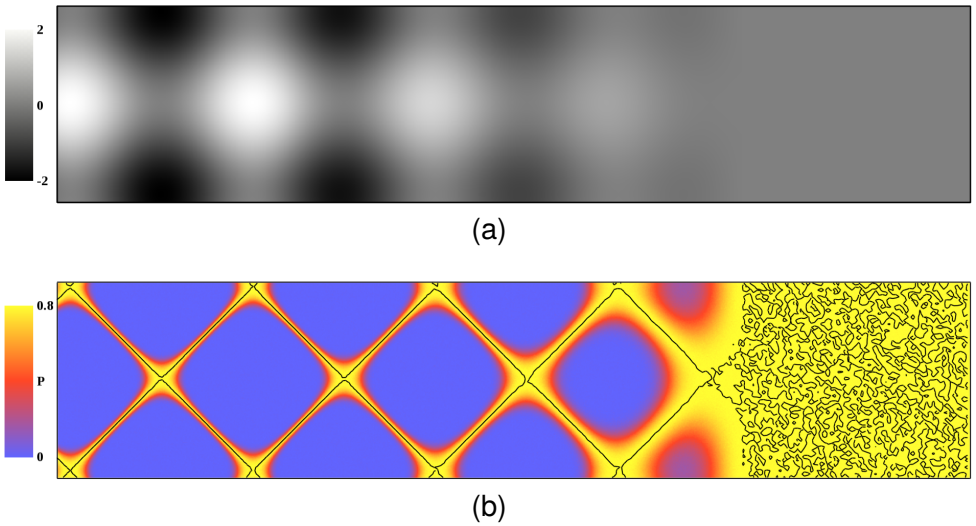



**Figure 6.4:** *Uncertain isolines for a synthetic 2D dataset. The expected values conform to a sine pattern (with a low amount of noise added) on the left that gradually approaches a plateau on the right as seen in (a). The variances and covariances are constant. In (b) the probabilities for $\vartheta = 0$ are color mapped while the crisp isoline of the expected values is shown in black. While the computation of isolines is ill-conditioned at critical points (especially plateaus) the probabilistic ansatz does not suffer from this problem and calculates high probabilities for the whole plateau.*

due to the interpolation of the PDFs (in contrast to the possible interpolation of *probabilities* in the discrete setting). We show crisp isosurfaces of the mean values because in many cases it represents the most probable shape of the isosurface. This surface is augmented with the volume rendered level-crossing probabilities. This visual design corresponds to traditional 2D plots with error bars where the mean value is shown like a crisp, certain value that is augmented with the standard deviation. The development of refined visual designs conveying uncertainty as well as assessment of their visual effectiveness is a promising area of research but it is beyond the scope of this thesis.

### 6.2.4 Results and Discussion

To illustrate the essential properties of the methods we apply them to synthetic datasets. We show the method's effectiveness for real world data by applying it to ensemble datasets from climate research and biofluid mechanics. The computations were performed on an Intel Xeon X5550 2.66 GHz system.

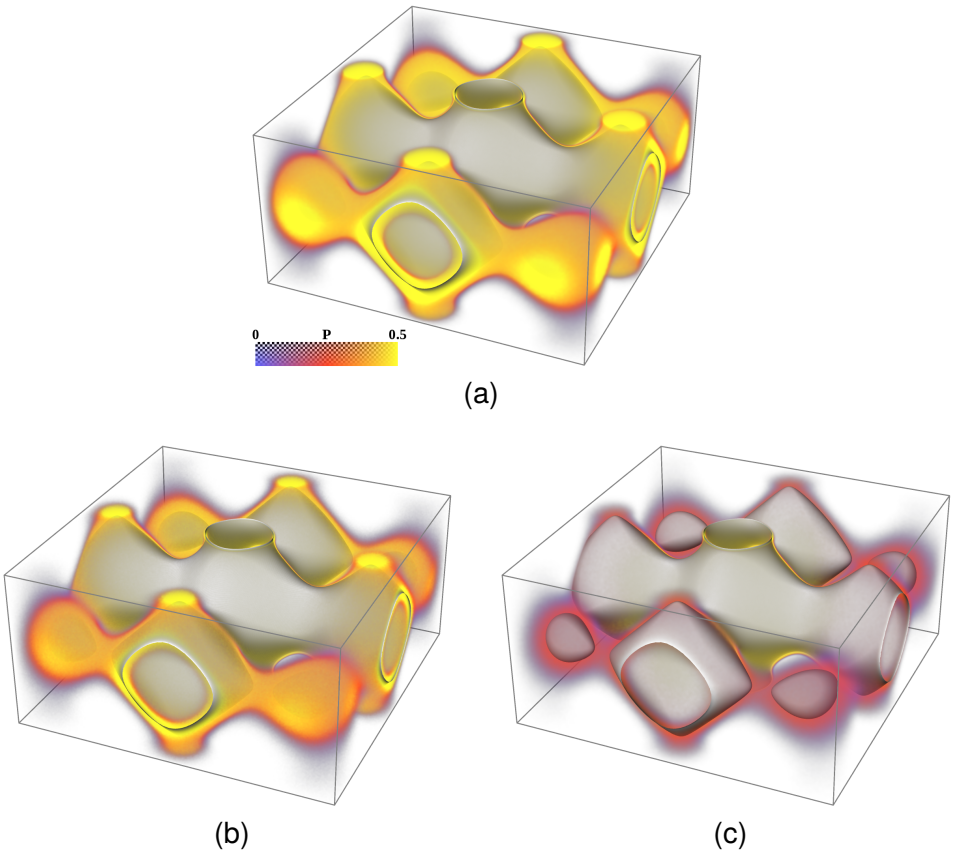



**Figure 6.5:** *Uncertain isosurfaces in a synthetic 3D dataset. The expected values are given by an analytic formula and the variances are constant. The correlation coefficient is globally set to* 0 *in (a), to* 0.65 *in (b) and to* 0.95 *in (c). The probabilities are displayed using direct volume rendering and a crisp isosurface of the expected values is shown in white. The results show that increasing correlation between the grid points (from (a) to (c)) decreases the level-crossing probabilities in the proximity of the mean surface and leads to more localized spatial distributions of uncertain isocontours.*

**Synthetic Datasets.** Fig. 6.4 shows uncertain isolines for a synthetic 2D dataset. The expected values of the input data (with a low amount of noise) correspond to a sine pattern on the left side of the image that gradually blends to a plateau on the right, see, shown in Fig. 6.4a. The variances and covariances are constant. The probabilities in the grid of $1000 \times 250$ pixels were computed using 4000 samples per pixel in 225 seconds. In Fig. 6.4b the level-crossing probabilities for $\vartheta = 0$ are mapped to color while the crisp isoline in the mean value field is shown in black. Uncertain isosurfaces in a synthetic 3D dataset are displayed in Fig. 6.5. The expected values of the

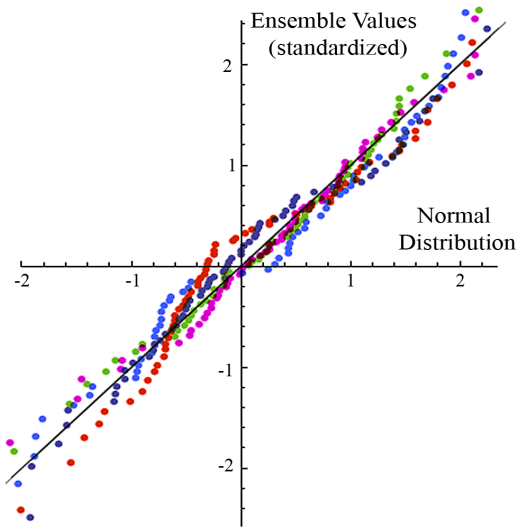



**Figure 6.6:** *Assessment of normality for 5 randomly chosen distributions from the temperature field ensemble using a Q-Q-Plot. The distributions do not show severe deviations from the normal distribution, i.e. small differences compared to a linear shape.*

input data are given by the simple analytic formula $\mu(x,y,z) = (\cos(7x) + \cos(7y) + \cos(7z))\exp(-4.5r)$, where $r = \sqrt{x^2+y^2+z^2}$. The variances are constant in the 3 images. A global correlation coefficient (for each pair of vertices) is varied to study the influence of correlation and set to 0 in 6.5a, to 0.65 in 6.5b and to 0.95 in 6.5c. The probabilities in the grid of $256 \times 256 \times 128$ voxels were computed using 1600 MC samples each in $\approx 45$ minutes for each result. The level-crossing probabilities ($\vartheta = 0.013$) are displayed using DVR. A crisp mean isosurface is shown in white.

Fig. 6.4b illustrates that the computation of crisp isolines is ill-conditioned at critical points (especially plateaus) while the probabilistic ansatz does not suffer from this problem and calculates high probabilities for the whole plateau.

**Results from Climate Simulations.** We apply the algorithm to daily average hindcast data from the DEMETER project [Pal04]. Data from this project was also employed for the results in Chap. 5. The means, variances and covariances are computed from a temperature field ensemble for Feb 20th, 2000. To check whether the modelling of the data using Gaussian distributions is appropriate we assess normality using a Q-Q plot [Job91, p. 63]. An example is shown in Fig. 6.6 where quantiles for 5 randomly chosen distributions from the temperature field ensemble are displayed. In this field the

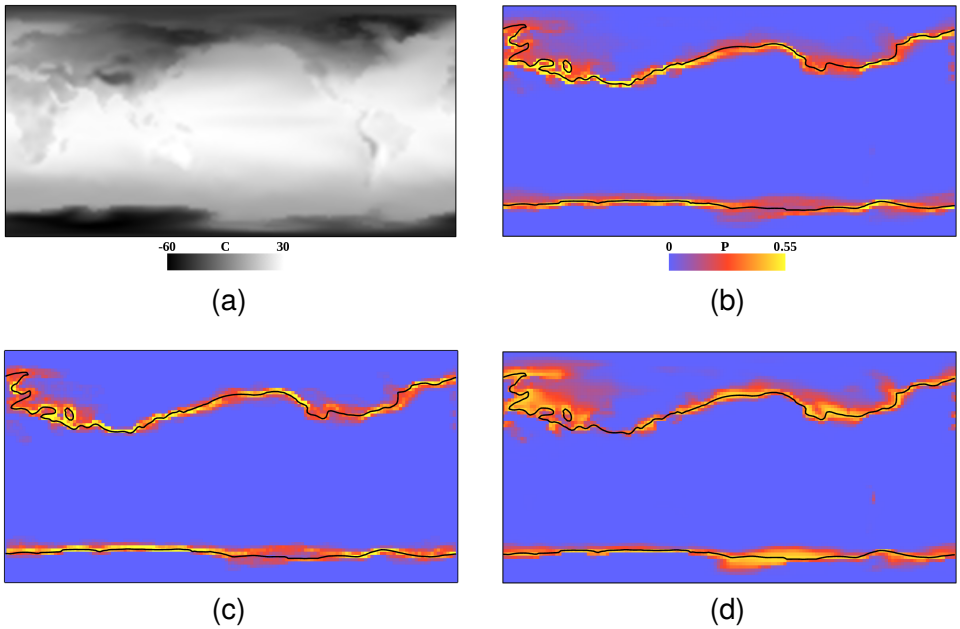
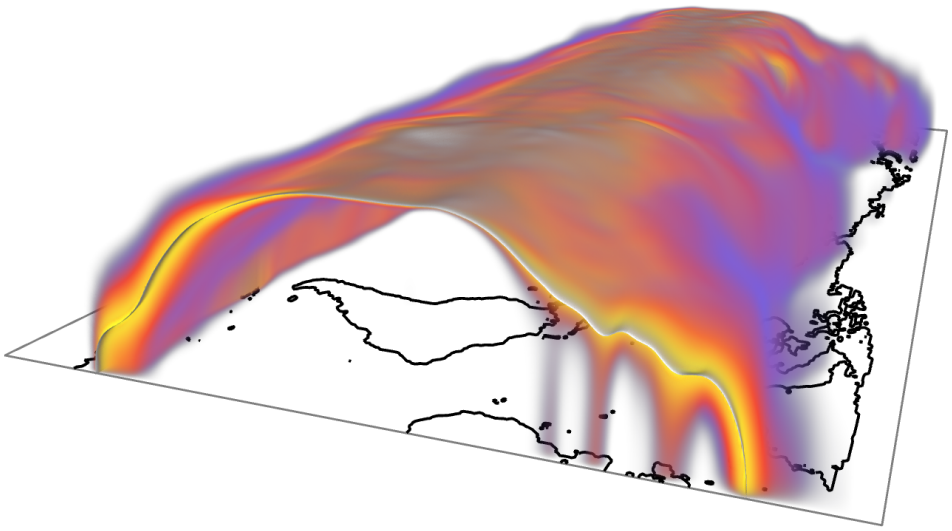


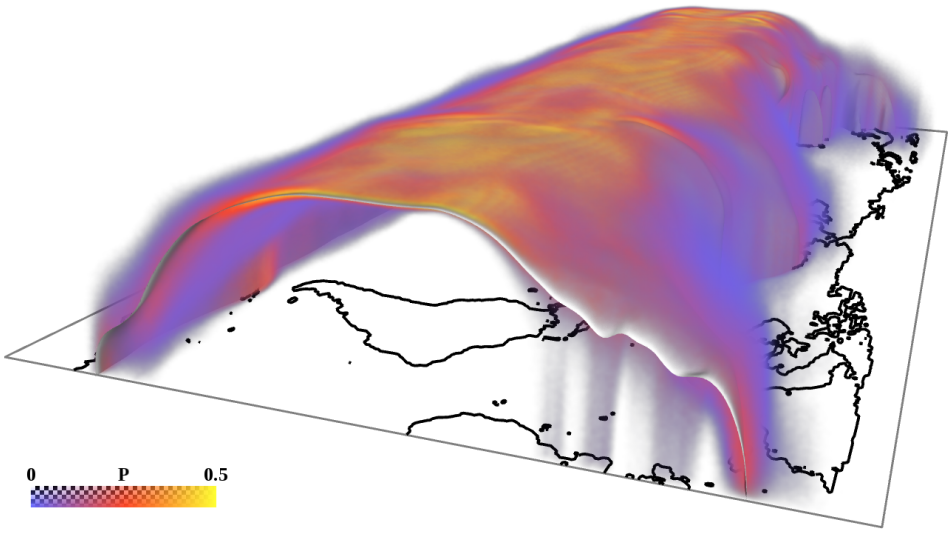

**Figure 6.7:** *Results for a 2 metre temperature field from climate simulations: The ensemble means are shown in (a). The level-crossing probabilities for $\vartheta = 0°C$ are colormapped in (b). For comparison the relative count of crisp isolines in the 63 ensemble members crossing the respective grid cell is shown in (c). The results computed using point-wise LCP (Eq. (5.10), i.e. not considering correlation) are shown in (d). While the latter result overestimates the spatial distribution of the uncertain isoline the distribution in (b) is more localized and similar to (c).*

data data approximately Gaussian distributed.

Results for the 2 meter temperature field are shown in Fig. 6.7. The ensemble means are shown in 6.7a. The level-crossing probabilities ($\vartheta = 0°C$) are mapped to color in 6.7b. The probabilities in the grid of $144 \times 73$ pixels were computed using 8.000 samples per pixel in 11 seconds. For comparison the relative count of crisp isolines in the 63 ensemble members crossing the respective grid cell is shown in 6.7c. The probabilities, computed using point-wise LCP employing Eq. (5.10) and not considering correlation, are displayed in 6.7d. The corresponding mean isoline is depicted in black. We combine the 2 meter data set with temperature fields of pressure levels 850, 500 and 200 hPa in the earth's atmosphere to obtain a 3D ensemble where the third coordinate represents air pressure. We compute means, variances and covariances for all hexahedral grid cells. Fig. 6.8 shows uncertain isosurfaces $\vartheta = 0°C$. For Fig. 6.8a the probabilities are computed using point-wise LCP (not considering correlation). For the probabilities that are displayed



(a)




(b)

**Figure 6.8:** *Uncertain isosurfaces $\vartheta = 0°C$ in a 3D temperature field. In Fig. (a) the probabilities computed using point-wise LCP (not considering correlation) are shown. For Fig. (b) correlation was considered and the level-crossing probabilities reveal a more localized spatial distribution of the uncertain isosurface.*

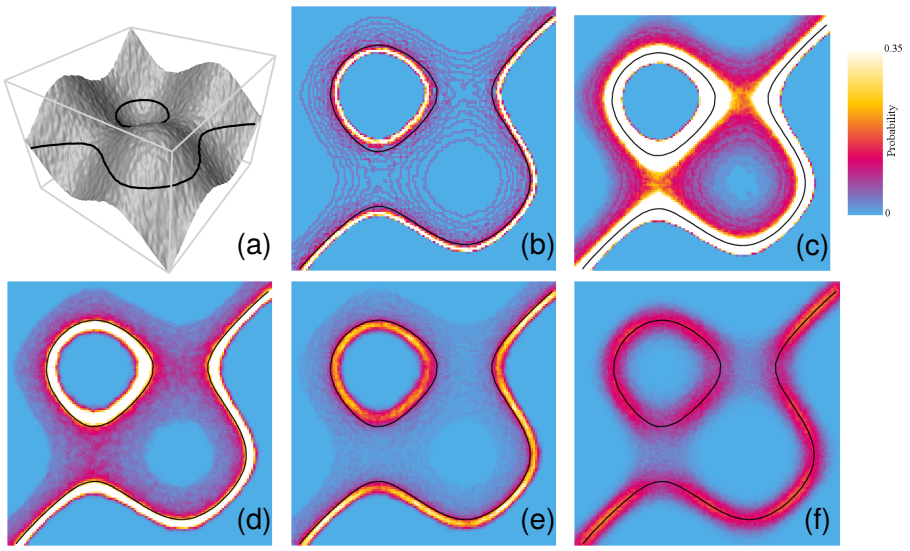



**Figure 6.9:** *(a) A single member of the scalar field ensemble displayed as a heightmap with an isoline. Color mapped level-crossing probabilities computed from (b) the empirical distributions, (c) histograms, (d) kernel density estimates with untransformed data and (e) kernel density estimates working with PC transformation. The isoline of the mean field is shown in black. (f) For comparison: Level-crossing probabilities computed using a parametric Gaussian model.*

in Fig. 6.8a correlation was taken into account. The probabilities in a grid resampled to $432 \times 219 \times 68$ voxels were computed using 8.000 samples per pixel in 194 minutes.

**Comparison of Probabilistic Models.** To illustrate the differences between the models described above we computed level-crossing probabilities for 5 different models that were derived from a synthetic ensemble dataset. The ensemble consists of 32 realizations of a sine pattern where Gaussian noise and a varying bias were added to all scalar values in every ensemble member, which leads to skewed distributions. 1D marginal distributions for a single vertex of the field are shown in Fig. 6.10. The empirical distribution is drawn in red, the histogram in blue, a parametric normal distribution in yellow and a kernel density estimate in violet. Fig. 6.9 (a) shows a single member of the ensemble displayed as a heightmap with an isoline. Level-crossing probabilities for the same isovalue are displayed using color mapping in Fig. 6.9 (b)-(f). Due to the skewed distribution the maximal probabilities are expected to occur off the isoline of the mean field. Indeed, this can be observed in Fig. 6.9 (b), (d) and (e). However, with the Gaussian model in Fig. 6.9 (f) the ridges of the probability field coincide with the isoline of the mean field.

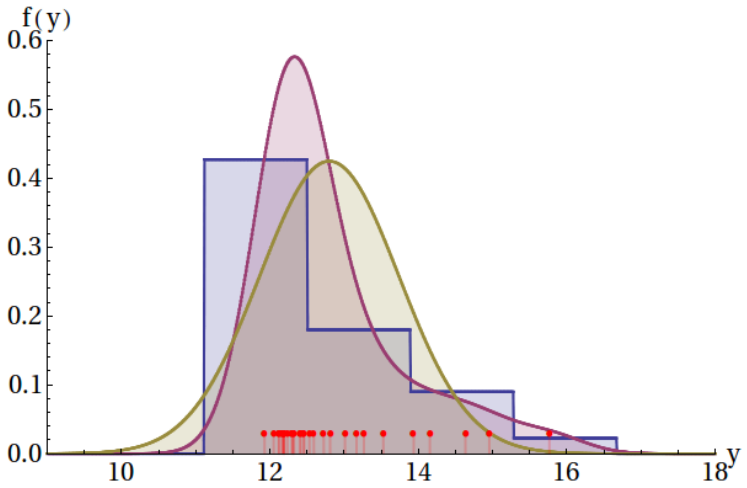



**Figure 6.10:** *1D marginal distributions, red: empirical distribution, blue: histogram, yellow: normal distribution (parametric model), violet: kernel density estimate*

In the results computed from histograms and KDE without PC transformation (Fig. 6.9 (c) and (d), respectively) we observe that the probabilities are higher and the spatial distributions are wider, compared to the other models. This is due to fact that these models cannot adequately represent the correlations present in the data. For this reason we did not consider these models in the subsequent examples.

We also compared the models computed from the 2 meter temperature ensemble, day 90 of the 2000-02 hindcast, from a climate simulation of the DEMETER project [Pal04]. To quantify the goodness of fit of the data to the Gaussian distribution we performed the Shapiro-Wilk test on 1D marginal distributions at each vertex of the grid. The resulting $p$-values that indicate the probability that the ensemble values were drawn form a Gaussian distribution are visualized in Fig. 6.11 (b). Fig. 6.11 (a), (c) and (d) show level-crossing probabilities for the isovalue $\vartheta = 24°\text{C}$ using color mapping. We can observe that the most significant differences between KDE and the Gaussian model occur in the center and lower center of the dataset – regions where the ensemble tends to be non-Gaussian, as indicated by the low $p$-values of the Shapiro-Wilk test. Fig. 6.11 (c) contains thin features with more detail, whereas Fig. 6.11 (d) appears more smoothed out. The regions with significant differences between KDE and the parametric Gaussian model are highlighted by white boxes in the figures.

Level-crossing probabilities for the pressure field on the vessel and aneurysm wall are mapped to color in Fig. 6.12. In Fig. 6.12 there are also structures in the KDE result (b) that are not present in the Gaussian result (c).

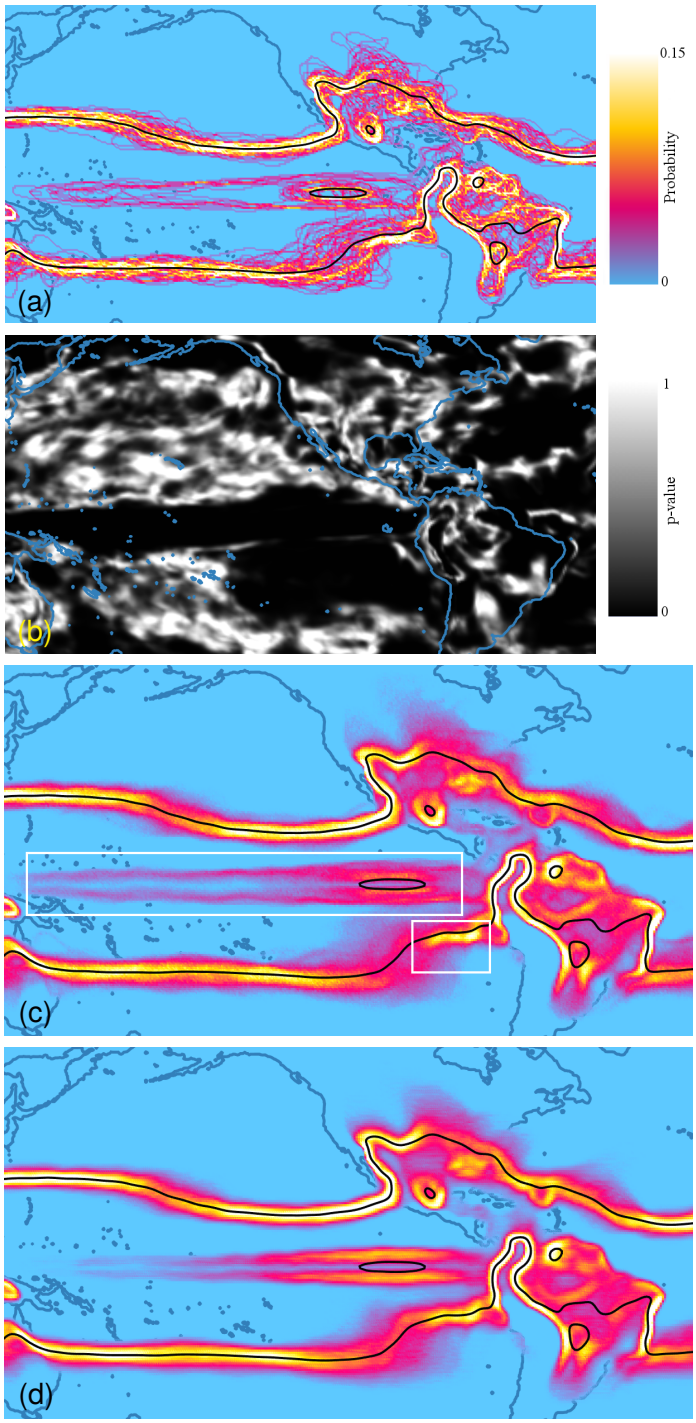



**Figure 6.11:** *Level-crossing probabilities ($\vartheta = 24°\mathrm{C}$) in the 2 meter temperature field from a climate simulation of the DEMETER project [Pal04] are mapped to color. The results were computed using (a) empirical distributions, (c) KDE and (d) a parametric Gaussian model. The results of vertex-wise Shapiro-Wilk tests (p-values) are visualized in (b). The regions with the most significant differences between (c) and (d) are highlighted by white boxes.*

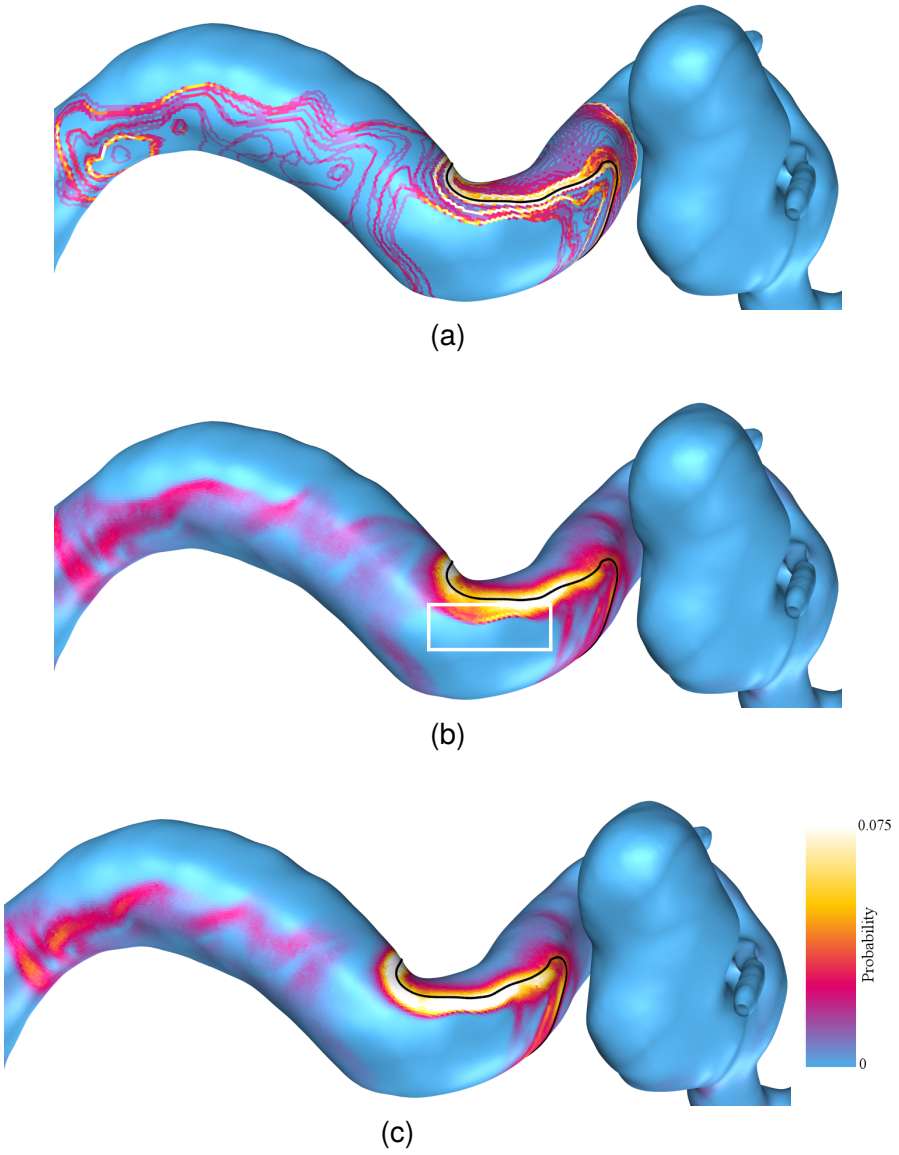



**Figure 6.12:** *Level-crossing probabilities for the pressure field on the vessel and aneurysm wall are mapped to color. The probabilities were computed using (a) empirical distributions, (b) KDE and (c) a parametric Gaussian model.*

## 6.3 Feature Probabilities in Uncertain Vector Fields

In previous work several approaches to visualize uncertain vector fields have been proposed, see Sect. 2.2. However, most of the methods assume the vectors of the field to be statistically independent, i.e. only point-wise marginal are considered distributions but no spatial correlations. Taking the corre-

lations into account increases the dimensionality of the data compared to independent distributions. Direct visualization (using e.g. glyphs) is therefore very challenging or not feasible at all.

### 6.3.1 Feature Types and Models for Vector-Valued Random Fields

We propose methods to compute spatial distributions of local features from uncertain vector fields considering the local correlation structure. These distributions can be used to display important structures of the data. In this thesis we focus on critical points (sources, sinks and saddles) and swirling motion vortex cores, but the proposed general approach can be applied to other local features as well.

The following methods are based on discretized random vector fields that are sampled on cells or nodes of structured or unstructured grids as described in Chap. 3. The resulting probabilities for the presence of a feature at spatial locations are defined in terms of *feature indicator functions* which were introduced in Sect. 6.1 and computed using a Monte Carlo method. An interesting question from the application point of view is the locality of features, i.e. the question how far regions, in which some feature is notably present, are extended.

For the computation of feature probabilities both parametric Gaussians and nonparametric models can be employed. Whether or not a given field is Gaussian, can either be statistically tested or assured by empirical knowledge and statistical considerations. An example of the second case are measurements of blood flow and tissue velocity by phase contrast MRI; due to the inherent noise in MRI the resulting vector fields are uncertain and can be shown to be correlated Gaussian random fields [FHH*11]. However, complex flows can also exhibit clearly non Gaussian distributions, e.g., in some cases there may be multimodal distributions due to the multistability of the system.

### 6.3.2 Critical Points in 2D

Isolated critical points in a crisp vector field $\mathbf{v}(x)$ are defined to be the points of the domain where the vector field is 0, e.g., the set of $x$ with $\mathbf{v}(x) = 0$, and the vector field is non-zero in an infinitesimal small neighborhood of those points.

The Poincaré index relates the occurrence of critical points in a volume to the vectors on the surface of that volume. It measures the signed winding number of the vectors along the surface of an oriented topological sphere. If the index is non-zero, the sphere encloses an isolated critical point. The sign of the Poincaré index is the same as the sign of the determinant of the Jacobian matrix of the vector field at the critical point.

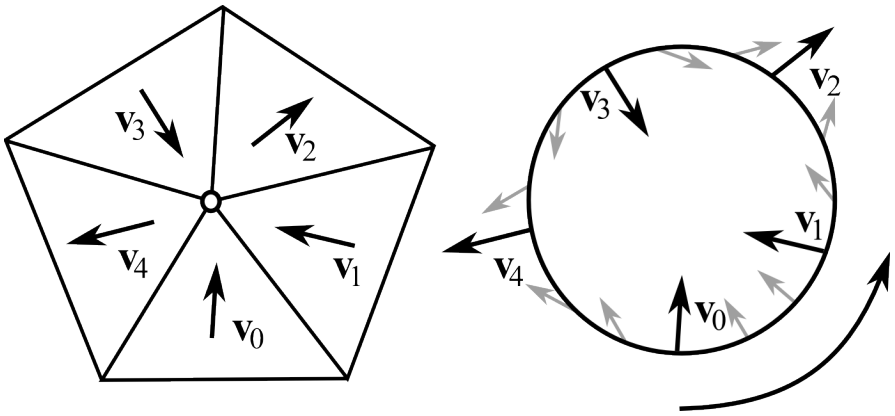



**Figure 6.13:** *Winding number calculation of a saddle point in a piecewise constant triangulated vector field. Following vectors in counter-clockwise direction from $v_0$ to $v_4$, vectors rotate clockwise, yielding the winding number of $-1$.*

In 2D, the vectors along a path are traversed in counter-clockwise direction as illustrated in Fig. 6.13, vector directions along the path cover a circle an integer multiple of times.

We use the Poincaré index for the identification of critical points by considering the vectors on the surface of the neighborhood. We compute discrete angles between adjacent vectors in $[-\pi, \pi)$, i. e., the smaller angle of the two possible rotation directions. This is equivalent to component-wise linear interpolation of vectors, and can be seen as follows: consider two linearly interpolated vectors $\mathbf{v}_0$ and $\mathbf{v}_1$,

$$\mathbf{v}(t) = (1-t)\mathbf{v}_0 + t\mathbf{v}_1, \tag{6.12}$$

with $t \in [0,1]$. The vector product $\mathbf{v}_0 \times \mathbf{v}(t)$ is

$$\mathbf{v}_0 \times \mathbf{v}(t) = (1-t)\mathbf{v}_0 \times \mathbf{v}_0 + t\mathbf{v}_0 \times \mathbf{v}_1 = t\mathbf{v}_0 \times \mathbf{v}_1. \tag{6.13}$$

Thus, for $t \in (0,1]$, rotation angles are in the same direction as the smaller angle between $\mathbf{v}_0$ and $\mathbf{v}_1$, i.e., the angular range covered by two linearly interpolated vectors is the same as the angle between the start- and end vector of the interpolation.

**Piecewise Constant Fields.** For piecewise constant tangent vector fields of a curved triangulated domain, we compute the Poincaré index for the nodes of the triangulation by considering vectors and triangles of the node's oriented star. The star of a node consists of all triangles with the node being one of the triangle's vertices, ordered in counter-clockwise orientation w.r.t.

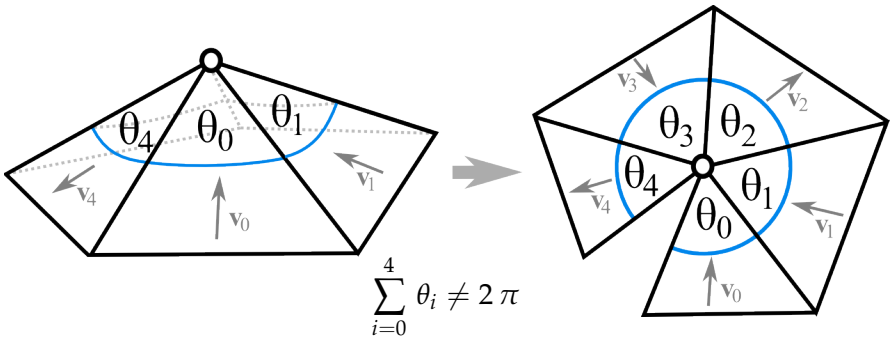



**Figure 6.14:** *Star of a curved surface, incident angles around the center sum to less than $2\pi$. Vector angles are measured in flattened space.*

the triangle normals. The angle of adjacent tangent vectors are measured in tangent space, i.e., triangles and vectors are first transformed in common flat space by unfolding the triangles along their common edge. For computing the Poincaré index in curved surface domains, the Gaussian curvature of the geometry needs to be considered. The Poincaré index is then

$$\mathrm{idx}(c, \mathbf{v}) = \frac{\sum_{i=0}^{K-1} \measuredangle(\mathbf{v}_i, \mathbf{v}_{(i+1)\%K})}{\sum_{i=0}^{K-1} \theta_i}, \tag{6.14}$$

for a node with an oriented star of $K$ triangles with incident angles $\theta_i$ and tangent vectors $v_i$. The index is integer valued and denotes the number of oriented windings of the vectors around the center. Technically, the index is rounded to the nearest integer to account for floating point issues.

**Continuous Fields.** The discrete formula is correct as well if vectors are interpolated linearly along the edges of the grid. This holds especially for triangular grids and bilinear rectangular grids, the most common cases. For tangent vectors in 2D flat space given on the nodes of a grid, the denominator of Eq. (6.14) is $2\pi$, the index is computed for the 2-cells; vectors of the nodes are traversed in counter-clockwise order.

**Classification.** The sign of the Poincaré index in 2D allows to discriminate between source/sink/center types of critical points (index $> 0$) and saddle type critical points (index $< 0$). To further distinguish between sources and sinks, we compute the divergence of the vector field. According to Gauss' theorem the total divergence of a volume element can be computed by considering the flux through a closed surface. For piecewise constant tangent vector fields, Polthier and Preuß [PP02] defined the divergence operator for

a vertex $c$ by

$$\mathrm{div}(c, \mathbf{v}) = \frac{1}{2} \int_{\partial \mathrm{star}(c)} < \mathbf{v}_i, \mathbf{n}_i >, \tag{6.15}$$

the sum over the triangles of c's star, and $\mathbf{n}_i$ the exterior normal for each triangle along the star. For interpolated 2D fields, the divergence is computed as the sum of fluxes through the edges. The indicator functions are then

$$I_{\mathrm{source}}(c, \mathbf{v}) = \begin{cases} 1 & \mathrm{idx}(c, \mathbf{v}) > 0 \ \wedge \ \mathrm{div}(c, \mathbf{v}) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{6.16}$$

$$I_{\mathrm{sink}}(c, \mathbf{v}) = \begin{cases} 1 & \mathrm{idx}(c, \mathbf{v}) > 0 \ \wedge \ \mathrm{div}(c, \mathbf{v}) < 0 \\ 0 & \text{otherwise} \end{cases} \tag{6.17}$$

$$I_{\mathrm{saddle}}(c, \mathbf{v}) = \begin{cases} 1 & \mathrm{idx}(c, \mathbf{v}) < 0 \\ 0 & \text{otherwise} \end{cases} . \tag{6.18}$$

Center type critical points with index 0 and divergence $= 0$ are not handled here. When dealing with numerical data, a divergence of exactly 0 does practically not occur. In the case of incompressible fluids that are known to be divergence-free center identification is performed by $I_{\mathrm{center}} = I_{\mathrm{source}} + I_{\mathrm{sink}}$.

### 6.3.3 Critical Points in 3D

Analogously to the 2D case, the Poincaré index in 3D is given by the sum of oriented solid angles of the vectors of the volume's faces [GTS04]. For triangular faces of a tetrahedron with linearly interpolated vectors defined on the tetrahedron nodes, the vectors of a face span a spherical triangle with solid angle in $[-2\pi, 2\pi)$. The solid angles of all 4 faces divided by $4\pi$ is the Poincaré index in $\{-1, 0, +1\}$. Index $-1$ identifies sinks and saddles with an one dimensional stable manifold, index $+1$ identifies sources and saddles with an one dimensional unstable manifold. This leads to the indicator functions

$$I_{+}(c, \mathbf{v}) = \begin{cases} 1 & \mathrm{idx}(c, \mathbf{v}) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{6.19}$$

$$I_{-}(c, \mathbf{v}) = \begin{cases} 1 & \mathrm{idx}(c, \mathbf{v}) < 0 \\ 0 & \text{otherwise} \end{cases} . \tag{6.20}$$

### 6.3.4 Swirling Motion

Swirling motion core-lines in 3D vector fields, as defined by Sujudi and Haimes [SH95], are lines where (i) the Jacobian $J$ of the vector field has 2

complex eigenvalues, and (ii) the real eigenvector is parallel to the vectors along the lines. In a tetrahedral grid with vectors given on the grid nodes, the Jacobian is constant and vectors interpolate linearly within each tetrahedron. Thus, swirling motion cores are straight lines within a tetrahedron. We thus define our swirling motion feature indicator $I_{\text{swirl}}$ on the vectors of a tetrahedron by 1 if a swirling motion core passes through the tetrahedron and 0 otherwise.

### 6.3.5 Computation of Feature Probabilities

Let $\mathbf{V}_c$ be a random vector in which all components of the vectors of the neighborhood of $c$ are combined and which is described by a local PDF $f_c$ which represents the the $(\ell K_c)$-variate distribution, where $\ell$ is the dimensionality of the vectors and $K_c$ is the number of degrees of freedom. The PDF can be a Gaussian or any nonparametric distribution introduced in Sect. 3.5. Then Eq. (6.2) can be adapted to conform to the vector notation leading to

$$\mathrm{P}_c = \int_D f_c(\mathbf{v}_c)\,\mathrm{d}\mathbf{v}_c = \int_{\mathbb{R}^{\ell K_c}} f_c(\mathbf{v}_c)\,I(c,\mathbf{v}_c)\,\mathrm{d}\mathbf{v}_c = \mathrm{E}(I(c,\cdot)), \tag{6.21}$$

where $D = \{\mathbf{v} \in \mathbb{R}^{\ell K_c} \mid I(c,\mathbf{v}) = 1\}$ for feature indicator $I$. Depending on the feature and the type of grid $\mathrm{P}_c$ can be defined for 0-cells (e.g. critical points at vertices in piecewise constant fields) or higher dimensional cells (such as tetrahedra for which $I_{\text{swirl}}$ is evaluated).

### 6.3.6 Visual Mapping

We employ two different methods to display the probabilities computed in 2D domains. The first method uses a heightmap with an additional colormapping. The second method uses an additive blending of distinct colors for source, sink and saddle probabilities. The first method is superior for flat domains, especially if they contain overlapping spatial distributions. The second method makes it possible to depict multiple probability fields simultaneously in a single visualization, and works well for fields with peaked and rather sparse spatial distributions. For 3D data we show colored nested transparent isosurfaces of the probability fields that indicate the spatial distribution of the respective features. To provide context and give a basic impression of the uncertain vector field's trends we display LIC visualizations of the mean field $\mu$ in 2D fields and streamlines for 3D fields, respectively.

### 6.3.7 Results and Discussion

We applied the probabilistic feature extraction methods to datasets from climate simulations and biofluid mechanics. Additionally, to illustrate basic

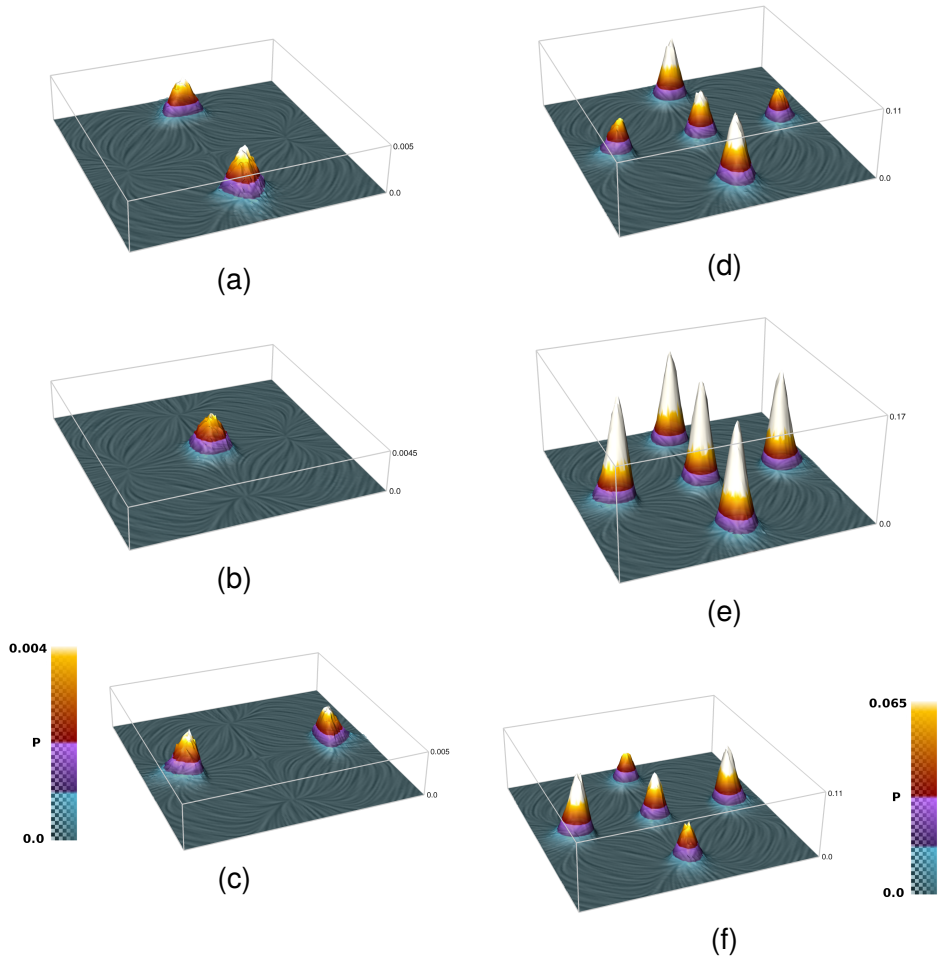



**Figure 6.15:** *Synthetic dataset. From left to right: sources, saddles and sinks; top: correlation considered, bottom: correlation neglected. Note that the ranges of probabilities differ between the correlated and uncorrelated case and have been scaled for visualization. LIC visualizations display the mean field $\mu$.*

properties of the methods and show the impact of spatial correlation, we present a synthetic example.

**Synthetic Dataset.** First, we applied our method to a dataset based on the formula proposed by Otto et al. [OGHT10]. With

$$v_c(x,y) = \begin{pmatrix} -x(1-x)(1+x)(1-y^2) - xy^2 \\ y(1-y)(1+y)(1-x^2) + yx^2 \end{pmatrix}, \tag{6.22}$$

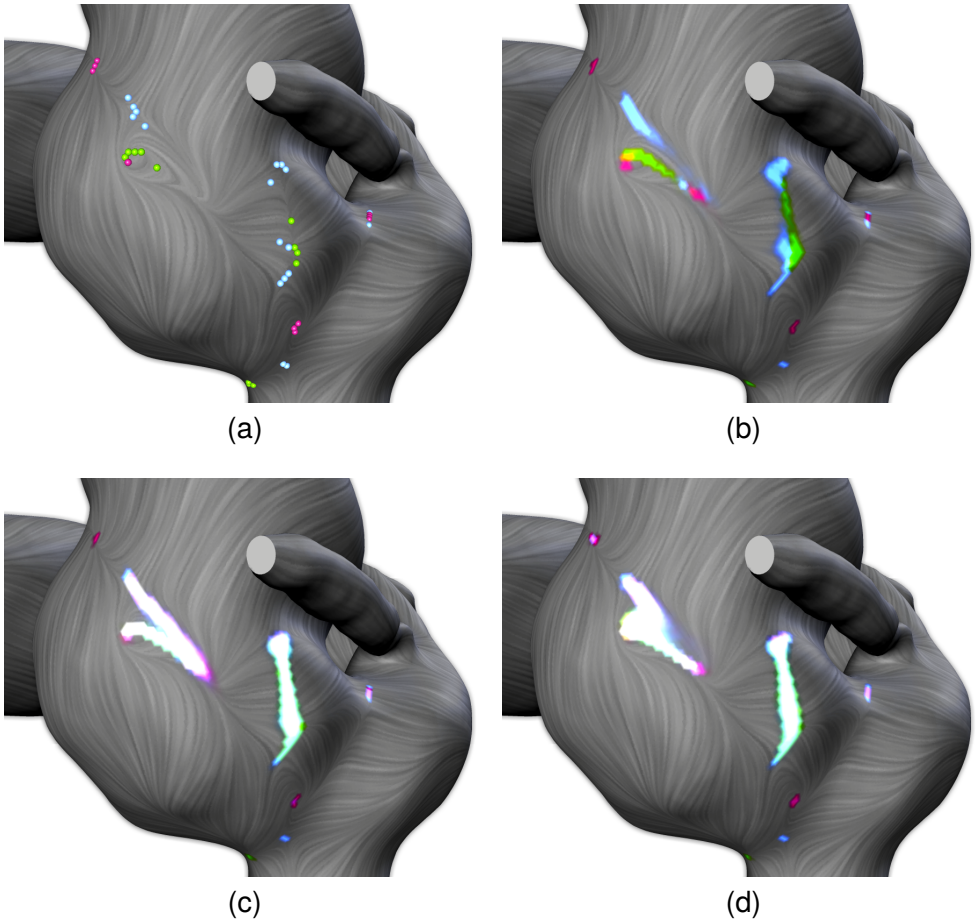



**Figure 6.16:** *Color coded probabilities for singularities in the wall shear stress vector field from a simulated cerebral aneurysm blood flow at a single simulation time step. The mean wall shear stress vector field $\mu$ is indicated by a low-contrast LIC visualization. Probabilities for the different critical point types are encoded by different colors: sinks in violet, sources in green and saddles in blue. Intensities are scaled by the probabilities. Colors are blended additively. Depicted are: All critical points of the 9 ensemble members (a), probabilities considering spatial correlations (b), probabilities with correlations of vector components only (c) and probabilities with correlations neglected (d).*

we created an ensemble dataset with $L = 32$ members and $r = 0.2$ by

$$v_i(x, y) = v_c(x + r\cos\phi_i, y + r\sin\phi_i), \tag{6.23}$$

with $i \in \{1, \dots L\}$, $\phi_i = \frac{2\pi i}{L}$, sampled on rectangular grids in $[-1, 1]^2$ with $128^2$ samples. We computed the sample mean and sample covariance from

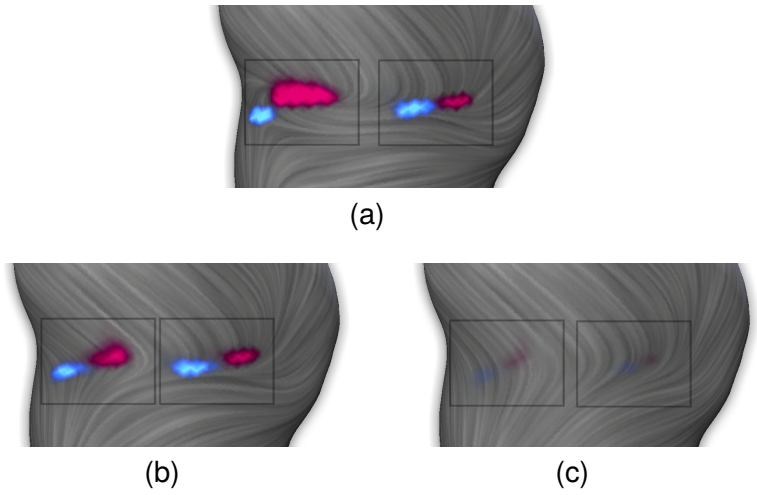



**Figure 6.17:** *Critical point probabilities of the aneurysm wall-shear stress vector field at three subsequent time steps.*

this ensemble. In Fig. 6.15 (a),(b) and (c) the resulting probabilities are presented. Probabilities for the singularities are very close to the singularity distribution of the ensemble dataset. For the results in (d),(e) and (f) the correlations were neglected (the non-diagonal covariances are set to zero). Consequences of that are misclassifications of critical points and overestimation of probabilities. The expected values for the total number of critical points (equal to the sum of cell-wise probabilities) in the whole domain are:

| correlation | $E(\#source)$ | $E(\#saddle)$ | $E(\#sink)$ |
|---|---|---|---|
| considered | 1.9 | 1.0 | 1.9 |
| neglected | 57.3 | 111.6 | 57.4 |

Thus, by considering spatial correlations these numbers reproduce very closely the numbers of critical points in every ensemble member. Probabilities are significantly overestimated in the uncorrelated case.

**Blood Flow Fields from Hemodynamics Simulation Data.** We inspected uncertain features of the wall-shear stress vector field and the blood flow velocity field in a cerebral aneurysm, resulting from time-dependent biofluid mechanical simulations. Aneurysm geometry was reconstructed from an individual patient. Modelling parameters are affected by uncertainty: patient-specific flow-rates could not be measured in clinical practice and are taken from a textbook; the hematocrit value of the blood changes over time. To inspect the impact of these uncertainties on the positions of flow singularities, we studied an uncertain vector field defined by an ensemble of simulation

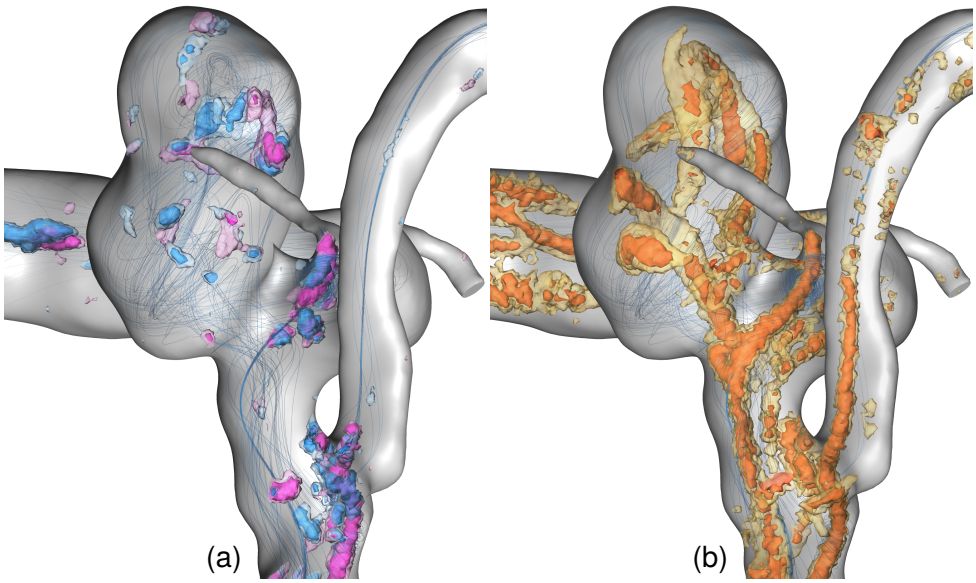



**Figure 6.18:** *Uncertain flow features over a full heart cycle in a cerebral aneurysm visualized by nested semi-transparent isosurfaces. Streamlines of the mean vector field reveals some context. (a) Critical point probabilities with Poincaré index $> 0$ (blue) and $< 0$ (violet). (b) Probabilities for swirling motion cores.*

results obtained with 9 different parameter configurations. From the ensemble we estimated the sample mean vectors $\widehat{\mu}$ and sample covariance matrices $\widehat{\Sigma}$.

Fig. 6.16 depicts probabilities for source, sink and saddle type critical points at a single time step of the simulation. Fig. 6.16 (a) depicts the critical points of all 9 ensemble members by colored spheres. The points are computed for the nodes of the triangulated surface, multiple occurrences of critical points at the same nodes are possible. Critical points of different ensemble members are close-by. In (b), critical point probabilities of the uncertain vector field with $\widehat{\mu}$ and $\widehat{\Sigma}$ are depicted. The similarity to (a) is high but non-vanishing probabilities also occur in other areas of low vector magnitude where the amount of uncertainty exceeds the mean. We studied the influence of the covariances by assuming statistical independence of the vectors among one another in (c), and of all components of the random vector in (d), i.e., by dropping all the covariances. In both cases, the distinctive power for the type of critical point diminishes, indicated by the white color, resulting from an additive blending of colors associated to high probabilities for all critical point types.

Three subsequent time steps at $t = 0.57s$, $0.61s$ and $0.63s$ at a heart cycle of $1s$ are displayed in Fig. 6.17. During time, the probabilities for critical

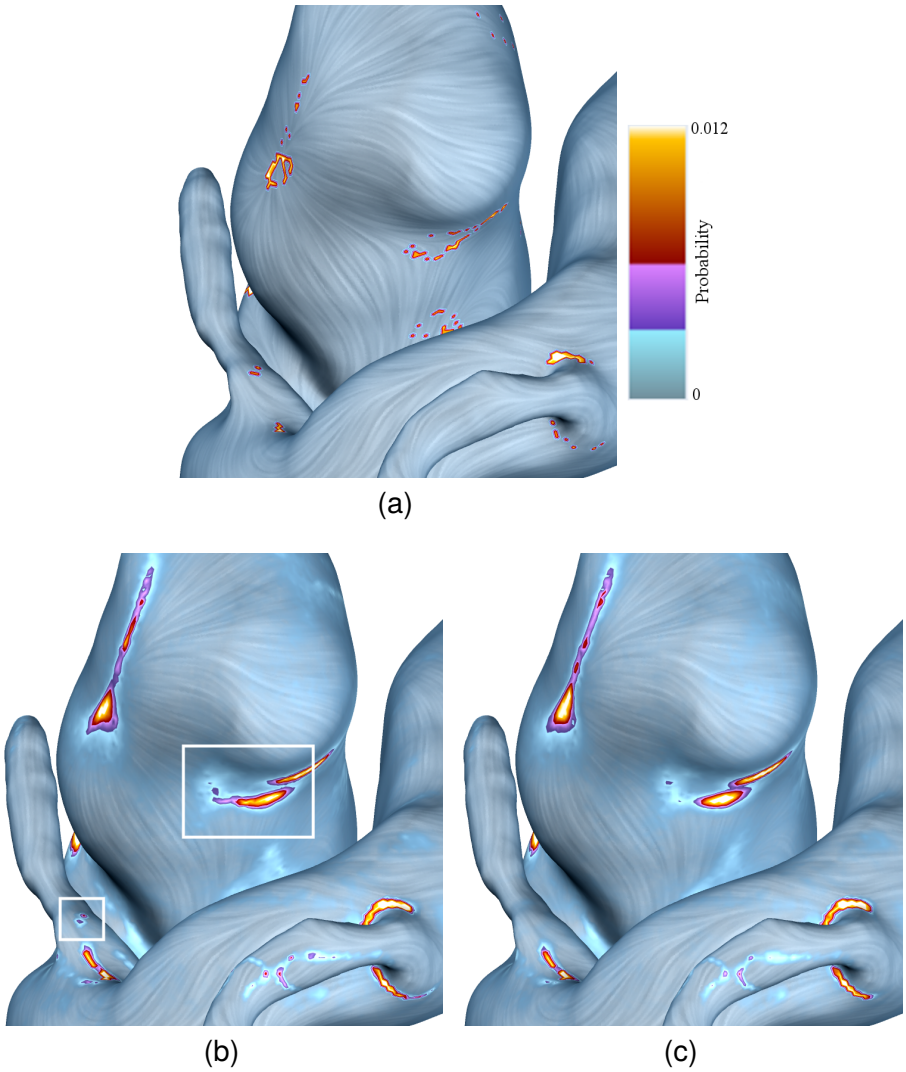



**Figure 6.19:** *Probabilities for the existence of critical points in the wall shear stress field on the vessel and aneurysm wall employing (a) empirical distributions, (b) KDE and (c) parametric Gaussian models. The mean vector field is shown using LIC.*

points of sinks and saddles disappear pair-wise. The following table lists the expected values for the number of saddles and sinks on the left ($E_l$) and right ($E_r$), for the time steps of Fig. 6.17. The areas over which the expectations were computed are indicated by rectangles.

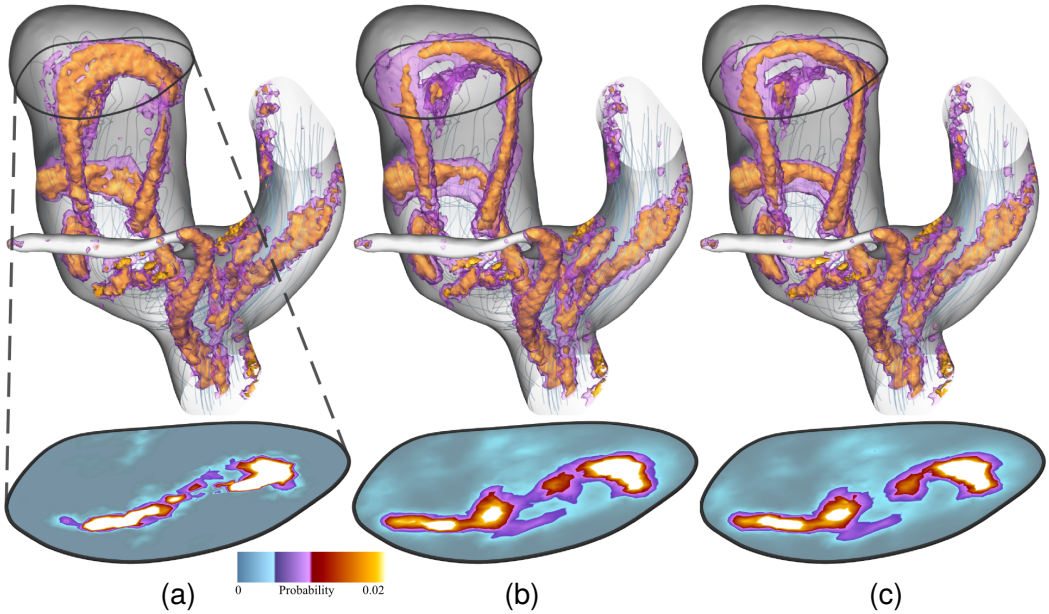



**Figure 6.20:** *Probabilities for the existence of vortex cores in the blood flow velocity field are displayed using nested isosurfaces (top row). 2D slices through the probability fields in the dome region are shown in the second row. The probabilities were computed using (a) empirical distributions, (b) KDE and (c) a parametric Gaussian model.*

| $t/s$ | $E_l$(#saddles) | $E_l$(#sinks) | $E_r$(#saddles) | $E_r$(#sinks) |
|---|---|---|---|---|
| 0.57 | 0.99 | 0.99 | 0.26 | 0.26 |
| 0.61 | 0.23 | 0.22 | 0.2 | 0.19 |
| 0.63 | 0.03 | 0.03 | 0.01 | 0.01 |

Results in 3D are presented in Fig. 6.18. We computed means and covariances for a single parameter setting, but for all time-steps of a heart cycle simulation. Pointwise statistical analysis of physical quantities over heart cycles is common practice in the application domain, see e.g. [BRM*08, GSK*12]. A Gaussian model can be a too rough approximation for the velocities of the pumping flow. Thus, more flexible probabilistic models are beneficial, see below for comparisons between different models. In (a) the critical points are depicted. As blood is an incompressible fluid, all critical points are of saddle-type. Critical point probabilities are more focused at vessel bifurcations and more fuzzy in the dome region of the aneurysm. Critical point type distinction with Poincaré index $> 0$ (1D unstable manifold) and $< 0$ (1D stable manifold) is observable as well. Probabilities for Sujudi-Haimes swirling motion cores are depicted in (b). The time-variation of vortical structures is higher in the dome-region of the aneurysm in comparison to the vortical flow in the vessels.

To inspect the variability of the results with respect to probabilistic modelling we represent the variability of the vectors over one cardiac cycle using 3 different models. Fig. 6.19 shows probabilities for the existence of critical points in the WSS field on the vessel and aneurysm wall. In Fig. 6.19 we can see that some critical points which are present in the results in (a) and (b), but not present in (c). Here, the Gaussian model fails to represent the structure of the vector field correctly. In turn some of the features that are visible in (b) and (c) are not present in (a) because critical points are missed due to the sparse discrete sampling of the state space. Probabilities for the existence of vortex cores in the blood flow velocity field are displayed using nested isosurfaces in the top row of Fig. 6.20. 2D slices through the probability fields in the dome region are shown in the second row. There are big differences between the empirical distributions and both the KDE and the Gaussian model. In (a) the spatial distributions are more compact and the probabilities are higher while in (b) and (c) the spatial distributions are wider (less peaked). Note that the absolute probability values are rather low because of strong spatial correlation. Roughly speaking, strong correlations lead to smooth realizations of the vector field with few critical points and vortex cores whereas weak correlations result in more chaotic realizations with a higher relative number of features.

**Wind Velocity Fields from Climate Simulation Data.** We analyzed ensemble datasets from the DEMETER project [Pal04] to obtain distributions for singularities in the 10-meter-wind velocity vector fields. For each time step similar to the temperature fields, 63 realizations of daily average wind velocities constitute an ensemble, where the results are generated by 7 different climate models and 9 different sets of simulation parameters. From these results we compute the means $\mu$ and covariances $\Sigma$ for each grid cell of the rectilinear 2D grid.

In Fig. 6.21 the probabilities for singularities in the uncertain 10-meter wind vector field are shown as a heightfield with colormapping. The mean vector field $\mu$ is indicated as a LIC visualization below the height field. In Fig. 6.21 (a) the probabilities for the existence of sources, (b) saddles and (c) for sinks are shown. Source- and sink probabilities are mainly high over the landmass, especially mountains attract this behavior. This is reasonable, as source and sink behavior in a 2D slice denote lifting and falling wind. An overview of saddle probabilities for all time steps is included in the movie in the supplementary material.

Fig. 6.22 shows the influence of spatial covariances. Displayed are cutouts of sink probabilities in Middle America. In (a), spatial covariances are considered, in (b) only vector-wise covariances are considered, and in (c) no covariances are considered. Including spatial covariances drastically

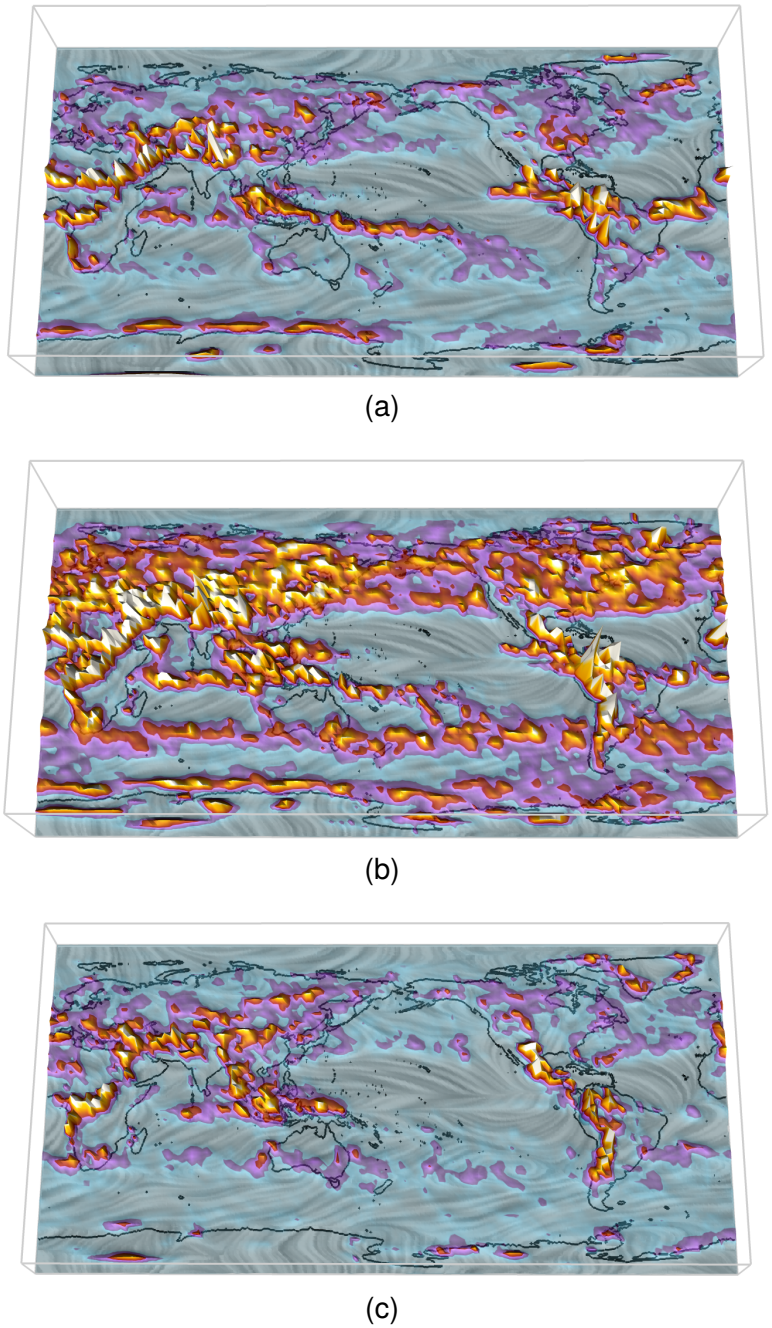
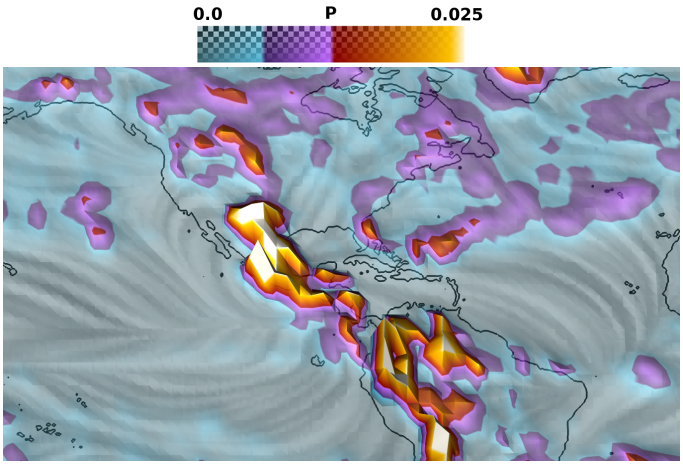
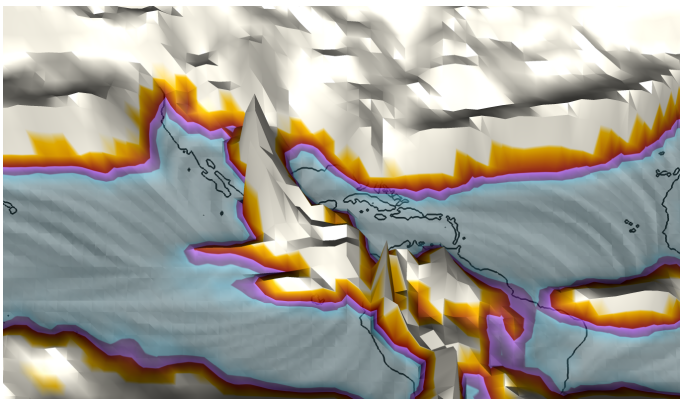


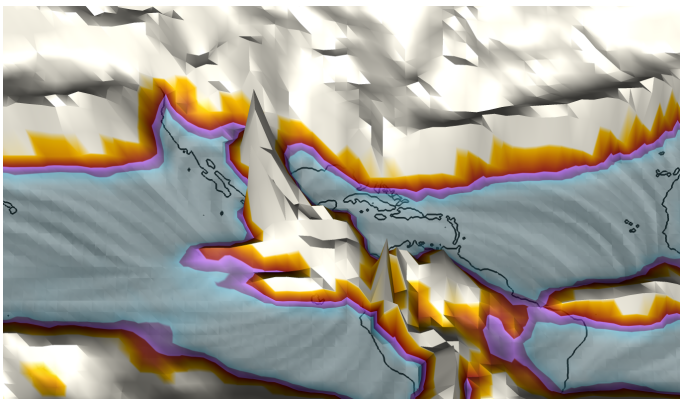

**Figure 6.21:** *Probabilities for singularities in the daily average wind vector field from the climate simulation dataset are shown as a heightfield with colormapping. (a) sources, (b) saddles, (c) sinks.*




(a)



(b)



(c)

**Figure 6.22:** *Cutout of the probability fields for sinks: (a) With correlations, (b) just vector-wise correlations, (c) without consideration of correlations.*

reduces the probabilities; the influence of the vector-wise covariances of (b) is practically not distinguishable from (c).

**Discussion.** The results show that our approach for the computation of probability fields for local features considering local covariance structures works very well in practice. As depicted in Fig. 6.16a and 6.16b, singularities in the aneurysm wall-shear stress vector field of the ensemble members agree very well with the computed probability fields. Ensemble critical points are clustered, the probability fields are localized as well. Following critical point probability fields over time as done in Fig. 6.17 confirms the fact of crisp singularities that critical points with index $+1$ and $-1$ appear or disappear pairwise. Probabilities for sinks and saddles fade out simultaneously with roughly the same amount. Recent studies have shown that low wall-shear stress in the dome region is connected to increased rupture risk of certain cerebral aneurysms [GSK*12]. The time evolution of probabilities of critical points, particularly their spatial stability and variability, as shown in the supplementary movie, provides information that domain experts considered as important.

Our approach allows to distinguish between different critical point types. This classification proved to produce meaningful results. The spatial covariances capture spatial derivatives very well. In the synthetic dataset of Fig. 6.15, practically no overlapping classifications are found. In the wall shear stress field of the aneurysm dataset, regions with high overlapping probabilities for different critical point types are rare and differentiation is precise, even if the different critical point probabilities are very close-by. Some more overlap is observable in 3D, but the different regions are still fairly separated. Singularities in the climate wind velocity dataset are much less separated. Notable differences between source, saddle and sink probabilities are also observable in Fig. 6.21.

We applied the Sujudi and Haimes swirling motion vortex core criterion on piecewise linear vector fields without filtering or strength assessment. In a crisp vector field, this leads to discontinuous line segments with a lot of clutter due to weak swirling behavior. Nevertheless, the resulting probabilities in the aneurysm dataset in Fig. 6.18 correspond very much to the main vortices. This is due to an implicit smoothing effect in the uncertainty setting, where weak features are spatially less concentrated and have lower probabilities. The chosen vortex criterion is more direct, compared to other methods using vortex indicating scalar fields, like e.g. $\lambda_2$ [JKJTM06]. The scalar fields would require additional processing steps (thresholding or ridge extraction) to yield Boolean feature detectors.

The proposed feature indicators are defined on the smallest possible scales represented by the sampling grid. The interpretation of the results

must account for this. In case features on larger scales are of interest, indicator function domain sizes would need to be adjusted accordingly.

Computation times differ depending on the type of model and the data set. Empirical distributions are the simplest model for which the computation of feature probabilities is the computationally least expensive. The calculations of the results above took only a few seconds on an Intel Xeon X5550. For the models that create continuous PDFs, the computational complexity is much higher, because Monte Carlo integration has to be performed. We chose the number of samples manually for each dataset, such that no Monte Carlo noise was observable anymore. The computation of 2D results took several minutes, the 3D results several hours. For the same number of samples, the sampling of KDE distributions takes slightly more time compared to the parametric Gaussian model due to the PC transformation and an additional random number that has to be computed for each sample to select the KDE component. An advantage of our local approach is that only local covariances are needed and thus the memory requirements are not quadratic in the number of cells as it would be if the complete correlation structure was stored.

Our method for critical point detection is closely related to the method proposed by Otto et al. [OGHT10, OGT11b, OGT11a]. Both methods compute scalar fields to indicate relative critical point strengths. We compute cell-wise probabilities, Otto et al. density fields that are normalized to an integral of 1. The approaches differ significantly. We consider critical points as local features, whereas Otto et al. consider them as global features. Thus, results are not directly comparable. Saddle detection ability is intrinsic in the local approach, more algorithmic effort (computation of the gradient of the squared velocity of the uncertain vector field) is needed in the global approach. The global approach seems to be insensitive to neglecting correlations; this deserves further research. The implementation complexity differs; we think that our local method is conceptually easier and easier to implement. However, our approach is limited to local features. Similar to crisp vector field analysis, features that are global by nature such as (closed) streamlines or separatrices are detectable with global methods only.

## 6.4 Fast Approximation Methods

A disadvantage of the estimation method presented above is the high computational cost of the Monte Carlo (MC) integration. In the following we introduce several approximation methods to overcome this drawback. In addition to two specific approaches for cell-wise level-crossing probabilities in discretized Gaussian fields we propose a flexible approximation method based on surrogate functions. The surrogate functions are estimated from

example grid cells and their feature probabilities (the training set) and can predict probabilities for new grid cells and datasets. We provide a quantitative and qualitative evaluation of the approximation errors and show that the results computed using surrogate functions converge to the ground truth for increasing sizes of the training sets.

### 6.4.1 Approximate Crossing Probabilities Based on Bivariate Distribution Functions

For the specific case of level-crossing probabilities in Gaussian fields with exponential correlation functions, Pfaffelmoser et al. [PRW11] presented a raycasting approach that computes first-crossing probabilities along rays using lookup-tables for fast evaluation. The results of this method depend on the viewing direction.

Our aim is to improve the computation of local cell-wise level-crossing probabilities considering *arbitrary* spatial correlations independently of the viewing direction, i.e. they are objective, in the sense that they are independent of the viewing and rendering conditions. Since the input data is usually given on some grid, it is a natural choice to consider grid cells and to compute cell-related probabilities. The numerical computation of high-dimensional integrals in general is expensive, both with deterministic and MC methods. There are two ways to deal with this problem: either utilize specific properties of the problem to facilitate the computation, or find good and fast approximations of the integrals. Here we consider the latter approach: we compute approximate univariate and bivariate distribution functions that can be evaluated in the rendering step using table lookups.

We will consider two possibilities for approximating the probabilities. The *maximum edge crossing method* considers pairwise correlations between two random variables at a time. The *linked-pairs method* iteratively traverses the vertices of a grid cell and considers joint and conditional probabilities between subsequent vertices. This algorithm induces an $n$-dimensional approximate distribution, where $n$ is the number of vertices. Depending on the order in which the vertices are traversed, different approximate probability distributions occur; an optimal distribution is selected by optimizing the Bhattacharyya distance to the original distribution.

#### 6.4.1.1 Standardization of the Bivariate Probability Integral

We define the events $Y_i^+ = (Y_i > \vartheta)$ and $Y_i^- = (Y_i \leq \vartheta)$. The edge-level-crossing and non-crossing probabilities given in Sect. 6.2.2 can also be expressed by

$$\mathrm{P}_c(\vartheta\text{-crossing}) = \mathrm{P}(Y_1^- \cap Y_2^+) \; + \; \mathrm{P}(Y_1^+ \cap Y_2^-)$$

and

$$\mathrm{P}_c(\vartheta\text{-non-crossing}) = \mathrm{P}(Y_1^- \cap Y_2^-) \; + \; \mathrm{P}(Y_1^+ \cap Y_2^+) \tag{6.24}$$

These results depend on the parameters $\mu_1$, $\mu_2$, $\mathrm{Cov}_{1,1}$, $\mathrm{Cov}_{2,2}$, $\mathrm{Cov}_{1,2}$ and $\vartheta$. The integral in Eq. (6.6) can be expressed in terms of the *standard normal* cumulative distribution function $\Phi(y_1, y_2, \rho)$, with *correlation coefficient* $\rho = \frac{\mathrm{Cov}_{1,2}}{\sigma_1 \sigma_2}$, standard deviation $\sigma_i = \sqrt{\mathrm{Cov}_{i,i}}$ and integration bounds given by the *stochastic distance function*

$$\Psi_i = \frac{\mu_i - \vartheta}{\sigma_i},$$

such that

$$\begin{aligned} \mathrm{P}_c(\vartheta\text{-crossing}) &= 1 - \big(\mathrm{P}(Y_1 \leq \vartheta, Y_2 \leq \vartheta) + \mathrm{P}(Y_1 > \vartheta, Y_2 > \vartheta)\big) \\ &= 1 - \big(\Phi(-\Psi_1, -\Psi_2, \rho) + \Phi(\Psi_1, \Psi_2, \rho)\big). \end{aligned} \tag{6.25}$$

This is a very convenient formulation because $\Phi(y_1, y_2, \rho)$ can be efficiently evaluated using a 3D lookup table [PRW11].

In general, integrals for distributions with $n > 2$ dimensions

$$\begin{aligned} \mathrm{P}_c(\vartheta\text{-crossing}) &= 1 - \big(\mathrm{P}(Y_1^- \cap Y_2^- \ldots \cap Y_n^-) + \\ &\qquad \mathrm{P}(Y_1^+ \cap Y_2^+ \ldots \cap Y_n^+)\big) \end{aligned} \tag{6.26}$$

can not be evaluated in closed form or using lookup tables (due to quickly increasing memory requirements). Numerical integration schemes, e.g. Monte Carlo methods, can be used for estimation.

#### 6.4.1.2 Vertex- and Edge-Based Approximations

To facilitate fast interactive visualization expensive numerical integration must be avoided. In addition to the trivial approach that simply neglects the correlation structure we propose two *approximations* for level-crossing probabilities that can be evaluated very efficiently, but consider correlations.

**Statistically Independent Vertices.** The first, highly simplified approach completely neglects the correlation structure and computes probabilities under the assumption that all random variables are statistically independent. The level-crossing probability for cell $c$ is then

$$\mathrm{Q}_c = 1 - \big(\mathrm{P}(Y_1^+)\mathrm{P}(Y_2^+)\ldots\mathrm{P}(Y_n^+) + \mathrm{P}(Y_1^-)\mathrm{P}(Y_2^-)\ldots\mathrm{P}(Y_n^-)\big). \tag{6.27}$$

However, this way the spatial distribution of uncertain isocontours is often overestimated [PRW11, PWH11].

**Maximum Edge Crossing Probability.** The second measure to approximate the cell-level-crossing probability is the maximum edge-level-crossing probability over all edges. Taking spatial correlations into account, the smallest grid entity to consider is an edge that connects any two points $(Y_1, Y_2)$ of a cell. If a cell with $n$ vertices and $m$ edges contains a level-crossing, at least one of its edges contains a level-crossings as well. The converse is obviously true as well: as soon as a level-crossing occurs between any two vertices, the cell has a level-crossing. Thus, the edge-wise level-crossing probability is a lower bound for the cell integral. As an approximation for the cell-wise level-crossing probability, we use the maximum lower bound, e.g.,

$$\mathrm{R}_c = \max_{i=1\ldots m} \left(1 - \left(\mathrm{P}(Y_{i,1}^+ \cap Y_{i,2}^+) + (\mathrm{P}(Y_{i,1}^- \cap Y_{i,2}^-)\right)\right), \tag{6.28}$$

where $Y_{i,1}$ and $Y_{i,2}$ are the random variables associated with the vertices that are connected by edge $i$. In other words, we reduce the $n$-dimensional distribution to 2D marginal distributions to find the edge with maximum level-crossing probability.

To get an intuition why the edge-wise level-crossing probability is indeed a lower bound consider the example of a single triangular cell with independent Gaussian distributions at the vertices $Y_{1,2,3} \sim \mathcal{N}(0,1)$. For the isovalue $\vartheta = 0$, the maximum edge crossing probability is $\mathrm{R}_c = 0.5$. In contrast, the cell-wise crossing probability is $\mathrm{P}_c = 1 - (\mathrm{P}(Y_1^+ \cap Y_2^+ \cap Y_3^+) + \mathrm{P}(Y_1^- \cap Y_2^- \cap Y_3^-)) = 0.75$. Generally, for cells with $n > 2$ vertices the cell-wise crossing probability is larger or equal to the maximum edge-wise probability because a crossing may also occur on other edges than the one corresponding to the maximum crossing probability.

#### 6.4.1.3 The Linked-Pairs Approximation

For the third approximation, more correlations are considered. Both 2D joint and conditional probabilities for level-crossings between any two variables of a cell can be evaluated using lookup tables. To exploit that, pairwise conditional probabilities are evaluated in a step by step fashion from vertex to vertex of a cell, see Fig. 6.23. We s4how that this method induces an approximate distribution that is again normally distributed. The approach has a degree of freedom in the choice of the traversal order of the vertices $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ that can be described by a spanning tree. The Bhattacharyya distance is used to compare the different choices to the original distribution.

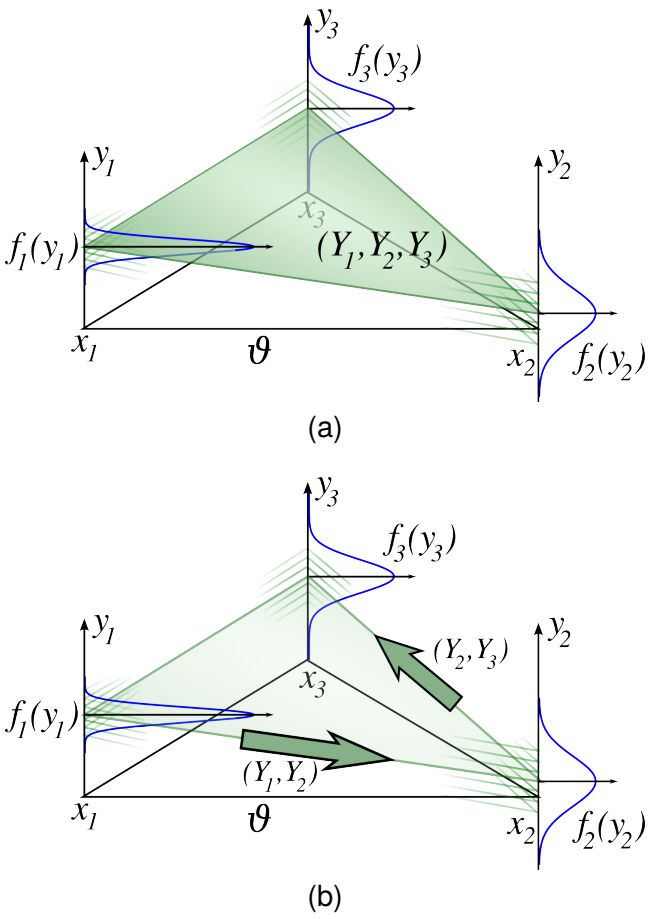



**Figure 6.23:** *Example for the computation of a level-crossing probability in a triangular cell. The marginal distributions at the grid points are shown in blue. Exemplarily, one realization of the interpolant is shown in green. Fig. (1) corresponds to the consideration of the complete covariance matrix in Eq. (6.26) while (2) shows the approximation using pairwise correlations in Eq. (6.31).*

**Approximate Probabilities.** We approximate $\mathrm{P}(Y_1^+ \cap Y_2^+ \ldots \cap Y_n^+)$ by

$$\tilde{\mathrm{P}}(Y_1^+, Y_2^+, \ldots, Y_n^+) \;:=\; \mathrm{P}(Y_1^+ \cap Y_2^+)\, \mathrm{P}(Y_3^+ | Y_2^+) \;\ldots\; \mathrm{P}(Y_n^+ | Y_{n-1}^+) \qquad (6.29)$$

with conditional probabilities

$$\mathrm{P}(Y_i^+ | Y_{i-1}^+) \;=\; \frac{\mathrm{P}(Y_{i-1}^+ \cap Y_i^+)}{\mathrm{P}(Y_{i-1}^+)}.$$

The choice of pairwise joint probabilities that need to be evaluated in Eq. (6.29) was chosen arbitrarily, but influences the result. The joint proba-

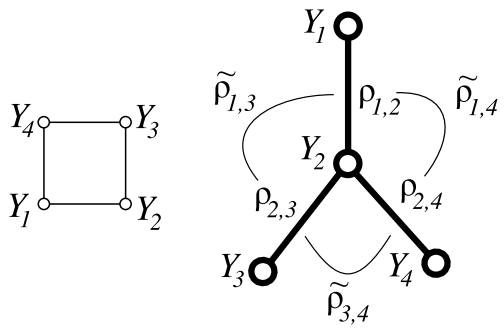



**Figure 6.24:** *One example for a spanning tree of a rectangular grid cell is shown. The choice of pairs of vertices on the edges of the tree determines the pairwise correlations that are taken from the input distribution and used to compute the remaining correlations according to Eq. (6.33). In this particular case* $\tilde{\rho}_{1,3} = \rho_{1,2}\,\rho_{2,3}$, $\tilde{\rho}_{1,4} = \rho_{1,2}\,\rho_{2,4}$ *and* $\tilde{\rho}_{3,4} = \rho_{2,3}\,\rho_{2,4}$.

bilities in each term determine which of the pairwise correlations are considered. A natural choice does not exist. To make the choice of pairs of vertices explicit, we reformulate (6.29) such that the parameter $k \in \{1, 2, \ldots, n^{n-2}\}$ identifies a spanning tree over all vertices of the cell. The approximate probability is given by

$$\tilde{\mathrm{P}}(Y_1^+, Y_2^+, \ldots, Y_n^+; k) = \mathrm{P}(S_1^k \cap S_2^k)\, \frac{\mathrm{P}(S_3^k \cap S_4^k)}{\mathrm{P}(S_3^k)}\, \frac{\mathrm{P}(S_5^k \cap S_6^k)}{\mathrm{P}(S_5^k)} \ldots \tag{6.30}$$

where $\{S^1, \ldots, S^{n^{n-2}}\}$ are the $n^{n-2}$ possible *spanning trees* over $n$ vertices, and each tree is given as an edge list $S^k = \{\{S_1^k, S_2^k\}; \{S_3^k, S_4^k\}; \ldots\}$. A method for the optimal choice of $k$ will be derived below. Analogously we can define $\tilde{\mathrm{P}}(Y_1^-, Y_2^-, \ldots, Y_n^-; k)$. The *approximate level-crossing probability* is given by

$$\tilde{\mathrm{P}}_c = 1 - \big(\, \tilde{\mathrm{P}}(Y_1^+, Y_2^+, \ldots, Y_n^+; k) + \tilde{\mathrm{P}}(Y_1^-, Y_2^-, \ldots, Y_n^-; k) \,\big). \tag{6.31}$$

**Approximate Distribution.** For the evaluation of the approximation $\tilde{\mathrm{P}}_c$ in Eq. (6.30) only the pairwise correlations between random variables as given by the spanning tree $S^k$ are used. This algorithm induces a new joint distribution for all variables. Starting from the original Gaussian random vector $\mathbf{Y}_c \sim \mathcal{N}(\mu, \Sigma)$ of a grid cell $c$, we derive the approximate distribution and show that the approximated distribution is again a multivariate normal distribution

$$\tilde{\mathbf{Y}}_c \sim \mathcal{N}(\mu, \tilde{\Sigma}).$$

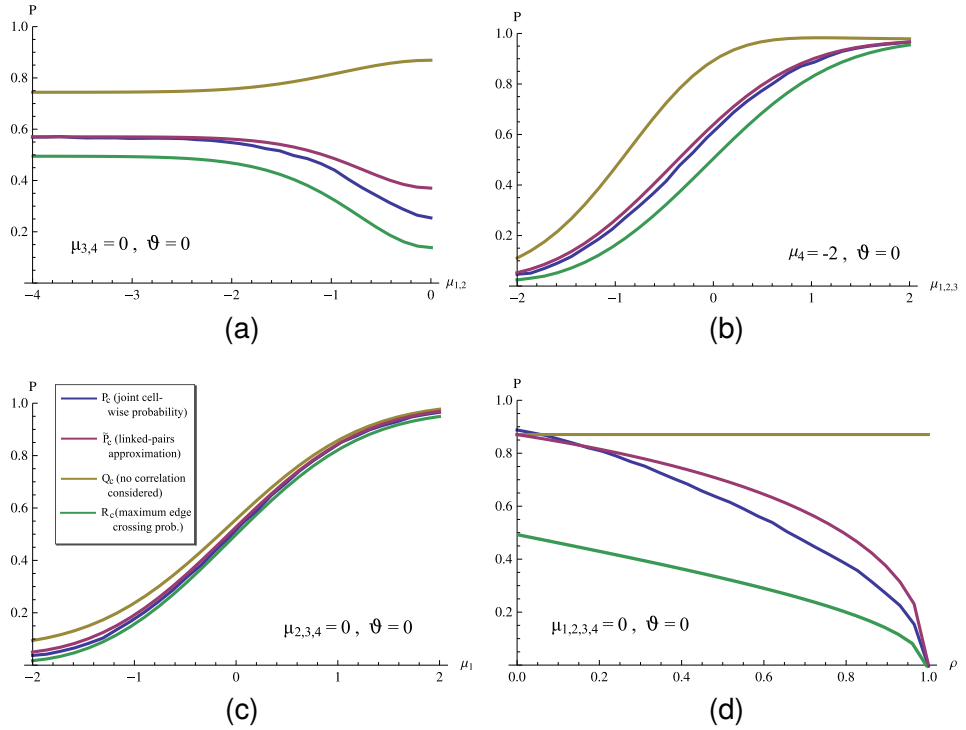



**Figure 6.25:** *(Approximate) crossing probabilities are plotted for a rectangular grid cell and random vector with constant unit variance, varying mean values in (1)-(3) and varying correlation coefficient in (4). The isovalue $\vartheta = 0$ is constant. In (1) $\mu_1 = \mu_2$ vary between $-4$ and $0$ with constant $\mu_3 = \mu_4 = 0$. In (2) $\mu_1 = \mu_2 = \mu_3$ vary between $-2$ and $2$ with $\mu_4 = -2$. In (3) $\mu_1$ varies between $-2$ and $2$ with $\mu_2 = \mu_3 = \mu_4 = -2$. In (4) all $\mu_i = 0$ are constant. In (1)-(3) the correlation coefficient $\rho = 0.9$ is constant. In (4) $\rho$ varies between $0$ and $1$.*

The expected values are identical for $\mathbf{Y}_c$ and $\tilde{\mathbf{Y}}_c$.

The covariance matrix $\tilde{\Sigma}$ that is induced by the approximation is computed as follows: Starting from $Y_1$ we evaluate the correlations of the cell in a step by step fashion. Traversing the spanning tree $S^k$ from a cell vertex $Y_1$ gives an ordered list of edges $\{(i,j)\}$. For each edge $(i,j)$ we extend the distribution iteratively with $\rho_{i,j}$ describing the correlation between $Y_i$ and $Y_j$. Thus, we extend the random vector $\mathbf{Y}_c$ by $Y_i$ or $Y_j$, respectively, depending on which one was not already included in a previous step. According to the derivations in Appendix B the correlation coefficients for this distribution are

$$\tilde{\rho}_{i,j} = \rho_{i,j}, \tag{6.32}$$

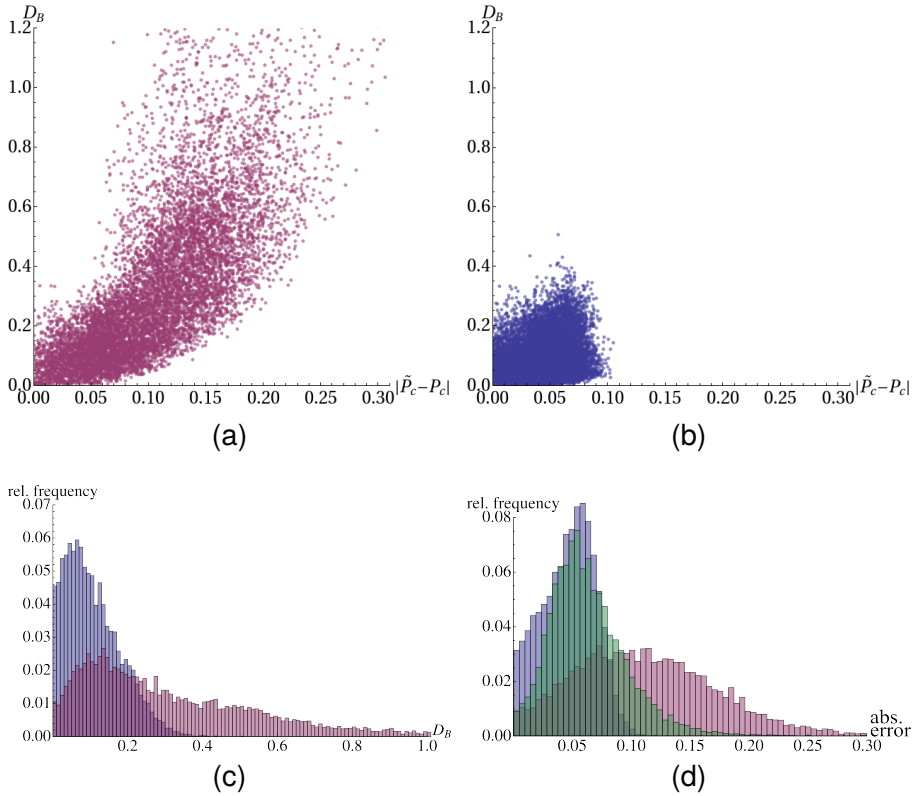



**Figure 6.26:** *Absolute approximation errors and Bhattacharyya distances for square cells with realistic covariance matrices (taken from the climate simulation dataset) and expected values $\mu_i = \vartheta$ are depicted in scatter plots (1),(2) and histograms (3),(4). In (1) the parameter k for Eq. (6.30) was chosen randomly while for (2) the optimal k with minimal $D_B$ was chosen for each cell. In (3) the histograms for $D_B$ with (blue) and without (purple) optimal choice of k is shown. In (4), The linked-pairs approximation error with (blue) and without (purple) optimal choice of k, and the maximum-edge approximation error (green) are shown.*

and

$$\tilde{\rho}_{l,j} = \tilde{\rho}_{l,i}\rho_{i,j}, \tag{6.33}$$

where $l \neq i$. After iterating over all edges we can compute $\tilde{\Sigma}$ from the correlation coefficients $\tilde{\rho}_{i,j}$.

In other words, $\tilde{\rho}_{i,j}$ is the product of the correlation coefficients along the path of the spanning tree connecting the variables $Y_i$ and $Y_j$, see Fig. 6.24 for an example.

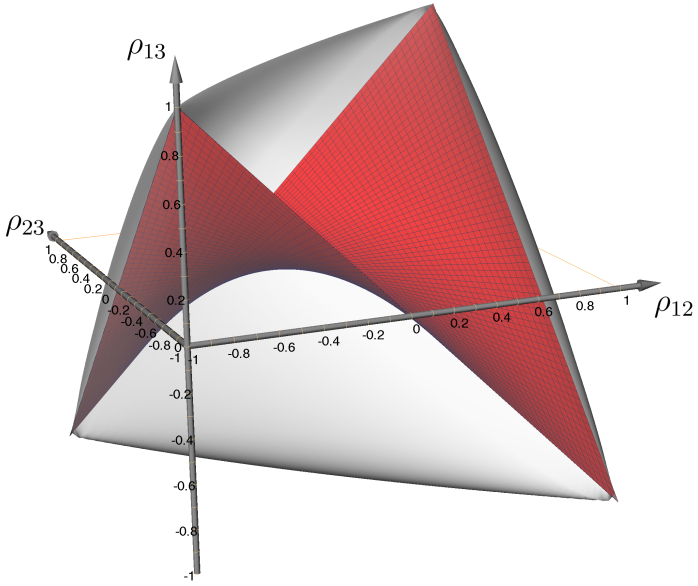



**Figure 6.27:** *Transparent grey surface encloses the space of all possible correlations between* 3 *random variables. Given the correlations between variable* 1 *and* 2 *by* $\rho_{12}$ *and correlation between variable* 2 *and* 3 *by* $\rho_{23}$*, the red surface depicts the computed correlation between variable* 1 *and* 3 *in the linked-pairs approximation.*

**Optimizing the Approximate Distribution.** The linked-pairs crossing probability $\tilde{P}_c$ computed in Eq. (6.31) depends on the choice of a specific spanning tree $k$ for the vertices of $c$. We expect the probability to be close to the true crossing probability if the approximate multivariate distribution is similar to the original distribution. Thus, it is our aim to choose $k$ such that the difference between the original distribution of $\mathbf{Y}_c$ and the approximate distribution of $\tilde{\mathbf{Y}}_c$ is minimal. As measure for the difference between the original and the approximate distribution, we use the Bhattacharyya distance. The Bhattacharyya distance for Gaussian distributions with identical means is given by

$$D_B(k) = \frac{1}{2} \ln \left( \frac{\det((\Sigma + \tilde{\Sigma}_k)/2)}{\sqrt{\det(\Sigma)\ \det(\tilde{\Sigma}_k)}} \right).$$

To obtain the optimal tree we create $\tilde{\Sigma}_k$ for all trees $k$ enumerated by the Prüfer sequence [Prü18], compute $D_B(k)$ and choose $k_{\min}$ such that $D_B$ is minimal, i.e. we solve

$$k_{\min} = \arg\min_k D_B(k), \tag{6.34}$$

and consider $k_{\min}$ in Eq. (6.31). The effect of the optimization is depicted in Fig. 6.26 where the approximation error in relation to $D_B$ is shown.

**Relationship to Graphical Models.** Each spanning tree of a cell can also be interpreted as a *graphical model* that describes the statistical dependencies between the corresponding random variables. More precisely, it is a Bayesian network that contains only the connections between random variables that are present as edges in the spanning tree. The model clarifies that the variables that are *not* connected are *conditionally independent* in the approximate distribution. The restriction to 2D marginal distributions means that each probability can only depend on one other variable.

#### 6.4.1.4 Results and Discussion

**Comparison of the Approximation Methods.** For a quantitative analysis of the approximation we compared the cell-wise level-crossing probability $\mathrm{P}_c$ (Eq. (6.26)) that was numerically estimated using MC sampling (see Sect. 6.1) to the corresponding values of the linked-pairs approximation $\tilde{\mathrm{P}}_c$, the maximum edge crossing probability $R_c$ and $Q_c$ (assuming independent vertices) for simple synthetic datasets.

The probabilities are plotted in Fig. 6.25 for rectangular grid cells and random vectors with constant unit variance, varying mean values in Fig. 6.25 (1)-(3) and varying correlation coefficient in Fig. 6.25 (4). The cell-wise probabilities $P_c$ are drawn in blue, the approximation $\tilde{P}_c$ in magenta, $R_c$ in green and $Q_c$ in yellow. The isovalue $\vartheta = 0$ is constant. In Fig. 6.25 (1) $\mu_1 = \mu_2$ vary between $-4$ and 0 with constant $\mu_3 = \mu_4 = 0$. In Fig. 6.25 (2) $\mu_1 = \mu_2 = \mu_3$ vary between $-2$ and 2 with $\mu_4 = -2$. In Fig. 6.25 (3) $\mu_1$ varies between $-2$ and 2 with $\mu_2 = \mu_3 = \mu_4 = -2$. In Fig. 6.25 (4) all $\mu_i = 0$ are constant. In Fig. 6.25 (1)-(3) the correlation coefficient $\rho = 0.9$ is constant. In Fig. 6.25 (4) $\rho$ varies between 0 and 1.

From the DEMETER 2 meter temperature ensemble we estimated the mean values and covariances for all rectangular grid cells. Fig. 6.28 depicts the correlation structure of the grid cell distributions by displaying the square roots of the eigenvalues of the correlation matrices, i.e., the standard deviations of the distribution in the spaces of their eigenvectors. Values close to 0 denote a flat distribution in the corresponding eigenvector direction, i.e., a high correlation. As depicted, correlations in the dataset are on average very high in at least two eigenvector directions.

In Fig. 6.29 the uncertain isotherm contour for $0^\circ$ $C$ in the temperature field from a climate simulation is displayed. Fig. 6.29 (a) shows the crossing probabilities $\mathrm{P}_c$ for all pixels estimated using a MC computation with 5000 samples. Fig. 6.29 (c) shows the probabilities of the linked-pairs approxi-

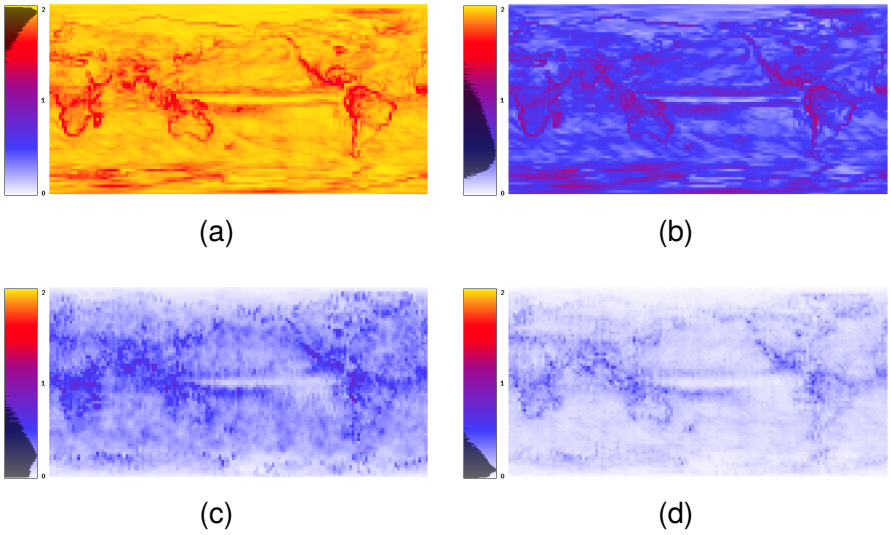



**Figure 6.28:** *Color mapped square roots of the eigenvalues (in decreasing order) of the correlation matrices of the 2D climate dataset. Histograms of the values are displayed on top of the colormap in logarithmic scale.*

mation $\tilde{P}_c$ while the absolute differences, i.e. $|\tilde{P}_c - P_c|$, are depicted in 6.29 (d). Analogously the crossing probabilities $Q_c$ assuming uncorrelated values and the difference image $|Q_c - P_c|$ as well as the maximum edge crossing probabilities $R_c$ and the difference image $|R_c - P_c|$ are displayed. Note that the ranges of the colormaps are individually adjusted for $Q_c$. In Fig. 6.29 (b) the probabilities along the green line indicated in Fig. 6.29 (a) are shown as 1D-graphs.

In Fig. 6.30 the crossing probabilities for a 3D temperature field from the same set of climate simulations are shown. The discretized random field for this example consists of hexahedral grid cells. In Fig. 6.30 (1) the joint cell-wise crossing probabilities $P_c$ estimated by MC sampling, (2) $Q_c$ assuming uncorrelated values, (3) approximate probabilities $\tilde{P}_c$, and (4) maximum edge crossing probabilities $R_c$ are displayed. To allow a quantitative comparison in (5) the probabilities along a straight line in the datasets are shown as 1D-graphs. The single-threaded computation times for the 3D results on an Intel i7 with 2.6 GHz are:

| | Time in seconds |
|---|---|
| Monte-Carlo integration (1000 samples/voxel) | 23 |
| max. edge method | 0.17 |
| linked-pairs method | 0.11 |

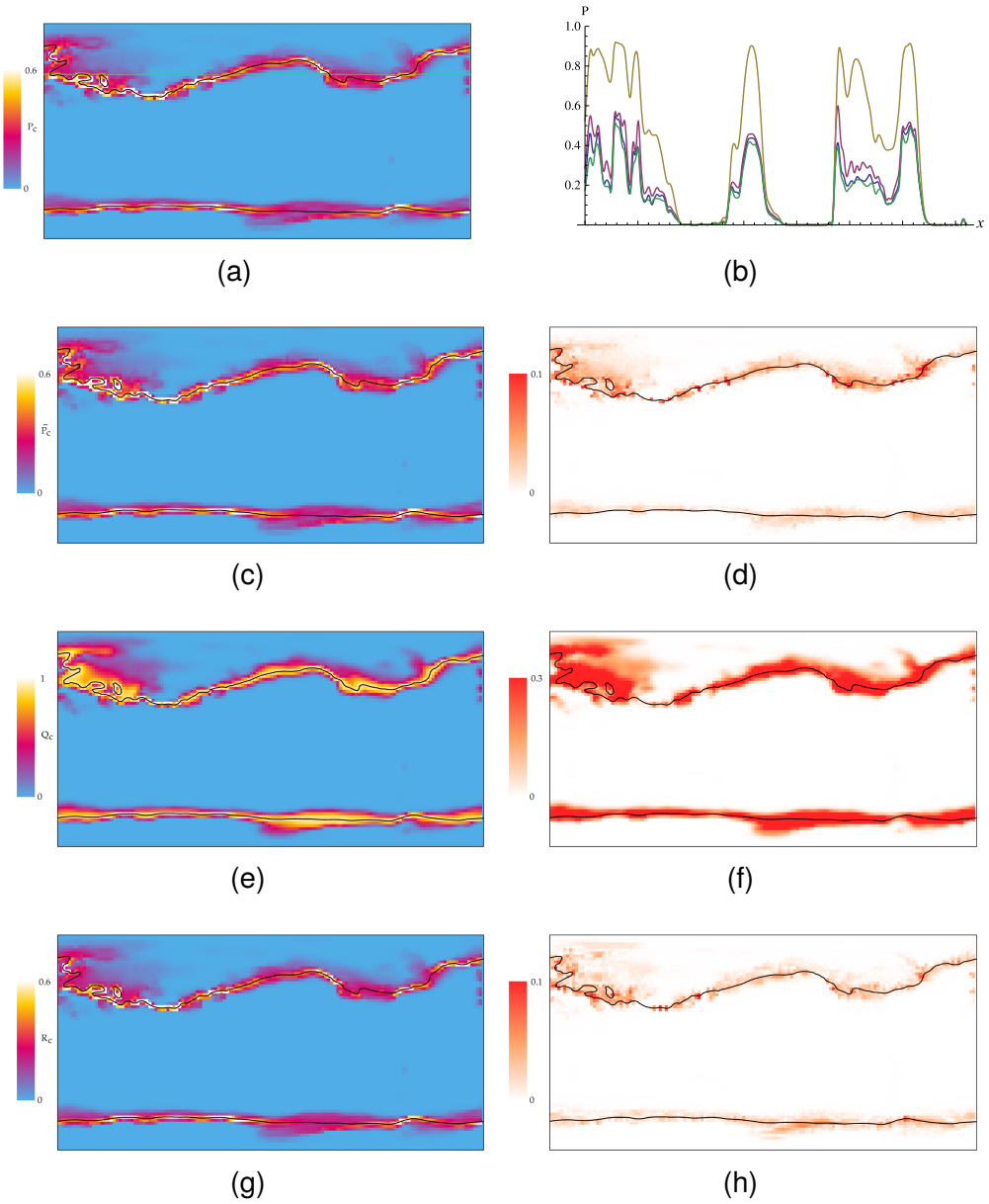



**Figure 6.29:** *Results for the 2D climate dataset: (1) joint cell-wise crossing probability* $\mathrm{P}_c$ *estimated by MC sampling. (3) approximate probabilities* $\tilde{\mathrm{P}}_c$*. (4) difference image* $|\tilde{\mathrm{P}}_c - \mathrm{P}_c|$*. (5) crossing probabilities* $\mathrm{Q}_c$ *assuming uncorrelated values. (6) difference image* $|\mathrm{Q}_c - \mathrm{P}_c|$ *(7) maximum edge crossing probabilities* $\mathrm{R}_c$*. (8) difference image* $|\mathrm{R}_c - \mathrm{P}_c|$*. Note that the ranges of the colormaps are individually adjusted for* $\mathrm{Q}_c$*. In (2) the probabilities along the green line indicated in (1) are shown as 1D-graphs.*

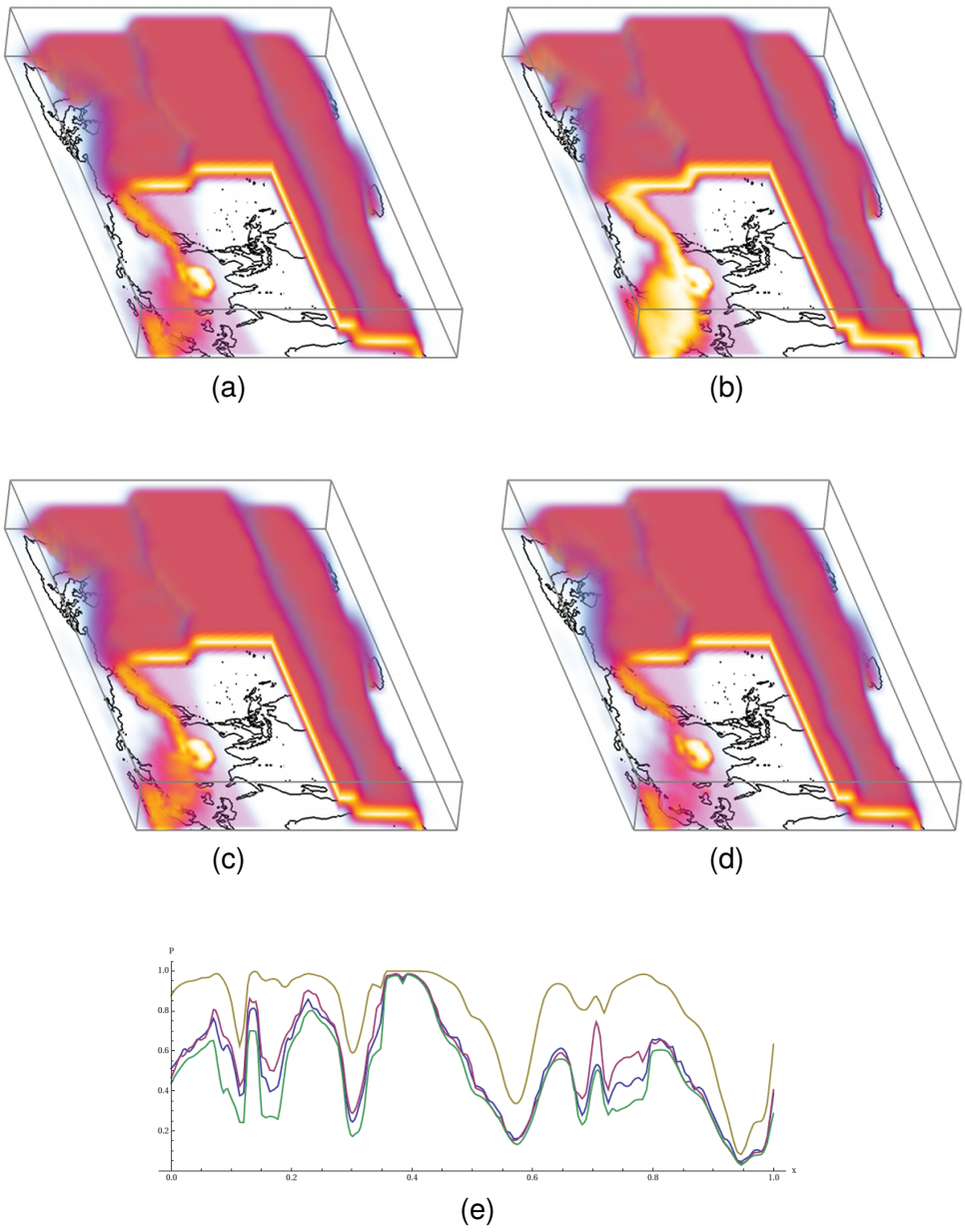



**Figure 6.30:** *Results for the 3D climate dataset: (1) joint cell-wise crossing probability* $\mathrm{P}_c$ *estimated by MC sampling. (2) crossing probabilities* $\mathrm{Q}_c$ *assuming uncorrelated values. (3) approximate probabilities* $\tilde{\mathrm{P}}_c$*. (4) maximum edge crossing probabilities* $\mathrm{R}_c$*. In (5) the probabilities along a straight line in the datasets are shown as 1D-graphs.*

**Discussion.** The major advantage of the approximation methods is that the crossing probabilities can be evaluated using lookup tables which results in much faster computation times compared to MC integration of the $n$-dimensional PDFs. Another advantage of the lookup method compared to the MC integration is that it does not suffer from MC noise.

Like Pfaffelmoser et al. [PRW11] the approximations employ a 3D lookup table for crossing probabilities considering correlation. The motivation for their work was to develop a fast raycasting solution. Instead of local cell-wise crossing probabilities they compute first-crossing probabilities along a ray which yields viewpoint-dependent results. In contrast, the results of our methods do not depend on any direction. By restricting correlation functions to the type $\exp(-|\text{distance}|)$ they could compute all probabilities along each ray. The resulting correlation coefficients corresponding to multiple pairs of random variables along a ray are similar to the covariance matrices induced by our approach to compute the approximate level-crossings using $\tilde{\mathrm{P}}_c$, i.e. the correlation coefficients along the path are multiplied, cf. Eq. (6.33). Probabilities computed with the linked-pairs approximation over- or underestimate the true level-crossing probability. In contrast, the maximum edge probability yields a true lower bound.

In Fig. 6.27, the volume enclosed by the semi-transparent grey surfaces illustrates the space of valid correlation matrices, e.g., positive-semidefinite matrices, for 3 random variables. The red opaque surface depicts the result for the computed correlation in the linked-pairs approximation. The approximation over- or underestimates the real correlation. Correlation matrices that are computable with the linked-pairs approximation are located on a 2d subspace of all valid correlation matrices. The approximation projects the unused correlation onto that subspace. Note that the red surface depicts only one traversal order $k$; matrices on two additional surfaces are used in the approximation for the remaining two traversal order choices.

In the optimization step for the choice of parameter $k$ in Eq. (6.30) we proposed using the Bhattacharyya distance as quality measure for the approximation. Fig. 6.26 (1) confirms that this is a good measure, as a positive correlation between Bhattacharyya distance $D_B$ and approximation error exists. Choosing the optimal parameter $k$ significantly decreases the Bhattacharyya distances and the approximation errors, as can be read off in Figs. 6.26 (2), (3) and (4). In that example, the linked-pairs approximation outperforms the maximum-edge approximation for optimized choice of $k$, but not for randomly chosen $k$. Searching for the linked-pairs approximation with lowest Bhattacharyya distance as proposed in Eq. (6.34) requires a traversal over all spanning trees for each cell. The number of spanning trees increases with $O(\exp(n \ln n))$, with $n$ the number of cell vertices, what can be costly for cells with many vertices. The search is independent from

a specific threshold value $\vartheta$, and can thus be performed in a preprocessing step, allowing interactive evaluation afterwards.

A limitation of the proposed approximations is that they do not allow to improve accuracy using a parameter or additional terms of a series expansion. As we focused on the evaluation of one and two dimensional distributions it is clear that we can not reach arbitrary accuracy.

The results of the synthetic datasets in Fig. 6.25 and the climate data in Fig. 6.29 reproduce the result from previous work that spatial correlations have a significant impact on crossing probabilities. In all results the differences between $P_c$ and $Q_c$ are significant. In Fig. 6.25 (a) the graph also shows qualitatively different behavior compared to the other methods. For rectangular grid cells (2D data), see Fig. 6.25, $\tilde{P}_c$ overestimates $P_c$ in most cases, while $R_c$ underestimates it. $\tilde{P}_c$ is closer to $P_c$, a large approximation error of $R_c$ from $P_c$, up to over 0.2, can be observed in Fig. 6.25 (d). For the climate dataset, see Fig. 6.29 however, almost all deviations are below 0.1. The approximation of the crossing probability $P_c$ both using $\tilde{P}_c$ and the maximum edge-crossing probability $R_c$ yields quantitatively and qualitatively good results. Visual impressions are true to the results of $P_c$. The 3D example yields similar results. In Fig. 6.30 we can observe that neglecting the correlation leads to much overestimated probabilities $Q_c$ while $\tilde{P}_c$ and $R_c$ approximate $P_c$ quite well.

The approximation methods reduce the computation of level-crossing probabilities to evaluations of uni- and bivariate CDFs. This is implemented using lookup tables to avoid expensive numerical integration during rendering. In the maximum edge crossing approximation the edge-related crossing probabilities of a grid cell are computed and the maximum is taken. In the linked-pairs approximation, pairwise correlations are evaluated step by step, spanning all random variables of a cell. This induces an approximate distribution that is again normally distributed. We used the Bhattacharyya distance for choosing the optimal approximation and showed that it is a good measure for minimizing the approximation error.

Above results confirm that it is essential to consider spatial correlations. Both approximation methods, the maximum edge crossing method and the linked-pairs method show comparable good results for real world data. While the maximum edge crossing method is conceptually simpler and provides a lower-bound for real cell-wise probabilities, linked-pairs requires a preprocessing step and – for the datasets analyzed – outperforms the maximum edge crossing method in terms of accuracy. The approximated level-crossing probabilities are in good agreement with the true cell-wise crossing probabilities, although pathological cell configurations exist where the error can be high. Experiments show that the approximation works very well in practice, and differences are hardly observable in the visualizations.

### 6.4.2 Surrogate Functions

In previous sections $P_c$ was either computed using computationally expensive Monte Carlo (MC) integration or approximated using pre-computed lookup tables for low dimensional distribution functions, see Sect. 6.4.1.

Computation times of several minutes to hours were reported for MC integration. While the approximation methods are much faster, they have two important drawbacks: (i) they are fixed formulations for level-crossing probabilities with no obvious way to adapt them to other types of features and (ii) they exhibit a fixed approximation error that cannot be reduced systematically.

In this section we propose *surrogate functions* for significantly accelerated estimation of feature probabilities. In domains like statistical data analysis or experimental design surrogate functions (or surrogate models) are often used as a tool to efficiently predict unobserved outcomes that would otherwise be costly to obtain [QHS*05,GCD*10,ONK03]. For feature probabilities – instead of solving the integral for each grid cell – we construct a function that maps cell attributes to the resulting probability and evaluate this function for all grid cells. Interpreting it as a regression problem we can also refer to the attributes as the independent variables and to the probability as the dependent variable. The model can either be built in a preprocessing step (by performing MC integration for a large number of distributions) and stored on disk or it can be incrementally refined in an online-learning procedure. A variety of models such as generalized linear models or support vector machines can be used for this purpose. For our implementation we have chosen $\mathcal{K}$-nearest-neighbors ($\mathcal{K}$-NN) regression [Alt92] because it can be quickly evaluated and the results only depend on the single parameter $\mathcal{K}$. This parameter matches our intuitive conception of a smoothing constant w.r.t. attribute space.

The approach has three main advantages. First, the computation of feature probabilities for a field using the surrogate function is orders of magnitude faster than MC integration. Second, the approach is flexible; surrogate functions for various types of features can be constructed. And third, as the accuracy of the surrogate model increases by adding more sample points to the training set, the resulting probability fields approach the ground truth. Thus, this method overcomes the main disadvantages of previous approximation methods.

#### 6.4.2.1 General Formulation

The aim for the formulation of a surrogate function is to quickly and efficiently estimate results that are otherwise costly to obtain. For our specific problem of probabilistic feature extraction we want to define a function that

estimates probabilities much faster than MC integration while still being able to reach high accuracy. Depending on the definitions, different features will lead to different surrogate functions but the overall approach is generic.

Consider a grid cell in a discretized random field with a PDF $f_c(\mathbf{y}_c)$ for which the resulting probability $P_c$ can be computed according to Eq. (6.2). In practice, $f_c$ and $I$ are uniquely defined by a finite number of *attributes*, e.g. expected values and covariances for Gaussian distributions, a sample of realizations for nonparametric distributions or parameters specifying the feature indicator. Let $\mathbf{u}$ be a vector with $d_\mathbf{u} = \dim(\mathbf{u})$ that combines all these attributes in a serialized fashion, then we can rewrite the probability $P_c$ as $P(\mathbf{u})$ to make its dependency on the attributes explicit.

A function

$$\zeta : \mathbb{R}^{d_\mathbf{u}} \to [0, 1], \tag{6.35}$$

that maps the attribute vector $\mathbf{u}$ to a probability that is approximately equal to the 'true' probability

$$\zeta(\mathbf{u}) \approx P(\mathbf{u}) \tag{6.36}$$

is called a *surrogate function* or *surrogate model* for $P(\mathbf{u})$. The approximation error depends on the complexity of $\zeta$ and on the choice of its parameters. Candidates for the choice of $\zeta$ include regression methods based on support vector machines, generalized linear models and $\mathcal{K}$-NN.

#### 6.4.2.2 Creating the Training Set

Before $\zeta$ can be used to predict probabilities it has to be defined by processing a set of examples using a model fitting/training procedure. We call the set of tuples

$$\mathbb{U} = \{ \, (\mathbf{u}_i, \, P(\mathbf{u}_i)) \mid i \in \{1, \ldots, n_\zeta\} \, \} \tag{6.37}$$

a *training set*, where $\mathbf{u}_i$ is the $i$-th attribute vector for a grid cell selected from some input data set (discretized random field), $P(\mathbf{u}_i)$ is the feature probability and $n_\zeta$ is the number of training examples. In a training step $\zeta(\mathbf{u})$ is determined to reflect the training data as good as possible in order to predict the unknown value of $P(\mathbf{u})$ for *new* examples with attribute vectors $\mathbf{u}$.

In some applications the acquisition of each training example can be very expensive, e.g. in cases where experiments have to be carried out. For feature probabilities however, we can compute $P(\mathbf{u}_i)$ for any $\mathbf{u}_i$ using MC integration. Thus, the only cost in our case is processing time. There are two different ways to perform the model fitting, pre-computation and incremental learning.

**Pre-Computation.** To construct the surrogate function $\zeta$ in a preprocessing step we randomly draw $n_\zeta$ attribute vectors $\mathbf{u}_i$ of cells from a set of fields. The pre-computation consists of the following steps, to be repeated for all $n_\zeta$ training examples:

1. randomly select a cell from a discretized random field
2. combine the attributes for the cell into vector $\mathbf{u}_i$
3. compute $P(\mathbf{u}_i)$ using MC integration
4. add $(\mathbf{u}_i, P(\mathbf{u}_i))$ to $\mathbb{U}$

After that the training algorithm can be executed for $\mathbb{U}$ to obtain $\zeta(\mathbf{u})$ which is then ready to be evaluated. Either $\mathbb{U}$ or the model that is derived by the training algorithm can be stored on disk to make it available for later use and avoid repeated pre-computation.

**Incremental Learning.** Instead of computing a large number of probability values in advance we can also incrementally refine $\zeta(\mathbf{u})$. We define a maximum error $\epsilon_{\max}$ and start with an empty set $\mathbb{U}$. For each cell in the input data set we determine $\mathbf{u}$ and estimate the prediction error $\epsilon$ for the current version of $\zeta(\mathbf{u})$. For some types of surrogate functions the error can be estimated in terms of function values. Alternatively we can consider the distance of $\mathbf{u}$ to the training examples as an error metric, assuming that the prediction error increases proportionally. If $\epsilon$ is larger than $\epsilon_{\max}$ we compute $P(\mathbf{u})$ using MC integration, store it in the resulting field and also store $(\mathbf{u}_i, P(\mathbf{u}_i))$ in $\mathbb{U}$. In case $\epsilon_{\max} < \epsilon$ we evaluate $\zeta(\mathbf{u})$ and store the value in the resulting field. Some learning algorithms can efficiently update the surrogate model using new training examples, e.g. online Gaussian processes [CO02, DW13]. For methods for which this is impossible (i.e. where the learning algorithm always processes the complete training set) the training algorithm can be re-run as soon as the size of $\mathbb{U}$ has grown a predefined fraction compared to the last run.

#### 6.4.2.3 Estimation of Feature Probabilities using $\mathcal{K}$-Nearest-Neighbors ($\mathcal{K}$-NN)

The $\mathcal{K}$-nearest-neighbors ($\mathcal{K}$-NN) algorithm is one of the simplest supervised learning methods. It stores all training examples and does not explicitly derive a generalization. Its predictions for regression problems are based on (weighted) averages of known examples. In the following we present two specific formulations of $\mathcal{K}$-NN surrogate functions for feature probabilities.

#### 6.4.2.4 $\mathcal{K}$-NN Surrogate Functions for Level-Crossing Probabilities

**Attribute Vectors.** To formulate a surrogate function for the computation of level-crossing probabilities we need to define attribute vectors that completely characterize the probability integral to be approximated. For Gaussian fields the uncertainty of all scalar values corresponding to a cell $c$ can be modeled by a joint random vector $\mathbf{Y}_c \sim \mathcal{N}(\mu_c, \Sigma_c)$. Then, a simple approach is the direct representation of the threshold $\vartheta$, all expected values $\mu_i$ and all covariances $\sigma_{i,j}$ of the PDF $f_c$

$$\mathbf{u} = (\vartheta, \mu_1, \ldots, \mu_{d_c}, \sigma_{1,1}, \sigma_{1,2}, \ldots, \sigma_{d_c,d_c}).$$

However, the dimensionality of $\mathbf{u}$ can be reduced by expressing the level-crossing probability in a standardized form, i.e. in terms of stochastic distance functions $\Psi_i = \frac{\mu_i - \vartheta}{\sqrt{\sigma_{i,i}}}$, and correlation coefficients $\rho_{i,j} = \frac{\sigma_{i,j}}{\sqrt{\sigma_{i,i}\,\sigma_{j,j}}}$, see [PRW11, PPH13]. The attribute vector can now be defined using $\Psi_i$ and $\rho_{i,j}$ (all off-diagonal entries of the correlation matrix)

$$\mathbf{u} = (\Psi_1, \ldots, \Psi_{K_c}, \rho_{1,2}, \ldots, \rho_{K_c,K_c-1}) \tag{6.38}$$

For triangular cells this reduces the dimensionality $d_\mathbf{u}$ of $\mathbf{u}$ from 10 to 6. For grids with other cell types (e.g. tetrahedra or hexahedra) attribute vectors can be defined analogously but with larger numbers of attributes.

Using this definition of $\mathbf{u}$ we can perform one of the learning approaches described in Sect. 6.4.2.2. For $\mathcal{K}$-NN the learning algorithm does not create a generalization of $\mathbb{U}$ but it creates efficient data structures such that searching for nearest neighbors can be performed quickly.

**Function Evaluation.** To evaluate $\zeta_\mathcal{K}(\mathbf{u})$ for a given vector $\mathbf{u}$ and predict the corresponding level-crossing probability, we look up the $\mathcal{K}$ members $(\mathbf{u}_i, P(\mathbf{u}_i)) \in \mathbb{U}$ where $\mathbf{u}_i$ are closest to $\mathbf{u}$. The function is then evaluated as

$$\zeta_\mathcal{K}(\mathbf{u}) = \frac{1}{\mathcal{K}} \sum_{i=1}^{\mathcal{K}} P(\mathbf{u}_i). \tag{6.39}$$

An alternative approach is to weigh the summands depending on the distance

$$\zeta_\mathcal{K}(\mathbf{u}) = \sum_{i=1}^{\mathcal{K}} w_i P(\mathbf{u}_i). \tag{6.40}$$

where

$$w_i = \frac{\beta(|\mathbf{u} - \mathbf{u}_i|)}{\sum_{j=1}^{\mathcal{K}} \beta(|\mathbf{u} - \mathbf{u}_j|)}$$

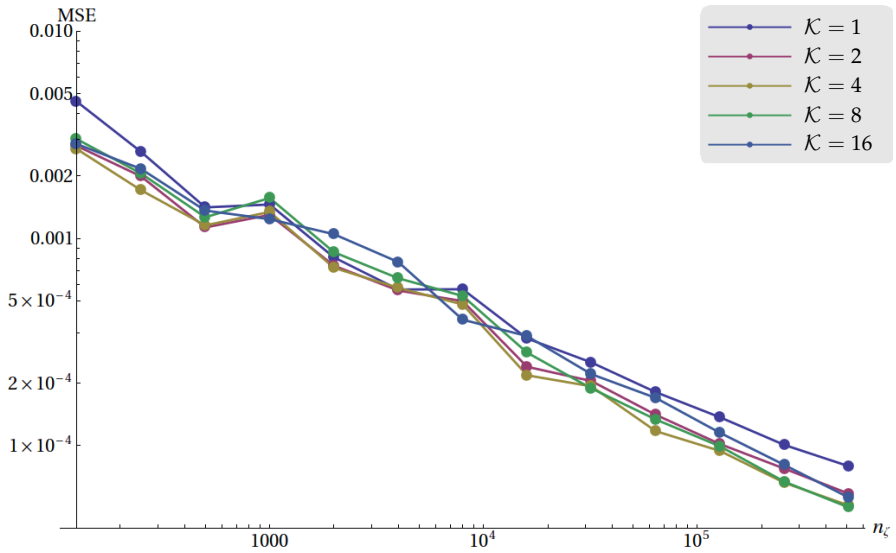



**Figure 6.31:** *Generalization performance (level-crossing probabilities): Mean squared prediction errors (MSE) of level-crossing probabilities estimated using 5-fold cross validation for increasing sizes of training sets of the $\mathcal{K}$-NN surrogate functions are shown in a log-log plot. The number $n_\zeta$ of training examples which are randomly subsampled from a large training set increases from* 125 *to* 512 000 *(from left to right). The varying results for different values of $\mathcal{K}$ are indicated using distinct colors.*

and $\beta : \mathbb{R}^+ \to \mathbb{R}$ is a monotonically decreasing function, e.g., a inverse (reciprocal), half-Gaussian or linear function.

#### 6.4.2.5 $\mathcal{K}$-NN Surrogate Functions for Critical-Point Probabilities

To represent uncertain vector fields and compute probabilities for the existence of critical points we need to consider random vectors of higher dimensionality. The uncertainty of all vectors adjacent to a cell $c$ can be modeled by a joint random vector $\mathbf{Y}_c \sim \mathcal{N}(\mu_c, \Sigma_c)$ with PDF $f_c$, see Eq. (3.8). For Gaussian fields we could characterize this distribution using an attribute vector consisting of all expected values and covariances. However, with a larger number of dimensions it is becoming more and more challenging to sample the space of attribute vectors and to obtain a reasonably accurate surrogate function $\zeta_\mathcal{K}(\mathbf{u})$.

For 2D Gaussian random vector field given on a regular grid with rectangular grid cells and four 2D vectors located at the vertices of each cell we model the correlations using exponential functions, cf. Eq. (3.9). This allows us to represent $\mathbf{Y}_c$ using an attribute vector with relatively few dimensions. Locally, for each cell $c$ we fit two functions to the correlation coefficients that are empirically estimated from ensemble data. The first function we consider

is called *autocorrelation function* $R_1(h) = \exp(-\gamma_1 h)$, where $h = \|\mathbf{x}_i - \mathbf{x}_j\|$ is the Euclidean distance between the two respective vertices of $c$ and $\gamma_1$ is the parameter that describes the falloff rate of spatial correlation. $R_1$ quantifies the correlation between corresponding components of the vector field at different locations of the grid, e.g. $\mathrm{Corr}(y_{i,1}, y_{j,1})$ where $y_{i,1}, y_{j,1}$ are components of $\mathbf{Y}_c$. The second function is called the *cross-correlation function* $R_2(h) = r \exp(-\gamma_2 h)$, and quantifies the correlation between the different components of the vector field, e.g. $\mathrm{Corr}(Y_{x,i}, Y_{y,j})$. For $h = 0$ the value $R_2(0) = r = \mathrm{Corr}(Y_{x,i}, Y_{y,i})$ is the correlation between the two components at vertex $i$, i.e. between the x-direction and y-direction. We estimate the parameters $\gamma_1$, $\gamma_2$ and $r$ locally for $\mathbf{Y}_c$ using least squares.

These parameters together with the mean values and standard deviations constitute the attribute vector

$$\mathbf{u} = (\mu_1, \ldots, \mu_{d_c}, \sigma_1, \sigma_2, \ldots, \sigma_{d_c}, \gamma_1, \gamma_2, r). \tag{6.41}$$

that we use for 2D vector fields. The dimensionality is $d_{\mathbf{u}} = 8 + 8 + 3) = 19$. The modelling error of the exponential function approach compared to considering arbitrary correlations is investigated below.

For critical-point probabilities, instead of a single probability we can store probabilities for the existence of sources, saddles and sinks for each $\mathbf{u}_i$ which results in a 3-valued surrogate function. After applying one of the model-fitting approaches described in Sect. 6.4.2.2 we can evaluate $\zeta_{\mathcal{K}}(\mathbf{u})$ analogously to Eq. (6.39).

#### 6.4.2.6 Implementation

The methods were implemented to create and evaluate surrogate functions based on $\mathcal{K}$-NN in C++ following the derivations given above. To get high performance with regard to the nearest-neighbor search we utilized the approximate nearest-neighbor library *ANN* that was published under the LGPL. A detailed description of the method and its performance were published by Arya et al. [AMN*98]. The computational complexity for the estimation of the feature probability at one grid cell using the $\mathcal{K}$-NN method is $O(\mathcal{K} d_{\mathbf{u}} \log n_\zeta)$. The performance regarding accuracy and computation times for the surrogate functions of feature probabilities is evaluated empirically in the next section.

#### 6.4.2.7 Results

We evaluate the methods introduced above by computing feature probabilities for uncertain scalar and vector fields using $\mathcal{K}$-NN surrogate functions of varying complexity. The most important quality of surrogate functions is

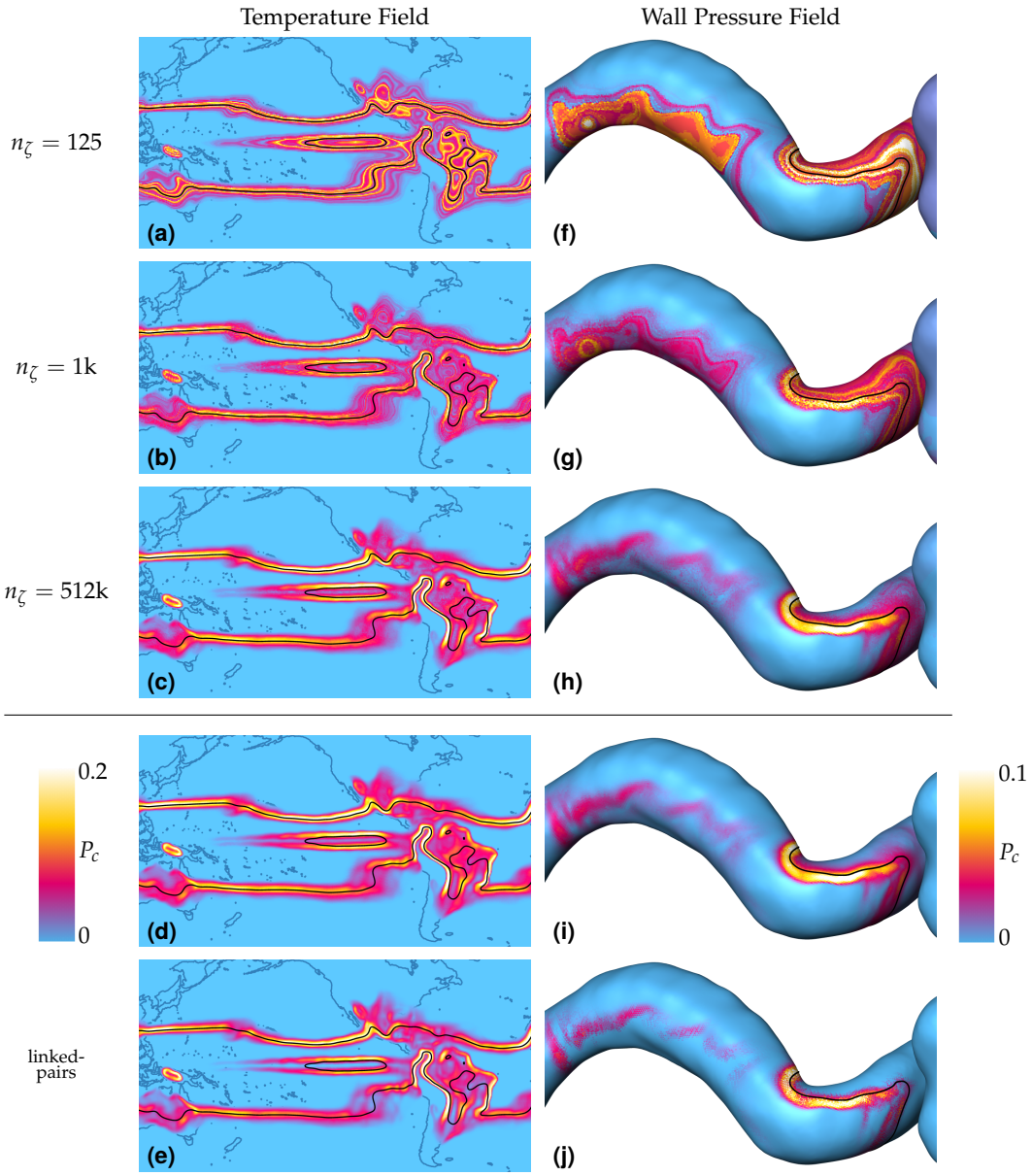



**Figure 6.32:** *Level-crossing probabilities in an uncertain temperature field for* $\vartheta = 24°C$ *in (a) - (e) and in a wall pressure field in (f) - (j) are depicted using color mapping with the mean isoline shown in black. In (a) - (c) and (f) to (h) the cell-wise probabilities are computed using a K-NN surrogate function where the number* $n_\zeta$ *of training examples increases from* 125 *to* 512 000 *with parameter* $K = 4$*. For comparison results computed using the linked-pairs approximation are shown in (e) and (j) and the benchmark result computed using MC integration is shown in (d) and (i). Note that the quality increases with larger* $n_\zeta$ *and that (c) and (h) approximate the MC results more closely than (e) and (j). Both data sets are not part of training data, i.e. the K-NN results are out-of-sample predictions.*

the ability to *generalize* to new (out-of-sample) data. In the following, we examine the generalization performance using two different approaches. First, we estimate mean prediction errors using cross validation of training sets with varying sizes and varying parameter $\mathcal{K}$. Second, we compare $\mathcal{K}$-NN results to a reference result that was calculated using MC integration with many samples and that we consider the ground truth for this evaluation. All computations were performed on an Intel Xeon X5650 with 2.66 GHz using a single-threaded implementation.

**Pre-Computation.** The training sets for $\zeta_{\mathcal{K}}(\mathbf{u})$ were computed using the pre-computation approach, see Sect. 6.4.2.2. For level-crossing and critical-point probabilities we took training examples from time series of 2 meter temperature fields and 10 meter wind fields, respectively, which were provided by the DEMETER project [Pal04]. For level-crossing probabilities the creation of the training set $\mathbb{U}$ with $n_\zeta = 512\,000$ took several minutes. For critical-point probabilities the creation of the training set $\mathbb{U}$ with $n_\zeta = 2\,048\,000$ took several hours. Most of the computation time is necessary for Monte Carlo sampling (50000 MC samples per cell for level-crossing probabilities, 80000 MC samples per cell for critical point probabilities). The smaller training sets that are considered in the following are random subsamples of the large training sets.

Note that the resulting surrogate functions are *not* specific to particular data sets. By definition of the attribute vector in Eq. (6.38) the surrogate function for level-crossing probabilities does not depend on the parameter $\vartheta$. The training sets are stored on disk to avoid repeated pre-computation.

**Level-Crossing Probabilities.** To estimate the *generalization performance* of $\zeta_{\mathcal{K}}$ for level-crossing probabilities we computed mean squared prediction errors (MSE) using 5-fold cross validation for increasing sizes of training sets. That means that we divided the data set into 5 parts where four-fifths that are used for training $\zeta_{\mathcal{K}}$ and one fifth is used for testing. For all examples $(\mathbf{u}_i, P(\mathbf{u}_i))$ in the test subset the squared error of the prediction

$$(P(\mathbf{u}_i) - \zeta_{\mathcal{K}}(\mathbf{u}_i))^2$$

is computed and averaged. This is repeated and averaged for the 5 possible choices of the test subset. The results are shown in Fig. 6.31 using a log-log plot. The number $n_\zeta$ of examples which are randomly subsampled from the large, initially created training set increases from 125 to 512 000 (from left to right). The test subset is taken from the $n_\zeta$ examples so the actual training sets contain $100, \ldots, 409\,600$ examples. The varying results for different values of $\mathcal{K}$ are indicated using different colors.

| $n_\zeta$ | time (seconds) | MSE |
|---|---|---|
| 125 | 1.30 | $3.52 \times 10^{-4}$ |
| 1000 | 1.54 | $1.26 \times 10^{-4}$ |
| 512 000 | 3.25 | $6.76 \times 10^{-6}$ |
| *ground truth (MC)* | 696.68 | – |
| *linked-pairs* | 1.13 | $5.52 \times 10^{-5}$ |

(a)

| $n_\zeta$ | time (seconds) | MSE |
|---|---|---|
| 125 | 1.90 | $7.17 \times 10^{-4}$ |
| 1000 | 2.30 | $1.77 \times 10^{-4}$ |
| 512 000 | 3.81 | $9.71 \times 10^{-6}$ |
| *ground truth (MC)* | 7340.97 | – |
| *linked-pairs* | 1.64 | $8.80 \times 10^{-5}$ |

(b)

**Table 6.1:** *Computation times and approximation errors (MSE) for level-crossing probabilities in the results depicted in Fig. 6.32. Values are given for the (a) temperature field and (b) wall pressure field.*

We also computed level-crossing probabilities for two discretized random fields from different data sources: an uncertain temperature field ensemble that was not part of the cross validation data set and a vessel wall pressure ensemble from a blood flow simulation. Fig. 6.32 shows results computed for the temperature field for $\vartheta = 24°C$ in Fig. 6.32 (a) - (e) and in the wall pressure field in Fig. 6.32 (f) - (j) that are depicted using color mapping with the mean isoline shown in black. In Fig. 6.32 (a) - (c) and (f) to (h) the cell-wise probabilities are computed using a $\mathcal{K}$-NN surrogate function where the number $n_\zeta$ of training examples increases from 125 to 512 000 with parameter $\mathcal{K} = 4$. For comparison results computed using the linked-pairs approximation are shown in Fig. 6.32 (e) and (j) and the benchmark results computed using MC integration is shown in Fig. 6.32 (d) and (i). Note that the quality increases with larger $n_\zeta$ and that the $\mathcal{K}$-NN results with the largest training set approximate the MC results much more closely than the linked-pairs method. Both data sets are not part of training data, i.e. the $\mathcal{K}$-NN results are out-of-sample predictions. The computation times and mean squared prediction errors (MSE) estimated for all cells of the respective fields are given in Table 6.1.

**Critical-Point Probabilities.** We investigated the generalization performance of $\zeta_{\mathcal{K}}$ for critical point probabilities analogously to the level-crossing case. The mean squared errors (MSE) of probabilities for the existence of critical points that are estimated using 5-fold cross validation for increasing sizes of training sets of the $\mathcal{K}$-NN surrogate functions are shown in Fig. 6.33 in a log-log plot. The number $n_\zeta$ of examples which are randomly subsampled from a large training set increases from 250 to 2 048 000 (from left to right). The varying results for different values of $\mathcal{K}$ are indicated using different colors.

Probabilities for the existence of sinks (critical points) in an uncertain wind velocity field in Fig. 6.34 (a) - (e) and in a blood flow field in (f) - (j) are depicted using color mapping. For the blood flow field the parameters of the Gaussian field were estimated from an simulation ensemble of 100 realizations of turbulent flow through an artificial heart valve. Both data sets are not part of training data, i.e. the $\mathcal{K}$-NN results are out-of-sample predictions. In Fig. 6.34 (a) - (c) and (f) to (h) the cell-wise probabilities are computed using a $\mathcal{K}$-NN surrogate function where the number $n_\zeta$ of training examples increases from $1k$ to $2M$ with parameter $\mathcal{K} = 4$. For comparison the benchmark results computed using MC integration are shown in Fig. 6.34 (d) - (e) and (i) - (j). The results in Fig. 6.34 (e) and (j) are computed for random fields with arbitrary correlations. For the results in all the other subfigures, spatial correlation is modelled using exponential correlation functions. Note that the approximation quality increases with larger $n_\zeta$ and that there are only minor observable differences between the two correlation models (e.g. in the vicinity of the north and the south pole). The computation times and mean squared approximation errors (MSE) estimated for all cells of the respective fields are given in Table 6.2.

#### 6.4.2.8 Discussion

The results show that the feature probabilities estimated using the $\mathcal{K}$-NN surrogate functions can achieve good accuracy. The generalization performance, as investigated using 5-fold cross validation of the training sets, improves for increasing numbers $n_\zeta$ of training examples. This applies to both level-crossing and critical-point probabilities. The rate of improvement w.r.t. increasing $n_\zeta$, however, depends on the dimensionality of the attribute vectors $\mathbf{u}$. The MSE decreases more quickly for the level-crossing probabilities with $d_{\mathbf{u}} = 6$ compared to the critical-point probabilities ($d_{\mathbf{u}} = 19$). The difference in the amount of training examples that is is needed to significantly improve the accuracy is due to the *curse of dimensionality*, see, e.g. [Bis06, pp. 33-38]. For a given number of points distributed in space the 'sparsity' increases quickly for an increasing number of dimensions. Note that the differ-

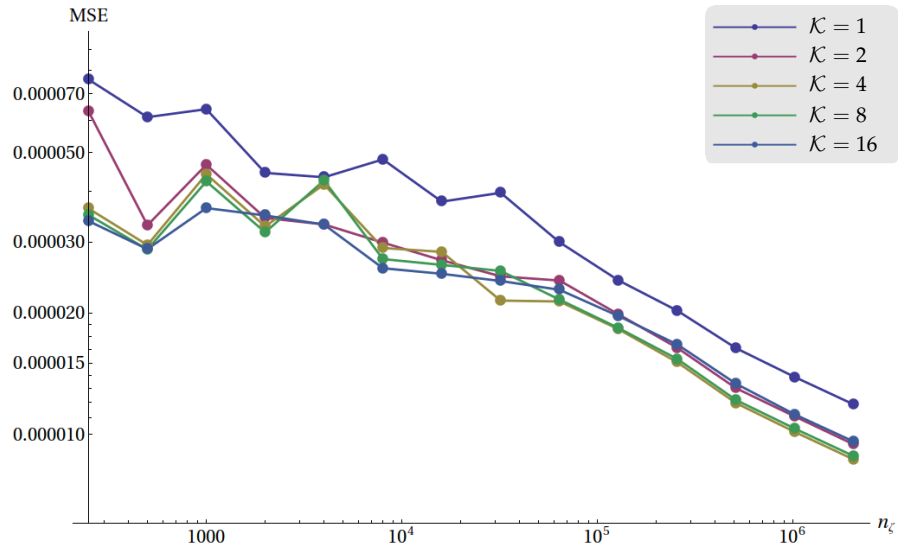



**Figure 6.33:** *Generalization performance (critical-point probabilities): Mean squared prediction errors (MSE) of probabilities for the existence of critical points that are estimated using 5-fold cross validation for increasing sizes of training sets of the $\mathcal{K}$-NN surrogate functions are shown in a log-log plot. The number $n_\zeta$ of training examples which are randomly subsampled from a large training set increases from 250 to 2M (from left to right). The varying results for different values of $\mathcal{K}$ are indicated using different colors.*

ent orders of magnitude of the absolute MSE values are due to the different ranges of probabilities in the results as well as due to the different counts of cells/training examples with zero or almost-zero probabilities.

The parameter $\mathcal{K}$ can be interpreted as a smoothing parameter w.r.t. the attribute space. The cross validation shows that the choice of $\mathcal{K}$ has an impact on the accuracy of the predictions: on average $\mathcal{K} = 1$ lead to the poorest performances, $\mathcal{K} = 4$ and $\mathcal{K} = 8$ performed well and for $\mathcal{K} = 16$ the accuracy decreased again, see Fig. 6.31 and 6.33. We chose $\mathcal{K} = 4$ for the subsequent examples because it performs well and it is computationally advantageous as it leads to relatively few queries for nearest neighbors. To create surrogate function that generalize well to new fields it is beneficial to express the attribute vectors $\mathbf{u}$ in a normalized way. For example, using the attributes defined in Eq. (6.38) the feature probability is expressed in terms of the standard normal distribution which makes the resulting $\zeta_\mathcal{K}$ scale invariant.

The results computed for the out-of-sample data sets show that the accuracy and generalization performance is good – both qualitatively and quantitatively – if the training sets are large enough. Generalization performance

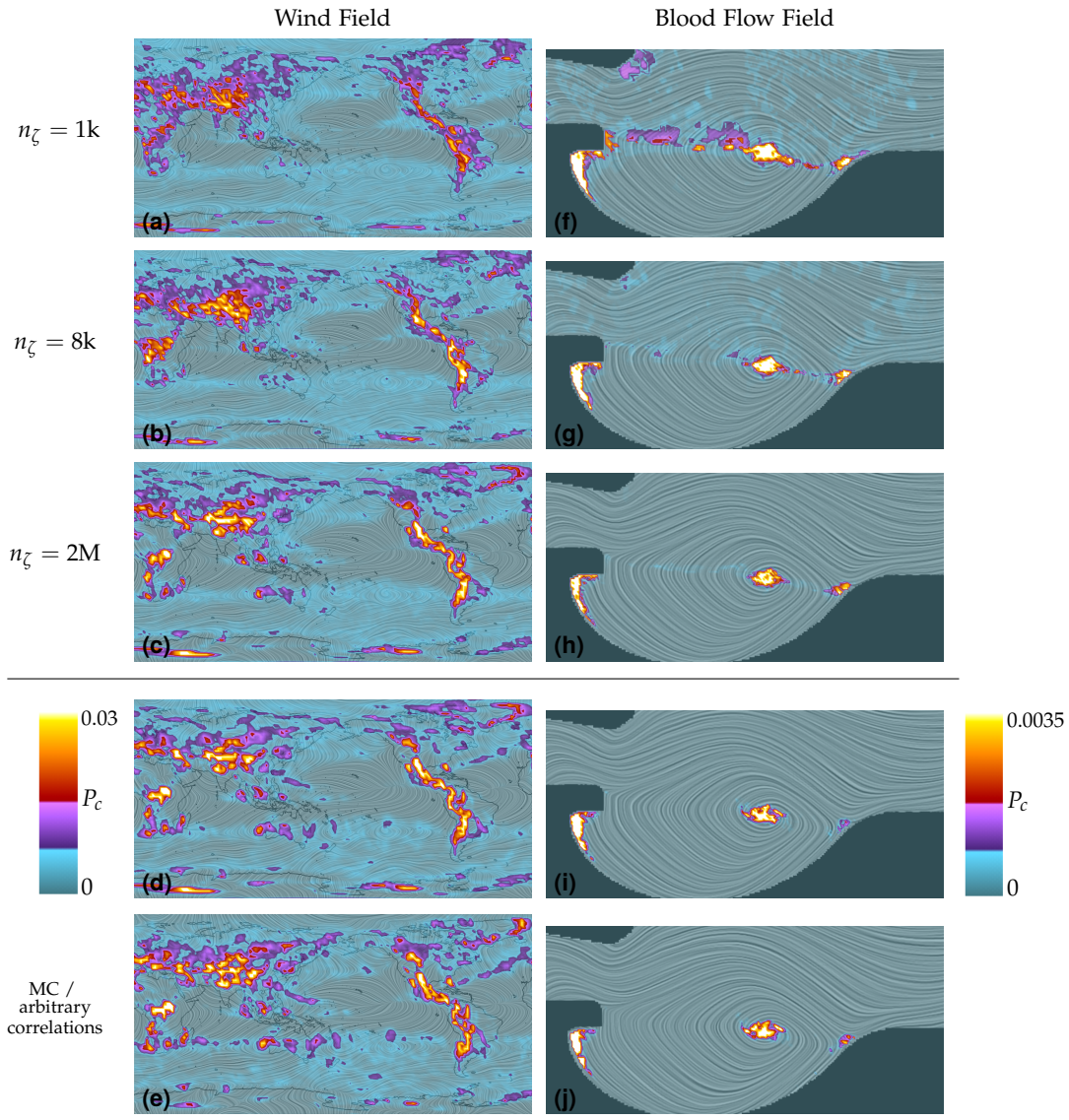



**Figure 6.34:** *Probabilities for the existence of sinks (critical points) in an uncertain wind velocity field in (a) - (e) and in a blood flow field in (f) - (j) are depicted using color mapping. In (a) - (c) and (f) to (h) the cell-wise probabilities are computed using a $\mathcal{K}$-NN surrogate function where the number $n_\zeta$ of training examples increases from 1k to 2M with parameter $\mathcal{K} = 4$. For comparison the benchmark results computed using MC integration are shown below the separator line. The results in (e) and (j) are computed for random fields with arbitrary correlations. For the results in all the other figures, spatial correlation is modelled using exponential correlation functions. Note that the approximation quality increases with larger $n_\zeta$ and that there are only minor observable differences between the two correlation models. Both data sets are not part of training data, i.e. the $\mathcal{K}$-NN results are out-of-sample predictions.*

is good, even if the data arise from completely different data sources, i.e. when the training is performed on climate simulation results and feature probabilities are predicted for blood flow simulations. For level-crossing probabilities the accuracy employing the complete training set (512k examples) is significantly better compared to the linked-pairs method. While the linked-pairs method is faster the controlled approximation is the major advantage of the surrogate function approach. The critical-point probabilities shown in Fig. 6.32 and 6.34 show that the smaller training sets lead to artifacts and low approximation accuracy. However, when the large training sets are considered, only very minor differences to the reference results that were computed for the exponential correlation model are observable. There are also some minor differences between the MC results using the exponential correlation model and those using arbitrary correlations, but all significant features (i.e. regions with relatively high feature probabilities) are very similar. This is evidence that the exponential correlation functions are an appropriate model for both types of CFD ensembles. Note that none of the distributions of the fields shown in Fig. 6.32 and 6.34 are part of the training sets, i.e. the probabilities are out-of-sample predictions.

An important limitation of the surrogate function approach is the difficulty to represent more complex probability distributions (e.g. given by an nonparametric estimator) efficiently in an attribute vector $\mathbf{u}$. While it is theoretically possible to represent a distribution using kernel density estimation and store a set of realizations in $\mathbf{u}$, we need to consider values for the complete cell for each realization such that even a moderate amount of realizations will lead to very high dimensional $\mathbf{u}$. A possible option to overcome this limitation in the future is to employ sparse kernel density estimation [CHH04] for nonparametric discrete random fields. Another limitation is that a single function $\zeta$ cannot estimate feature probabilities defined for cells that have varying numbers of neighbors.

We chose $\mathcal{K}$-NN regression for our surrogate functions because it (i) works well with large training sets, provided they fit into memory, (ii) can model nonlinear functions, (iii) has only one hyperparameter that needs to be optimized, and (iv) is rather straightforward to implement. An investigation of other regression methods for the estimation of surrogate functions is left for future work. A major challenge will be to find regression methods that perform well with such large training sets. Candidates include generalized linear models (binomial regression) and nonparamtric regression approaches employing kernel approximations. Standard SVM regression with nonlinear kernels as well as Gaussian process regression does not scale well enough to hundreds of thousands or millions of training examples. The computational complexity of SVMs is between quadratic and cubic for the training set size [BL07].

| $n_\zeta$ | time (seconds) | MSE |
|---|---|---|
| 1000 | 0.44 | $3.14 \times 10^{-5}$ |
| 8000 | 0.86 | $2.16 \times 10^{-5}$ |
| 2048000 | 9.71 | $5.24 \times 10^{-6}$ |
| *ground truth (MC)* | 525.28 | – |

(a)

| $n_\zeta$ | time (seconds) | MSE |
|---|---|---|
| 1000 | 3.07 | $4.54 \times 10^{-7}$ |
| 8000 | 8.04 | $7.63 \times 10^{-8}$ |
| 2048000 | 37.17 | $2.25 \times 10^{-8}$ |
| *ground truth (MC)* | 7340.97 | – |

(b)

**Table 6.2:** *Computation times and approximation errors (MSE) for critical-point probabilities in the results depicted in Fig. 6.34. Values are given for the (a) the wind field and (b) the blood flow field, both modeled using exponential correlation functions.*

The main advantage of the surrogate function approach is the significant reduction of computation times compared to MC integration. Depending on the data set the computation can be accelerated by up to 3 orders of magnitude. Further acceleration is expected to be possible using parallelization of the computations. The approach opens up the possibility for the user to employ probabilistic feature extraction during exploratory data visualization which was otherwise prevented by the high computational cost of MC sampling. Compared to previous approximation methods surrogate functions are a general approach that can be employed for many types of features. In particular, it allows to quickly estimate critical point probabilities which was not possible using previous approaches. Another advantage is that surrogate functions provide *controlled* approximations of the feature probabilities i.e. the approximation error can be systematically reduced by adding more examples to the training sets.

## 6.5 Model Selection for Discrete Random Fields

### 6.5.1 Spatial Correlation

In the results, the impact of spatial correlation on the feature probabilities is clearly visible. In Fig. 6.35 the impact of changing covariance between two adjacent grid points on the level-crossing probability is shown for two

pairs of input distributions with unit standard deviation. As the covariance increases, the crossing probability slightly decreases for the first case and decreases significantly for the second case. Recall that high covariance $\mathrm{Cov}(Y_1, Y_2)$ means that, for example, a positive deviation from the mean of a realization of $Y_1$ also implies a positive deviation of a realization of $Y_2$. For that reason, the effect of decreasing probability is larger for the second case while for the first case the effect is smaller because of the different mean values of the Gaussians.

Fig. 6.5 shows that increasing correlation between the grid points decreases the probabilities surrounding the mean crisp surface which leads to thinner spatial distributions of uncertain isocontours. As we can see in Fig. 6.7 neglecting correlation leads to overestimation of the uncertain isolines' spatial distributions. The distribution in Fig. 6.7b is thinner than the distribution in Fig. 6.7d and it is similar to Fig. 6.7c which we regard as a ground truth. In Fig 6.8 the probabilities for the cuboid case show thinner spatial distributions compared to the LCP-method described in Chap. 5 that neglects correlation.

Neglecting spatial correlations for the estimation of feature probabilities in vector fields has two notable effects. First, the ability to distinguish between critical point types is reduced, and second, probabilities are significantly over-estimated, see Sect. 6.3.7. The reason for this is that neglecting correlation corresponds to a white-noise-model for the uncertainty. Disturbance of vector fields by white noise leads to a much larger number of critical points and thus higher cell-wise probabilities. Spatial covariances between adjacent samples are more important, the differences between fully uncor-

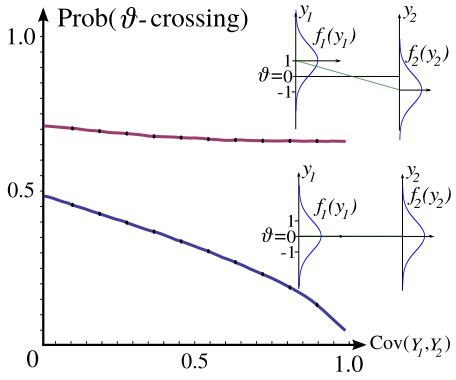



**Figure 6.35:** *The impact of changing covariance between two adjacent grid points on the level-crossing probability is shown for two pairs of input distributions with $\sigma_i = 1$. As the covariance increases the probability decreases slightly for the first case and significantly for the second case.*

related random variables and vector-wise correlated random variables are not that large. Localized high probability areas for critical points do not necessarily have counterparts in the mean vector field, as can be observed in Fig. 6.17 (b) and (c). It is therefore not sufficient to annotate features of the mean field with probabilities.

### 6.5.2 Parametric or Nonparametric Models?

The feature probabilities that were computed using the different probabilistic models show distinct quantitative and qualitative characteristics. In case empirical distributions are employed, the resulting probability fields are sparse and non-smooth. Empirical distributions that are used to compute level-crossing probabilities lead to results that are very similar to spaghetti plots, which are commonly used to visualize climate data. The usage of KDE leads to smooth probability fields that nicely capture the variability and the detailed structure of the features. The parametric model also leads to smooth results, but it introduces errors where the data is non-Gaussian. Depending on the data characteristics, the differences between KDE and the parametric Gaussian model can be significant or only subtle.

Selecting a suitable statistical model for a given dataset is crucial for achieving results that are neither biased, nor otherwise erroneous. The usage of empirical distributions for probabilistic feature extraction has the advantage that the results directly reflect the underlying numerical ensemble data. A disadvantage is that, depending on the condition number of the feature extraction problem, the position of features with non-zero probability can be unstable, cf. Sect. 4.2. The actual ensemble members are arbitrary in the sense that they are just one possible sampling of the underlying distribution and a different set of samples would be valid as well. Additionally, the results may suggest that the uncertain field is modelled by discrete probability distributions when the quantities in fact vary according to continuous distributions. KDE methods are used to get smooth reconstructions of the continuous distributions underlying a dataset and they can represent skewed or multimodal data. Jones compared CDFs that were computed using kernel smoothing to empirical distribution functions, which can be regarded as a special case with bandwidth $H = 0$ [Jon90]. His investigation revealed that *any* smoothing $H_u > H > 0$ (for some upper bound $H_u$) decreases the integrated mean squared error. Parametric models are constructed by determining optimal estimates for a fixed set of parameters from a dataset. While the estimated parameters converge for increasing sample sizes, the error that can be introduced by incorrectly specifying the model (for a phenomenon that is really non-Gaussian) cannot be removed by taking more samples. Statistical tests should be used to measure the goodness of fit.

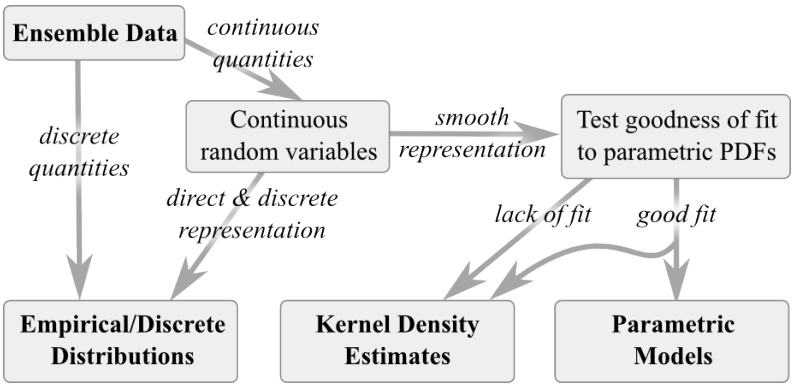



**Figure 6.36:** *Schematic diagram for the model selection task for ensemble data.*

The aim should be to obtain a good balance between goodness of fit and model simplicity. Empirical distributions are a simple model that is useful if the aim is the direct representation of ensemble data or if fast computation times are required. Parametric models should be used if a reasonable fit to the data can be ascertained. KDE is the best choice for datasets that are known to describe continuous phenomena but for which no assumptions about a fixed type of parametric distribution can be made. However, with KDE it is also possible to accurately approximate parametric distributions; this makes it an all-purpose tool for continuous random variables. The process of model selection for ensemble data is summarized in a schematic in Fig. 6.36.

# 7 Conclusions & Outlook

For the interpretation of scientific data, the uncertainties associated with it must be taken into account. While quantitative indications of uncertainty are ubiquitous in science and engineering (provided in tables and 1D plots for example), uncertainty cannot be adequately represented using standard feature extraction and visualization methods for 2D and 3D fields in most cases.

This thesis introduced methods to analyze field data that is afflicted with uncertainty. Our approaches facilitate the estimation of the propagation of uncertainty from the raw data to derived quantities and features that are important for the interpretation of the data. All approaches are well-founded in probability theory and statistics.

As mathematical model for uncertain scalar, vector and tensor fields, we employed discrete *random fields*. In this framework, a variety of probabilistic models can be employed. Different models have proven to be useful for various applications. In the simplest case we assume that all random variables in a field conform to some type of parametric probability distribution, e.g. Gaussian, and are statistically independent. More complex models consider fields with *arbitrary spatial correlations* and/or *nonparametric distributions* (empirical distributions, histograms and kernel density estimates (KDE)). In contrast to several previous approaches that were restricted to Gaussian fields our approach is more flexible and able to work with various types of distributions. The models described in Chap. 3 ensure that correct (consistent) marginal distributions can be obtained from the random field in order to perform local evaluation of the data. For KDE, we proposed an approach to perform a principal component transformation in order to efficiently capture correlations and use automatic bandwidth selection. We used these models a basis for local feature extraction.

We introduced the concept of *condition numbers* to feature-based visualization (Chap. 4). Specifically, we applied it to isocontour extraction from

scalar fields and to anisotropy index computation from diffusion tensor (DT) fields to examine the sensitivity to uncertainties in the input data. Using condition numbers of the isocontour problem the amplification or attenuation of uncertainty can be locally estimated independently of the contouring algorithm. The average condition numbers of isocontours has been shown to aid the selection of thresholds that correspond to robust isocontours. Previous work on DT-MRI [PXH*99, HAN04, CKPB07] had shown that fractional anisotropy ($FA$) yields higher $SNR$ than relative anisotropy ($RA$), i.e., that $FA$ is more immune to noise than $RA$. However, when extracting isosurfaces from the anisotropy fields our condition analysis in Sect. 4.3 indicates that the propagation of uncertainty from the DT eigenvalues to the isosurface position is *approximately equal* for $FA$ and $RA$. This equality was shown both analytically using a first order approximation and empirical results.

Based on the probabilistic models for uncertain fields, we derived statistically founded point-wise measures for the spatial distribution of uncertain isocontours in *continuous domains* (Chap. 5). Assuming that the random variables are uncorrelated, we spatially interpolate probability density functions between the grid points and define two quantities, the *isocontour density* and the *level-crossing probability field*. The measures are employed in interactive visualization methods for uncertain 2D and 3D volume data. For the 3D case, the introduced quantitative measures are used as procedural transfer functions in GPU assisted ray casting. No preprocessing is necessary and interactive frame rates are achieved.

In Chap. 6, we presented a general framework for the extraction of probabilistic local features from uncertain scalar and vector fields. In contrast to the approach presented in Chap. 5, this method associates feature probabilities to each *grid cell* of a discrete random field and not to all points in a continuous domain. This enables (i) the definition of any type of feature using indicator functions and (ii) the consideration of the spatial correlation structure. The feature probabilities are computed using Monte Carlo (MC) integration of the local multivariate marginal distributions.

Using this generic framework, we devised methods to estimate *level-crossing probabilities* in scalar random fields (Sect. 6.2), i.e., probabilities for the existence of an isocontour in given grid cells. Due to the consideration of correlation, this approach leads to more accurate results compared to Chap. 5. The probabilistic procedure does not have to deal with degenerate or ambiguous cases separately like it is the case for marching cubes and related algorithms. Probabilities for the occurrence of level crossings for critical isovalues or non-Morse functions (with respect to the mean values, for example) are computed correctly without treating any special cases.

For vector-valued random fields, we defined probabilistic equivalents to critical points and cores of swirling motion (Sect. 6.3). Our results indicate

significant differences in the spatial locality of features and show that the consideration of correlation is essential for obtaining correct results. In contrast to previous global methods [OGHT10,OGT11a,OGT11b], we take a different perspective on the topic. Our local method is able to extract features even in divergence-free fields and to detect saddle points in a straightforward way. It works on different grid types including surface vector fields.

To overcome the high computational cost of the MC integration, we introduced fast approximation methods for the estimation of feature probabilities. For cell-wise crossing probabilities in Gaussian fields, we proposed methods based on univariate and bivariate distribution functions that can be evaluated in the rendering step using lookup tables. A more general approach is the construction of *surrogate functions* from training data. The functions map the attributes describing the probability distributions that correspond to grid cells to feature probabilities. This way, the computation times are reduced by multiple orders of magnitude compared to MC integration. We demonstrated the utility of surrogate functions based on nonparametric $\mathcal{K}$-NN regression by showing good generalization performance for unobserved data.

The methods were applied to scalar, vector and tensor fields from engineering, medicine, and climate research. Several features that are useful for domain-specific analysis were detected, e.g., probabilities for the existence of critical points in the wall shear stress field of simulated blood flow in an aneurysm. Regions of low wall shear stress have been linked to rupture risk for specific types of cerebral aneurysms [GSK*12]. Comparing the results computed from multiple probabilistic models, we found some significant differences regarding spatial distribution, smoothness and the ability to represent subtle details in the data. This highlights the importance of careful model selection depending on the respective application as discussed in Sect. 6.5.

Challenges for future work mainly lie in two areas. The first area is the extension of the framework to features that are defined *globally*. The big impact of correlations observed with the local method raises the question whether the results of global methods, if extended to consider correlations, would also be affected significantly. Here, the numerical solution of stochastic differential equations based on correlated noise will be key. Global features that are defined in terms of stochastic paths, similar to the approach by Otto et al. [OGHT10], can possibly be extended to consider correlations but should be carefully studied regarding the convergence of the resulting distributions. The methods presented in this thesis are formulated for additive perturbations only. To account for multiplicative noise, e.g., as speckle noise, or for other phenomena that can not be modelled as additive noise, an adapted approach will be necessary. Another challenge is the extension

to data that is inherently multi-scale and, e.g., defined on hierarchies of discretizations (multi-grids).

The evaluation of the perceptual effectiveness of visualization methods for uncertainty in 2D and 3D data is beyond the scope of this thesis and constitutes a second area of further research. To investigate the influence of various visual representations on inference and decision making, formal user studies are required. MacEachren et al. provided a survey of previous work about the utility of uncertainty visualizations on decision making [MRH*05]. The authors also provide a list of challenges that should be considered for visual design. However, the majority of the papers and approaches address specific tasks in geography and cartography. A more general analysis and empirical evaluation of visualization methods for various application domains will be highly useful to enable better understanding of uncertain data.

# Appendices

# A

# Basics of Random Variables and Probability Distributions

In this appendix we revisit some fundamental concepts of probability theory that are necessary for the formulation of the uncertainty models in Chap. 3. For a more detailed exposition, refer to probability textbooks, e.g. [Fel71].

## A.1 Events

An *event* $E$ consists of a set of outcomes of a non-deterministic experiment with an associated probability $P(E)$. The set of all possible outcomes is called *sample space* $\Omega$ which is finite or countable in case of discrete phenomena and uncountably infinite in case of continuous phenomena. The probabilities associated to all events of a sample space must satisfy the Kolmogorov axioms [For08, p. 91].

## A.2 Random Variables

A random variable $Y$ is a variable that does not have a fixed, single value. It can take on different all possible values that are elements of the sample space $\Omega$, i.e. the value is subject to random variations. This randomness can be used as a mathematical representation of uncertainty about the true value of a physical quantity, e.g., due to measurement uncertainty and/or quantification errors. Lists of random variables describing multi-valued phenomena are called *multivariate random variables* or *random vectors*.

## A.3 Probability Distributions

Obviously, in most casis not all realizations $y \in \Omega$ of $Y$ are equally likely. The representation of the *distribution* of probabilities for all possible outcomes depends on whether $Y$ is a discrete or continous random variable.

**Probability Mass Functions (PMFs).** For a discrete random variable $Y$, a PMF is a functions that gives the probabilities for the event that $Y$ is equal to some value. For example, if

$$Y \sim \text{Bernulli}(p), \tag{A.1}$$

then the PMF is

$$f(y, p) = \begin{cases} p & \text{if } y = 1, \\ 1 - p & \text{if } y = 0. \end{cases} \tag{A.2}$$

The probability for the event that $Y = 1$ is

$$P(Y = 1) = 1 - (P = 0) = p. \tag{A.3}$$

**Probability Density Functions (PDFs).** In case $Y$ is a continuous random variable a different approach is necessary because there are uncountably infinite outcomes and each single-point outcome has probability zero. Instead of a probability a *probability density* is assigned to each value using a PDF. Integrating over subsets of the sample space (e.g. intervals) gives the probability that the random variable take a value from this subset. For example for a random variable $Y \in \mathbb{R}$ with an associated PDF $f$ we can compute the probability that $Y$ is contained in an interval using

$$P(a \leq Y \leq b) = \int_a^b f(y)\,\mathrm{d}y. \tag{A.4}$$

The function

$$F(y) = \int_{-\infty}^{y} f(v)\,\mathrm{d}v \tag{A.5}$$

gives the probability that $Y$ is less or equal to $y$ and is called *cumulative distribution function* (CDF) of $Y$. For example if $Y \geq 0$, $\lambda > 0$ and

$$Y \sim \text{Exponetial}(\lambda), \tag{A.6}$$

then the PDF is

$$f(y) = \lambda \exp(-\lambda\, y) \tag{A.7}$$

and the CDF is

$$F(y) = 1 - \exp(-\lambda\, y). \tag{A.8}$$

In case $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ is a *random vector* then the joint distribution can be described by a multivariate PDF

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{Y}}(y_1, y_2, \ldots, y_n). \tag{A.9}$$

The multivariate CDF as well as probabilities related to $\mathbf{Y}$ can be computed using $n$-dimensional integration over $f_{\mathbf{Y}}$.

## A.4 Marginals of Multivariate Gaussian Distributions

In this section we determine the marginal distribution of a multivariate Gaussian. The marginal is needed to obtain correct local cell-wise distributions in discrete random fields as discussed in Chap. 3 and 6. We mainly follow the approach of Bishop [Bis06, p. 88] but the proof that the marginal is again a Gaussian that can be expressed using the partitioned mean vector and covariance matrix of the joint distribution is a well known result and can also be found in, e.g. [MKB79, p. 63] or [Sun04, p. 23].

Let $Y = (y_1, y_2, \ldots, y_n)^T$ be a $n$-dimensional random vector that conforms to a multivariate normal distribution

$$Y \sim \mathcal{N}(\mu, \Sigma). \tag{A.10}$$

The density is given by

$$f(\mathbf{y}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y}-\mu)^T\Sigma^{-1}(\mathbf{y}-\mu)\right). \tag{A.11}$$

We can rewrite the vector as

$$Y = (Y_a, Y_b)^T, \tag{A.12}$$

where $Y_a = (y_1, y_2, \ldots, y_m)^T$ and $Y_b = (y_{m+1}, \ldots, y_n)^T$ are components of $Y$. We are now interested in the marginal distribution of $Y_a$. That means that we want to compute the density

$$f(\mathbf{y}_a) = \int f(\mathbf{y}_a, \mathbf{y}_b)\, \mathrm{d}\mathbf{y}_b. \tag{A.13}$$

We separate the mean vector as $\mu = (\mu_a, \mu_B)^T$, the covariance matrix $\Sigma$ such that

$$\Sigma = \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix} \tag{A.14}$$

and the precision matrix $\Lambda = \Sigma^{-1}$ such that

$$\Lambda = \begin{bmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{bmatrix}. \tag{A.15}$$

Then we can rewrite the quadratic terms in the exponent of the Gaussian as

$$\begin{aligned} -\frac{1}{2}(\mathbf{y}-\mu)^T\Sigma^{-1}(\mathbf{y}-\mu) \quad &= \\ &-\frac{1}{2}(\mathbf{y}_a-\mu_a)^T\Lambda_{aa}(\mathbf{y}_a-\mu_a) - \frac{1}{2}(\mathbf{y}_a-\mu_a)^T\Lambda_{ab}(\mathbf{y}_b-\mu_b) \\ &-\frac{1}{2}(\mathbf{y}_b-\mu_b)^T\Lambda_{ba}(\mathbf{y}_a-\mu_a) - \frac{1}{2}(\mathbf{y}_b-\mu_b)^T\Lambda_{bb}(\mathbf{y}_b-\mu_b). \end{aligned} \tag{A.16}$$

We want to integrate out $\mathbf{y}_b$ so we first consider the terms involving $\mathbf{y}_b$. Let

$$\mathbf{m} = \Lambda_{bb}\mu_b - \Lambda_{ba}(\mathbf{y}_a - \mu_a),$$

then the terms of Eq. (A.16) involving $\mathbf{y}_b$ can be written as

$$\begin{aligned} -\frac{1}{2}\mathbf{y}_b^T\Lambda_{bb}\mathbf{y}_b + \mathbf{y}_b^T\mathbf{m} = \\ -\frac{1}{2}(\mathbf{y}_b - \Lambda_{bb}^{-1}\mathbf{m})^T\Lambda_{bb}(\mathbf{y}_b - \Lambda_{bb}^{-1}\mathbf{m}) + \frac{1}{2}\mathbf{m}^T\Lambda_{bb}^{-1}\mathbf{m}. \end{aligned} \tag{A.17}$$

This way the dependence on $\mathbf{y}_b$ is transformed into the standard quadratic form of a multivariate Gaussian PDF and one term that does not depend on $\mathbf{y}_b$. We can exponentiate the quadratic form and see that the integration to compute Eq. (A.13) will take the form

$$\int \exp\left(-\frac{1}{2}(\mathbf{y}_b - \Lambda_{bb}^{-1}\mathbf{m})^T\Lambda_{bb}(\mathbf{y}_b - \Lambda_{bb}^{-1}\mathbf{m})\right) \mathrm{d}\mathbf{y}_b$$

which is an unnormalized Gaussian. Thus, the result of this integral is the reciprocal of the normalization constant which depends on the determinant of the covariance matrix but neither on the mean vectors nor $\mathbf{y}_a$ and $\mathbf{y}_b$. We can complete the square and integrate out $\mathbf{y}_b$ and the only term depending on $\mathbf{y}_a$ that remains is $\frac{1}{2}\mathbf{m}^T\Lambda_{bb}^{-1}\mathbf{m}$. We combine this with the other terms from Eq. (A.16) which depend on $\mathbf{y}_a$ and plug in $\mathbf{m}$ to obtain

$$\begin{aligned} &\frac{1}{2}(\Lambda_{bb}\mu_b - \Lambda_{ba}(\mathbf{y}_a-\mu_a))^T\Lambda_{bb}^{-1}(\Lambda_{bb}\mu_b - \Lambda_{ba}(\mathbf{y}_a-\mu_a)) \\ &-\frac{1}{2}\mathbf{y}_a^T\Lambda_{aa}\mathbf{y}_a + \mathbf{y}_a^T(\Lambda_{aa}\mu_a + \Lambda_{ab}\mu_b) + \text{const.} \\ = \quad &-\frac{1}{2}\mathbf{y}_a^T(\Lambda_{aa} - \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba})\mathbf{y}_a + \mathbf{y}_a^T(\Lambda_{aa} - \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba})^{-1}\mu_a + \text{const.} \end{aligned} \tag{A.18}$$

where 'const.' represents quantities that do not depend on $\mathbf{y}_a$. Using the fact that the quadratic form in the exponent of any Gaussian can be written as

$$-\frac{1}{2}(\mathbf{y}-\mu)^T\Sigma^{-1}(\mathbf{y}-\mu) = -\frac{1}{2}\mathbf{y}^T\Sigma^{-1}\mathbf{y} + \mathbf{y}^T\Sigma^{-1}\mu + \text{const.}$$

we can see from Eq. (A.18) that

$$\Sigma_{\mathbf{y}_a} = (\Lambda_{aa} - \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba})^{-1} \tag{A.19}$$

and

$$\mu_a = \Sigma_{aa}(\Lambda_{aa} - \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba})\mu_a. \tag{A.20}$$

Using matrix identities for Eq. (A.14) and (A.15) we can confirm [Bis06, p. 89] that

$$\text{cov}(\mathbf{y}_a) = \Sigma_{aa} = \Sigma_{\mathbf{y}_a}. \tag{A.21}$$

Similarly the expected value is

$$\text{E}(\mathbf{y}_a) = \mu_a. \tag{A.22}$$

That means that the marginal density $f(\mathbf{y}_a)$ is again a Gaussian distribution that can be expressed using the partitioned covariance matrix and mean vector of $f(\mathbf{y}_a, \mathbf{y}_b)$.

# B

# The Approximate Distribution Induced by the Linked-Pairs Approximation

In the following, we show that the approximate distribution for $\tilde{\mathbf{Y}}$ is again a multivariate normal distribution, and derive a formula for computing the covariance matrix of that approximate distribution. Without loss of generality, distributions with zero mean are assumed. Pairwise correlations are iteratively added to a multivariate distribution, yielding again another multivariate distribution that serves as input for the next step. By re-ordering random variables, we always extend the last variable of a random vector.

Given a $n$-dimensional multivariate normal distribution of the random vector $\mathbf{Y} = [Y_1, Y_2, \ldots, Y_n]$ with covariance matrix $\Sigma = (\rho_{ij}\sigma_i\sigma_j)_{1\leq i,j\leq n}$, variances $\sigma_i^2$ and $\rho_{ii} = 1$, we consider a second random vector $\tilde{\mathbf{Y}} = [Y_n, Y_{n+1}]$ with a two-dimensional normal distribution that describes the extension of $\mathbf{Y}$ by another variable. The covariance matrix of $\tilde{\mathbf{Y}}$ is

$$\Sigma_{\tilde{\mathbf{Y}}} = \begin{pmatrix} \sigma_n^2 & \rho_{n,n+1}\sigma_n\sigma_{n+1} \\ \rho_{n,n+1}\sigma_n\sigma_{n+1} & \sigma_{n+1}^2 \end{pmatrix}. \tag{B.1}$$

Covariances between the first $n-1$ variables and $n+1$ are not explicitly stated. In the following, the PDF of the joint distribution of the two distributions is computed.

With $f_{\mathbf{Y}}(y_1, \ldots, y_n)$, the PDF describing $\mathbf{Y}$, and $f_{y_{n+1}}(y_{n+1}|y_n)$ the PDF of the conditional distribution of $y_{n+1}$ given $y_n$, the joint PDF is given by the product of both PDFs, as $f_{y_{n+1}}$ is independent from $y_1 \ldots y_{n-1}$:

$$f(y_1, \ldots, y_{n+1}) = f_{\mathbf{Y}}(y_1, \ldots, y_n) f_{y_{n+1}}(y_{n+1}|y_n) \tag{B.2}$$

The $n$ dimensional multivariate normal distribution is given by:

$$f_{\mathbf{Y}}(y_1,\ldots,y_n) = \frac{1}{(2\pi)^{n/2}\,|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(y_1\ldots y_n)\Sigma^{-1}(y_1\ldots y_n)^T\right) \quad \text{(B.3)}$$

The conditional density for $y_{n+1}$ given $y_n$ is again normally distributed with mean $\bar{\mu} = \alpha y_n$, where $\alpha = \rho_{n,n+1}\frac{\sigma_{n+1}}{\sigma_n}$ and variance $\bar{\sigma}^2 = \sigma_{n+1}^2(1-\rho_{n,n+1}^2)$. It is then

$$f_{y_{n+1}}(y_{n+1}|y_n) = \frac{1}{\sqrt{2\pi\bar{\sigma}^2}} \exp\left(-\frac{(y_{n+1}-\alpha y_n)^2}{2\bar{\sigma}^2}\right), \quad \text{(B.4)}$$

and can be written in matrix form as

$$f_{y_{n+1}}(y_{n+1}|y_n) = \frac{1}{\sqrt{2\pi\bar{\sigma}^2}} \exp\left(-\frac{1}{2}(y_n y_{n+1}) \begin{pmatrix} \alpha^2/\bar{\sigma}^2 & -\alpha/\bar{\sigma}^2 \\ -\alpha/\bar{\sigma}^2 & 1/\bar{\sigma}^2 \end{pmatrix} \begin{pmatrix} y_n \\ y_{n+1} \end{pmatrix}\right). \quad \text{(B.5)}$$

Using the following identity for the inverse of the $(n+1)\times(n+1)$ matrix and employing block matrix notation

$$\left[\begin{pmatrix} \Sigma^{-1} & 0 \\ & \vdots \\ 0 & \cdots \quad 0 \end{pmatrix} + \begin{pmatrix} 0 & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & 0 \\ 0 & \cdots & 0 & \alpha^2/\bar{\sigma}^2 & -\alpha/\bar{\sigma}^2 \\ 0 & \cdots & 0 & -\alpha/\bar{\sigma}^2 & 1/\bar{\sigma}^2 \end{pmatrix}\right]^{-1} = \begin{pmatrix} \Sigma & \alpha\Sigma_n \\ \alpha\Sigma_n^T & \alpha^2\Sigma_{n,n}+\bar{\sigma}^2 \end{pmatrix},$$

where $\Sigma_n$ is the $n$th column of $\Sigma$ and $\Sigma_{n,n}$ the entry at $(n,n)$ of $\Sigma$, the joint distribution $f$ can be written as

$$\begin{aligned} f(y_1,\ldots y_{n+1}) &= \frac{1}{(2\pi)^{\frac{n+1}{2}}\,|\Sigma|^{\frac{1}{2}}\,|\bar{\sigma}^2|^{\frac{1}{2}}} \\ &\exp\left(-\frac{1}{2}(y_1\ldots y_{n+1}) \begin{pmatrix} \Sigma & \alpha\Sigma_n \\ \alpha\Sigma_n^T & \alpha^2\Sigma_{n,n}+\bar{\sigma}^2 \end{pmatrix}^{-1} \begin{pmatrix} y_1 \\ \vdots \\ y_{n+1} \end{pmatrix}\right). \quad \text{(B.6)} \end{aligned}$$

With $\alpha^2\Sigma_{n,n}+\bar{\sigma}^2 = \sigma_{n+1}^2$ and

$$\Sigma^* = \begin{pmatrix} \Sigma & \alpha\Sigma_n \\ \alpha\Sigma_n^T & \sigma_{n+1}^2 \end{pmatrix} \quad \text{(B.7)}$$

the joint distribution is finally

$$f(y_1, \dots y_{n+1}) = \frac{1}{(2\pi)^{\frac{n+1}{2}} |\Sigma^*|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(y_1 \dots y_{n+1})\Sigma^{*-1}(y_1 \dots y_{n+1})^T\right). \tag{B.8}$$

It is left to show that $|\Sigma\bar{\sigma}^2| = |\Sigma^*|$. Applying the block matrix formula for determinants

$$\left|\begin{pmatrix} A & B \\ C & D \end{pmatrix}\right| = |A||D - CA^{-1}B| \tag{B.9}$$

to Eq. (B.7) yields the desired equality:

$$\begin{aligned} |\Sigma^*| &= |\Sigma||\sigma_{n+1}^2 - \alpha\Sigma_n^T\Sigma^{-1}\alpha\Sigma_n| \\ &= |\Sigma||\sigma_{n+1}^2 - \alpha^2\Sigma_n^T(0 \cdots 01)^T| \\ &= |\Sigma||\sigma_{n+1}^2 - \alpha^2\Sigma_{nn}| \\ &= |\Sigma||\sigma_{n+1}^2 - \alpha^2\sigma_n^2| \\ &= |\Sigma||\sigma_{n+1}^2(1 - \rho_{n,n+1}^2)| \\ &= |\Sigma||\bar{\sigma}^2| \\ &= |\Sigma\bar{\sigma}^2|. \end{aligned} \tag{B.10}$$

The last column of Eq. (B.7) contains the covariances between $Y_1, \cdots Y_n$ and the last variable $Y_{n+1}$. The covariances are

$$\alpha\Sigma_N = (\rho_{i,n}\rho_{n,n+1}\sigma_i\sigma_{n+1})_{1 \le i \le n}, \tag{B.11}$$

i.e., the correlations to variable $Y_n$ are multiplied by $\rho_{n,n+1}$ for variable $Y_{n+1}$.

# C

# Condition Numbers of Anisotropy Isosurface Computation

Let $D(x,y)$ be a diffusion tensor field in $\mathbb{R}^2$ where each tensor is described by its eigenvalues $\lambda_i$ and eigenvectors $\mathbf{e}_i$. The fractional anisotropy is given by

$$FA = \sqrt{\frac{1}{2}}\sqrt{\frac{(\lambda_1-\lambda_2)^2}{\lambda_1^2+\lambda_2^2}} \tag{C.1}$$

and the relative anisotropy by

$$RA = \sqrt{\frac{1}{2}}\frac{\sqrt{(\lambda_1-\lambda_2)^2}}{\lambda_1+\lambda_2}. \tag{C.2}$$

With the derivatives

$$\frac{\partial FA}{\partial \lambda_1} = \frac{\lambda_2\left(\lambda_1^2-\lambda_2^2\right)}{|\lambda_1-\lambda_2|\left(\lambda_1^2+\lambda_2^2\right)^{3/2}}$$

and

$$\frac{\partial FA}{\partial \lambda_2} = \frac{\lambda_1\lambda_2^2-\lambda_1^3}{|\lambda_1-\lambda_2|\left(\lambda_1^2+\lambda_2^2\right)^{3/2}}$$

we can determine the absolute normwise condition for $FA$ computation

$$\kappa_{FA}^{abs} = \left|\left|\left(\frac{\partial FA}{\partial \lambda_1}, \frac{\partial FA}{\partial \lambda_2}\right)^T\right|\right| = \frac{|\lambda_1+\lambda_2|}{\lambda_1^2+\lambda_2^2}\,. \tag{C.3}$$

Analogously, for $RA$ we can write

$$\frac{\partial RA}{\partial \lambda_1} = \frac{2\left(\lambda_1 - \lambda_2\right)\lambda_2}{\left|\lambda_1 - \lambda_2\right|\left(\lambda_1 + \lambda_2\right)^2}$$

$$\frac{\partial RA}{\partial \lambda_2} = \frac{2\lambda_1\left(\lambda_2 - \lambda_1\right)}{\left|\lambda_1 - \lambda_2\right|\left(\lambda_1 + \lambda_2\right)^2}$$

$$\kappa_{RA}^{abs} = \left|\left|\left(\frac{\partial RA}{\partial \lambda_1}, \frac{\partial RA}{\partial \lambda_2}\right)^T\right|\right| = \frac{2\sqrt{\lambda_1^2 + \lambda_2^2}}{\left(\lambda_1 + \lambda_2\right)^2}. \tag{C.4}$$

For an explicit formulation of Eq. (4.16) we use the gradient

$$\nabla FA = \left(\frac{\partial FA}{\partial x}, \frac{\partial FA}{\partial y}\right)^T \tag{C.5}$$

with

$$\frac{\partial FA}{\partial x} = -\frac{\left(\lambda_1^2 - \lambda_2^2\right)\left(\lambda_1 \frac{\partial \lambda_2}{\partial x} - \lambda_2 \frac{\partial \lambda_1}{\partial x}\right)}{\left|\lambda_1 - \lambda_2\right|\left(\lambda_1^2 + \lambda_2^2\right)^{3/2}} \tag{C.6}$$

and

$$\frac{\partial FA}{\partial y} = -\frac{\left(\lambda_1^2 - \lambda_2^2\right)\left(\lambda_1 \frac{\partial \lambda_2}{\partial y} - \lambda_2 \frac{\partial \lambda_1}{\partial y}\right)}{\left|\lambda_1 - \lambda_2\right|\left(\lambda_1^2 + \lambda_2^2\right)^{3/2}}, \tag{C.7}$$

leading to the condition number for isosurface extraction

$$\kappa_{FA^{-1}(\vartheta)}^{abs} = \sqrt{\frac{\left(\left|\frac{\left(\lambda_1^2-\lambda_2^2\right)\left(\lambda_1 \frac{\partial \lambda_2}{\partial y} - \lambda_2 \frac{\partial \lambda_1}{\partial y}\right)}{\left(\lambda_1^2+\lambda_2^2\right)^2}\right|^2 + \left|\frac{\left(\lambda_1^2-\lambda_2^2\right)\left(\lambda_1 \frac{\partial \lambda_2}{\partial x} - \lambda_2 \frac{\partial \lambda_1}{\partial x}\right)}{\left(\lambda_1^2+\lambda_2^2\right)^2}\right|^2\right)\left(\lambda_1^2 + \lambda_2^2\right)}{\left(\lambda_1 - \lambda_2\right)^2}}. \tag{C.8}$$

Similarly for RA

$$\frac{\partial RA}{\partial x} = -\frac{2\left(\lambda_1 - \lambda_2\right)\left(\lambda_1 \frac{\partial \lambda_2}{\partial x} - \lambda_2 \frac{\partial \lambda_1}{\partial x}\right)}{\left|\lambda_1 - \lambda_2\right|\left(\lambda_1 + \lambda_2\right)^2} \tag{C.9}$$

and

$$\frac{\partial RA}{\partial y} = -\frac{2\left(\lambda_1 - \lambda_2\right)\left(\lambda_1 \frac{\partial \lambda_2}{\partial y} - \lambda_2 \frac{\partial \lambda_1}{\partial y}\right)}{\left|\lambda_1 - \lambda_2\right|\left(\lambda_1 + \lambda_2\right)^2} \tag{C.10}$$

give the condition number

$$\kappa^{abs}_{RA^{-1}(\vartheta)} = \frac{2\sqrt{\left|\lambda_2 \frac{\partial \lambda_1}{\partial y} - \lambda_1 \frac{\partial \lambda_2}{\partial y}\right|^2 + \left|\lambda_2 \frac{\partial \lambda_1}{\partial x} - \lambda_1 \frac{\partial \lambda_2}{\partial x}\right|^2}}{\left(\lambda_1 + \lambda_2\right)^2}. \tag{C.11}$$

By elementary algebra it can be shown that the relation

$$\kappa^{abs}_{FA}\kappa^{abs}_{FA^{-1}(\vartheta)} = \kappa^{abs}_{RA}\kappa^{abs}_{RA^{-1}(\vartheta)} \tag{C.12}$$

holds. This means that in a first order approximation the propagation of uncertainties from the eigenvalues to uncertainties of isocontour-positions is equal for FA and RA.

# Bibliography


[AB08] Allendes Osorio R., Brodlie K.: Contouring with uncertainty. In *Theory and Practice of Computer Graphics 2008 – Eurographics UK Chapter Proceedings* (2008), Lim I. S., Tang W., (Eds.), Eurographics Association, pp. 59–66.

[Abr97] Abrahamsen P.: *A review of Gaussian random fields and correlation functions*. Norsk Regnesentral/Norwegian Computing Center, 1997.

[Adl81] Adler R. J.: *The Geometry of Random Fields*. John Wiley & Sons, Chichester, 1981.

[Alt92] Altman N.: An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician 46*, 3 (1992), 175–185.

[AMN*98] Arya S., Mount D. M., Netanyahu N. S., Silverman R., Wu A. Y.: An optimal algorithm for approximate nearest neighbor searching fixed dimensions. *Journal of the ACM (JACM) 45*, 6 (1998), 891–923.

[And01] Anderson A. W.: Theoretical analysis of the effects of noise on diffusion tensor imaging. *Magnetic Resonance in Medicine 46*, 6 (December 2001), 1174–1188.

[AOB09] Allendes Osorio R. S., Brodlie K. W.: Uncertain Flow Visualization using LIC. In *7th EG UK Theory and Practice of Computer Graphics: Proceedings. 7th Theory and Practice of Computer Graphics* (2009), Tang W., Collomosse J., (Eds.), Eurographics Association Eurographics UK Chapter.

[AT07] Adler R. J., Taylor J.: *Random Fields and Geometry*. Springer, Berlin, 2007.

[ATW09] Adler R. J., Taylor J. E., Worsley K. J.: *Applications of Random Fields and Geometry*. Preprint, Technion-Israel Institute of Technology Haifa, 2009.

[AW09] Azaïs J.-M., Wschebor M.: *Level Sets and Extrema of Random Processes and Fields*. John Wiley & Sons, Ch. 6, Chichester, 2009.

[BAOL12] Brodlie K., Allendes Osorio R., Lopes A.: A review of uncertainty in data visualization. In *Expanding the Frontiers of Visual Analytics and Visualization*. Springer, 2012, pp. 81–109.

[Bis06] Bishop C. M.: *Pattern recognition and machine learning*. Springer, New York, 2006.

[BJB*12] Bhatia H., Jadhav S., Bremer P.-T., Chen G., Levine J. A., Nonato L. G., Pascucci V.: Flow visualization with quantified spatial and temporal errors using edge maps. *IEEE Transactions on Visualization and Computer Graphics 18*, 9 (2012), 1383–1396.

[BL07] Bottou L., Lin C.-J.: Support vector machine solvers. In *Large Scale Kernel Machines*, Bottou L., Chapelle O., DeCoste D., Weston J., (Eds.). MIT Press, Cambridge, MA., 2007, pp. 1 – 28.

[BLS05] Bergmann O., Lundervold A., Steihaug T.: Generating a synthetic diffusion tensor dataset. In *Computer-Based Medical Systems, 2005. Proceedings. 18th IEEE Symposium on* (june 2005), pp. 277–281.

[BM58] Box G. E. P., Muller M. E.: A Note on the Generation of Random Normal Deviates. *The Annals of Mathematical Statistics 29*, 2 (June 1958), 610–611.

[BPS97] Bajaj C. L., Pascucci V., Schikore D. R.: The contour spectrum. In *Proceedings of IEEE Visualization 1997* (Los Alamitos, CA, USA, 1997), IEEE Computer Society Press, pp. 167–174.

[BRM*08] Boussel L., Rayz V., McCulloch C., Martin A., Acevedo-Bolton G., Lawton M., Higashida R., Smith W. S., Young W. L., Saloner D.: Aneurysm growth occurs at region of low wall shear stress: patient-specific correlation of hemodynamics and growth in a longitudinal study. *Stroke 39*, 11 (2008), 2997–3002.

[BVPtHR09] Brecheisen R., Vilanova A., Platel B., ter Haar Romeny B.: Parameter sensitivity visualization for DTI fiber tracking. *IEEE Transactions on Visualization and Computer Graphics 15* (2009), 1441–1448.

[BWE05] Botchen R. P., Weiskopf D., Ertl T.: Texture-based visualization of uncertainty in flow fields. In *Proceedings of IEEE Visualization 2005* (2005), IEEE, pp. 647–654.

[CHH04] Chen S., Hong X., Harris C. J.: Sparse kernel density construction using orthogonal forward regression with leave-one-out test score and local regularization. *IEEE Transactions on Systems, Man, and Cybernetics Part B-Cybernetics 34*, 4 (2004), 1708–1717.

[CKPB07] Chang L. C., Koay C. G., Pierpaoli C., Basser P. J.: Variance of estimated DTI-derived parameters via first-order perturbation methods. *Magnetic Resonance in Medicine 57*, 1 (January 2007), 141–149.

[CO02] Csató L., Opper M.: Sparse on-line gaussian processes. *Neural Computation 14*, 3 (2002), 641–668.

[DH03] Deuflhard P., Hohmann A.: *Numerical Analysis in Modern Scientific Computing: An Introduction*. Springer-Verlag New York, Inc., 2003.

[DKLP01] Djurcilov S., Kim K., Lermusiaux P. F., Pang A.: Volume rendering data with uncertainty information. In *Data Visualization – Joint Eurographics-IEEE TCVG Symposium on Visualization*. Springer, 2001, pp. 243–252.

[DKLP02] Djurcilov S., Kim K., Lermusiaux P., Pang A.: Visualizing scalar volumetric data with uncertainty. *Computers & Graphics 26*, 2 (2002), 239–248.

[DP00] Djurcilov S., Pang A.: Visualizing sparse gridded data sets. *IEEE Computer Graphics and Applications 20*, 5 (2000), 52–57.

[DP01] Dubois D., Prade H.: Possibility theory, probability theory and multiple-valued logics: A clarification. *Annals of mathematics and Artificial Intelligence 32*, 1-4 (2001), 35–66.

[DW13] Demir I., Westermann R.: Progressive high-quality response surfaces for visually guided sensitivity analysis. *Computer Graphics Forum 32*, 3pt1 (2013), 21–30.

[EK06] Ennis D. B., Kindlmann G.: Orthogonal tensor invariants and the analysis of diffusion tensor magnetic resonance images. *Magnetic Resonance in Medicine 55*, 1 (2006), 136–146.

[FCHW99] Firbank M., Coulthard A., Harrison R., Williams E.: A comparison of two methods for measuring the signal to noise ratio on MR images. *Physics in Medicine and Biology 44*, 12 (1999), N261–N264.

[Fed69] Federer H.: *Geometric Measure Theory*. Springer, New York, 1969.

[Fel71] Feller W.: *Introduction to Probability Theory and its Applications*. John Wiley & and Sons, 1968 and 1971. Vol. 1 and 2.

[FFW06] Friman O., Farneback G., Westin C.-F.: A Bayesian approach for stochastic white matter tractography. *TMI 25*, 8 (2006), 965–978.

[FHH*11] Friman O., Hennemuth A., Harloff A., Bock J., Markl M., Peitgen H.-O.: Probabilistic 4d blood flow tracking and uncertainty estimation. *Med. Image. Anal. 15*, 5 (2011), 720–728.

[Fil09] Filler A. G.: MR neurography and diffusion tensor imaging: Origins, history & clinical impact of the first 50,000 cases with an assessment of efficacy and utility in a prospective 5,000 patient study group. *Neurosurgery 65*, 4 Suppl (2009), A29–A43.

[FKLT10] Feng D., Kwock L., Lee Y., Taylor R.: Matching visual saliency to confidence in plots of uncertain data. *IEEE Transactions on Visualization and Computer Graphics 16* (2010), 980–989.

[FLS63] Feynman R. P., Leigthon R. B., Sands M.: *The Feynman Lectures on Physics*, vol. 1. Caltech, Caltech, 1963, ch. 6 Probability, p. 10.

[For08] Fornasini P.: *The Uncertainty in Physical Measurements: An Introduction to Data Analysis in the Physics Laboratory*. Springer, Berlin, 2008.

[GCD*10] Gorissen D., Couckuyt I., Demeester P., Dhaene T., Crombecq K.: A surrogate modeling and adaptive sampling toolbox for computer based design. *The Journal of Machine Learning Research 11* (2010), 2051–2055.

[Gen04] Gentle J. E.: *Random Number Generation and Monte Carlo Methods*, 2nd ed. Springer, New York, 2004.

[GR04] Grigoryan G., Rheingans P.: Point-based probabilistic surfaces to show surface uncertainty. *IEEE Transactions on Visualization and Computer Graphics 10*, 5 (2004), 564–573.

[GSK*12] Goubergrits L., Schaller J., Kertzscher U., van den Bruck N., Pöthkow K., Petz C., Hege H.-C., Spuler A.: Statistical wall shear stress maps of ruptured and unruptured middle cerebral artery aneurysms. *J. R. Soc. Interface 9*, 69 (2012), 677 – 688.

[GST14] Günther D., Salmon J., Tierny J.: Mandatory critical points of 2d uncertain scalar fields. *Computer Graphics Forum 33*, 3 (June 2014), 31–40.

[GTS04] Garth C., Tricoche X., Scheuermann G.: Tracking of vector field singularities in unstructured 3d time-dependent datasets. In *Proc. IEEE Vis. 2004* (2004), pp. 329–336.

[HAN04] Hasan K. M., Alexander A. L., Narayana P. A.: Does fractional anisotropy have better noise immunity characteristics than relative anisotropy in diffusion tensor MRI? An analytical approach. *Magnetic Resonance in Medicine 51*, 2 (2004), 413–417.

[HH89] Helman J. L., Hesselink L.: Representation and display of vector field topology in fluid flow data sets. *Computer 22* (August 1989), 27–36.

[HLNW11] Hlawatsch M., Leube P., Nowak W., Weiskopf D.: Flow radar glyphs - static visualization of unsteady flow with uncertainty. *IEEE Transactions on Visualization and Computer Graphics 17*, 12 (2011), 1949–1958.

[HSNHH10] Hotz I., Sreevalsan-Nair J., Hagen H., Hamann B.: Tensor field reconstruction based on eigenvector and eigenvalue interpolation. In *Scientific Visualization: Advanced Concepts* (2010), Hagen H., (Ed.), vol. 1 of *Dagstuhl Follow-Ups*, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Germany, pp. 110–123.

[Hun87] Hunt J. C. R.: Vorticity and vortex dynamics in complex turbulent flows. In *Canadian Society for Mechanical Engineering, Transactions* (1987), vol. 11, pp. 21–35.

[JKJTM06] Jankun-Kelly M., Jiang M., Thompson D., Machiraju R.: Vortex visualization for practical engineering applications. *IEEE Transactions on Visualization and Computer Graphics 12*, 5 (2006), 957–964.

[JMS96] Jones M., Marron J., Sheather S.: A brief survey of bandwidth selection for density estimation. *J. Am. Stat. Assoc. 91*, 433 (1996), 401–407.

[Job91] Jobson J.: *Applied Multivariate Data Analysis Volume 1: Regression and Experimental Design*. Springer, New York, 1991.

[Joi08] Joint Committee for Guides in Metrology: *Uncertainty of measurement – Part 3: Guide to the expression of uncertainty in measurement (GUM)*. International Organization for Standardization, Geneva, 2008.

[Jon90] Jones M.: The performance of kernel density functions in kernel distribution function estimation. *Stat. Probabil. Lett. 9*, 2 (1990), 129–132.

[Jon03] Jones D. K.: Determining and visualizing uncertainty in estimates of fiber orientation from diffusion tensor MRI. *Magnetic Resonance in Medicine 49* (2003), 7–12.

[JS03] Johnson C. R., Sanderson A. R.: A next step: Visualizing errors and uncertainty. *IEEE Computer Graphics and Applications 23*, 5 (2003), 6–10.

[JWSK07] Jänicke H., Wiebel A., Scheuermann G., Kollmann W.: Multifield visualization using local statistical complexity. *IEEE Transactions on Visualization and Computer Graphics 13* (2007), 1384–1391.

[Kal02] Kallenberg O.: *Foundations of modern probability*. Springer, 2002.

[Kan98] Kanwal R. P.: *Generalized Functions: Theory and Technique*. Birkhäuser, Basel, 1998.

[KCPB07] Koay C. G., Chang L.-C., Pierpaoli C., Basser P. C. J.: Error propagation framework for diffusion tensor imaging via diffusion tensor representations. *IEEE Transactions on Medical Imaging 26*, 8 (2007), 1017–1034.

[KMKS13] Kinkeldey C., Mason J., Klippel A., Schiewe J.: Assessing the impact of design decisions on the usability of uncertainty visualization: Noise annotation lines for the visual representation of attribute uncertainty. In *Proceedings of 26th International Cartographic Conference* (2013), ICA.

[KUS*05] Kniss J. M., Uitert R. V., Stephens A., Li G.-S., Tasdizen T., Hansen C.: Statistically quantitative volume visualization. In *Proceedings of IEEE Visualization 2005* (October 2005), pp. 287–294.

[KW03] Krüger J., Westermann R.: Acceleration techniques for GPU-based volume rendering. In *Proceedings IEEE Visualization 2003* (2003), pp. 287–292.

[KWTM03] Kindlmann G., Whitaker R., Tasdizen T., Möller T.: Curvature-based transfer functions for direct volume rendering: Methods and applications. In *Proceedings of IEEE Visualization 2003* (October 2003), pp. 513–520.

[LA03] Lazar M., Alexander A. L.: An error analysis of white matter tractography methods: synthetic diffusion tensor field simulations. *NeuroImage 20*, 2 (October 2003), 1140–1153.

[LBMP*01] Le Bihan D., Mangin J.-F., Poupon C., Clark C. A., Pappata S., Molko N., Chabriat H.: Diffusion tensor imaging: Concepts and applications. *Journal of Magnetic Resonance Imaging 13*, 4 (2001), 534–546.

[LC87] Lorensen W. E., Cline H. E.: Marching cubes: A high resolution 3D surface construction algorithm. *SIGGRAPH Computer Graphics 21*, 4 (1987), 163–169.

[LFC02] Lodha S. K., Faaland N. M., Charaniya A. P.: Visualization of uncertain particle movement. In *Proceedings of the Computer Graphics and Imaging Conference* (2002), pp. 226–232.

[LHZP07] Laramee R., Hauser H., Zhao L., Post F.: Topology-based flow visualization, the state of the art. *Topology-based Methods in Visualization* (2007), 1–19.

[Lir02] Lira I.: *Evaluating the Measurement Uncertainty: Fundamentals and Practical Guidance*. Institute of Physics Publishing, Bristol, 2002.

[LLBP12] Liu S., Levine J. A., Bremer P.-T., Pascucci V.: Gaussian mixture model based volume visualization. In *Large Data Analysis and Visualization (LDAV), 2012 IEEE Symposium on* (2012), IEEE, pp. 73–77.

[LLPY07] Lundstrom C., Ljung P., Persson A., Ynnerman A.: Uncertainty visualization in medical volume rendering using probabilistic animation. *IEEE Transactions on Visualization and Computer Graphics 13*, 6 (2007), 1648–1655.

[Lod08] Lodwick W. A.: *Fuzzy surfaces in GIS and geographical analysis: theory, analytical methods, algorithms and applications*. CRC Press, Boca Raton, 2008.

[LPK05] Love A. L., Pang A., Kao D. L.: Visualizing spatial multivalue data. *IEEE Computer Graphics and Applications 25* (2005), 69–79.

[LPSW96] Lodha S. K., Pang A., Sheehan R. E., Wittenbrink C. M.: Uflow: visualizing uncertainty in fluid flow. In *Proceedings IEEE Visualization 1996* (1996), pp. 249–254.

[LSMC11] Lee V. S., Steinbach L., Mukherji S., Celsohyginocruz L. (Eds.): *Magnetic Resonance Imaging Clinics of North America (special issue on Diffusion Imaging) 19, 1*. 2011.

[Mat02] Matsumoto Y.: *An Introduction to Morse Theory*. American Mathematical Society, 2002.

[Mil63] Milnor J. W.: *Morse Theory*. Princeton Univ Press, 1963.

[MKB79] Mardia K. V., Kent J. T., Bibby J. M.: *Multivariate analysis*. Academic press, 1979.

[MLP*10] McLoughlin T., Laramee R., Peikert R., Post F., Chen M.: Over two decades of integration-based, geometric flow visualization. In *Computer Graphics Forum* (2010), vol. 29:6, pp. 1807–1829.

[MME05] Moore D. S., McCabe G. P., Evans M. J.: *Introduction to the Practice of Statistics*. W. H. Freeman & Co., New York, NY, USA, 2005.

[MRH*05] MacEachren A., Robinson A., Hopper S., Gardner S., Murray R., Gahegan M., Hetzler E.: Visualizing geospatial information uncertainty: What we know and what we need to know. *Cartography and Geographic Information Science 32*, 3 (2005), 139–161.

[Mur01] Murphy D. B.: *Fundamentals of Light Microscopy and Electronic Imaging*. Wiley-Liss, New York, 2001.

[NY06] Newman T. S., Yi H.: A survey of the marching cubes algorithm. *Computers & Graphics 30*, 5 (2006), 854 – 879.

[OGHT10] Otto M., Germer T., Hege H.-C., Theisel H.: Uncertain 2D vector field topology. *Computer Graphics Forum 29* (2010), 347–356.

[OGT11a] Otto M., Germer T., Theisel H.: Closed stream lines in uncertain vector fields. In *Proc. Spring Conference on Computer Graphics (SCCG)* (2011).

[OGT11b] Otto M., Germer T., Theisel H.: Uncertain topology of 3d vector fields. In *Proceedings of 4th IEEE Pacific Visualization Symposium (PacificVis 2011)* (Hong Kong, China, March 2011), pp. 67–74.

[ONK03] Ong Y. S., Nair P. B., Keane A. J.: Evolutionary optimization of computationally expensive problems via surrogate modeling. *AIAA journal 41*, 4 (2003), 687–696.

[OT12] Otto M., Theisel H.: Vortex analysis in uncertain vector fields. *Computer Graphics Forum 31*, 3 (2012), 1035–1044.

[Pal04] Palmer T. N. et al.: *Development of a European Multi-Model Ensemble System for Seasonal to Inter-Annual Prediction (DEMETER)*. Technical memorandum, European Centre for Medium-Range Weather Forecasts, Reading, England, 2004.

[PB96] Pierpaoli C., Basser P. J.: Toward a quantitative assessment of diffusion anisotropy. *Magnetic Resonance in Medicine 36*, 6 (December 1996), 893–906.

[PB03] Pajevic S., Basser P. J.: Parametric and non-parametric statistical analysis of DT-MRI data. *Journal of Magnetic Resonance 161*, 1 (March 2003), 1–14.

[PH11] Pöthkow K., Hege H.-C.: Positional uncertainty of isocontours: Condition analysis and probabilistic measures. *IEEE Transactions on Visualization and Computer Graphics 17*, 10 (2011), 1393–1406.

[PH12] Pöthkow K., Hege H.-C.: Uncertainty Propagation in DT-MRI Anisotropy Isosurface Extraction. In *New Developments in the Visualization and Processing of Tensor Fields*, Laidlaw D., Vilanova A., (Eds.), Mathematics and Visualization. Springer, Berlin, 2012, pp. 209 – 225.

[PH13] Pöthkow K., Hege H.-C.: Nonparametric models for uncertainty visualization. *Computer Graphics Forum 32*, 3 (2013), 131 – 140.

[PH14] Pöthkow K., Hege H.-C.: Accelerated probabilistic feature extraction using surrogate functions. *in preparation* (2014).

[PKXJ12] Potter K., Kirby R., Xiu D., Johnson C.: Interactive visualization of probability and cumulative density functions. *International Journal for Uncertainty Quantification 2*, 4 (2012), 397–412.

[PMG04] Pauly M., Mitra N., Guibas L.: Uncertainty and variability in point cloud surface data. In *Eurographics Symposium on Point-Based Graphics* (2004), pp. 77–84.

[PMW13] Pfaffelmoser T., Mihai M., Westermann R.: Visualizing the variability of gradients in uncertain 2d scalar fields. *IEEE Transactions on Visualization and Computer Graphics 19*, 11 (2013), 1948 – 1961.

[PP02] Polthier K., Preuss E.: Identifying vector field singularities using a discrete hodge decomposition. In *Visualization and Mathematics III* (2002), Springer, pp. 112–134.

[PPH12] Petz C., Pöthkow K., Hege H.-C.: Probabilistic local features in uncertain vector fields with spatial correlation. *Computer Graphics Forum 31*, 3 (2012), 1045 – 1054.

[PPH13] Pöthkow K., Petz C., Hege H.-C.: Approximate level-crossing probabilities for interactive visualization of uncertain isocontours. *International Journal for Uncertainty Quantification 3:2* (2013), 101–117.

[PRH10] Prassni J.-S., Ropinski T., Hinrichs K.: Uncertainty-aware guided volume segmentation. *IEEE Transactions on Visualization and Computer Graphics 16* (2010), 1358–1365.

[PRJ12] Potter K., Rosen P., Johnson C.: From quantification to visualization: A taxonomy of uncertainty visualization approaches. *International Journal for Uncertainty Quantification* (2012), 226–249.

[Prü18] Prüfer H.: Neuer Beweis eines Satzes über Permutationen. *Arch. Math. Phys. 27* (1918), 742–744.

[PRW11] Pfaffelmoser T., Reitinger M., Westermann R.: Visualizing the positional and geometrical variability of isosurfaces in uncertain scalar fields. *Computer Graphics Forum 30*, 3 (2011), 951–960.

[PT88] Palais R. S., Terng C.-L.: *Critical Point Theory and Submanifold Geometry*, vol. 1353, Lecture Notes in Mathematics. Springer-Verlag, 1988.

[PVH*03] Post F., Vrolijk B., Hauser H., Laramee R., Doleisch H.: The state of the art in flow visualisation: Feature extraction and tracking. In *Computer Graphics Forum* (2003), vol. 22:4, pp. 775–792.

[PW12] Pfaffelmoser T., Westermann R.: Visualization of global correlation structures in uncertain 2d scalar fields. *Computer Graphics Forum 31*, 3 (2012), 1025–1034.

[PW13a] Pfaffelmoser T., Westermann R.: Correlation visualization for structural uncertainty analysis. *International Journal for Uncertainty Quantification 3*, 2 (2013).

[PW13b] Pfaffelmoser T., Westermann R.: Visualizing contour distributions in 2d ensemble data. In *EuroVis-Short Papers* (2013), The Eurographics Association, pp. 55 – 59.

[PWB*09] Potter K., Wilson A., Bremer P.-T., Williams D., Doutriaux C., Pascucci V., Johhson C. R.: Ensemble-vis: A framework for the statistical visualization of ensemble data. In *IEEE Workshop on Knowledge Discovery from Climate Data: Prediction, Extremes.* (2009), pp. 233–240.

[PWH01] Pekar V., Wiemker R., Hempel D.: Fast detection of meaningful isosurfaces for volume data visualization. In *Proceedings of IEEE Visualization 2001* (Washington, DC, USA, 2001), IEEE Computer Society, pp. 223–230.

[PWH11] Pöthkow K., Weber B., Hege H.-C.: Probabilistic marching cubes. *Computer Graphics Forum 30*, 3 (2011), 931 – 940.

[PWL97] Pang A. T., Wittenbrink C. M., Lodha S. K.: Approaches to uncertainty visualization. *The Visual Computer 13*, 8 (1997), 370–390.

[PXH*99] Papadakis N., Xing D., Houston G., Smith J., Smith M., James M., Parsons A., Huang C.-H., Hall L., Carpenter T.: A study of rotationally invariant and symmetric indices of diffusion anisotropy. *Magnetic Resonance Imaging 17* (1999), 881–892(12).

[QHS*05] Queipo N. V., Haftka R. T., Shyy W., Goel T., Vaidyanathan R., Kevin Tucker P.: Surrogate-based analysis and optimization. *Progress in aerospace sciences 41*, 1 (2005), 1–28.

[RB09] Ryu Y.-H., Baik J.-J.: Flow and dispersion in an urban cubical cavity. *Atmospheric Environment 43*, 10 (2009), 1721 – 1729.

[RLBS03] Rhodes P. J., Laramee R. S., Bergeron R. D., Sparr T. M.: Uncertainty visualization methods in isosurface rendering. In *Eurographics 2003, Short Papers* (2003), pp. 83–88.

[Sco92] Scott D. W.: *Multivariate Density Estimation: Theory, Practice, and Visualization (Wiley Series in Probability and Statistics)*. Wiley, 1992.

[SH95] Sujudi D., Haimes R.: *Identification of swirling flow in 3D vector fields*. Tech. rep., Dept. of Aeronautics and Astronautics, MIT, Cambridge, MA, 1995.

[SHM10] Saad A., Hamarneh G., Moller T.: Exploration and visualization of segmentation uncertainty using shape and appearance prior information. *IEEE Transactions on Visualization and Computer Graphics 16* (2010), 1366–1375.

[Sil92] Silverman B.: *Density estimation for statistics and data analysis*. Chapman & Hall/CRC, 1992.

[SJK04] Sanderson A. R., Johnson C. R., Kirby R. M.: Display of vector fields using a reaction-diffusion model. In *Proceedings of IEEE Visualization 2004* (2004), pp. 115–122.

[SKS12] Schlegel S., Korn N., Scheuermann G.: On the interpolation of data with normally distributed uncertainty for visualization. *IEEE Transactions on Visualization and Computer Graphics 18*, 12 (2012), 2305–2314.

[SPB07] Snook L., Plewes C., Beaulieu C.: Voxel based versus region of interest analysis in diffusion tensor imaging of neurodevelopment. *NeuroImage 34*, 1 (2007), 243 – 252.

[SS04] Scott D. W., Sain S. R.: Multi-dimensional density estimation. In *Handbook of Statistics—Vol 23: Data Mining and Computational Statistics*, Rao C. R., Wegman E. J., (Eds.). Elsevier, Amsterdam, 2004, pp. 229–263.

[SSD*08] Scheidegger C. E., Schreiner J. M., Duffy B., Carr H., Silva C. T.: Revisiting histograms and isosurface statistics. *IEEE Transactions on Visualization and Computer Graphics 14* (2008), 1659–1666.

[SSSSW13] Schultz T., Schlaffke L., Schölkopf B., Schmidt-Wilcke T.: HiFiVE: A Hilbert space embedding of fiber variability estimates for uncertainty modeling and visualization. *Computer Graphics Forum 32*, 3 (2013).

[STS07] Schultz T., Theisel H., Seidel H.-P.: Segmentation of DT-MRI anisotropy isosurfaces. In *EuroVis07: Joint Eurographics -*

*IEEE VGTC Symposium on Visualization 2007* (Norrköping, Sweden, May 2007), Museth K., Möller T., Ynnerman A., (Eds.), Eurographics, pp. 187–194.

[Sun04] Sung H. G.: *Gaussian mixture regression and classification*. PhD thesis, Rice University, Houston, TX, 2004.

[SWTH07] Sahner J., Weinkauf T., Teuber N., Hege H.-C.: Vortex and strain skeletons in Eulerian and Lagrangian frames. *IEEE Transactions on Visualization and Computer Graphics 13*, 5 (2007), 980–990.

[SZD*10] Sanyal J., Zhang S., Dyer J., Mercer A., Amburn P., Moorhead R.: Noodles: A tool for visualization of numerical weather model ensemble uncertainty. *IEEE Transactions on Visualization and Computer Graphics 16* (2010), 1421–1430.

[Szy11] Szymczak A.: Stable morse decompositions for piecewise constant vector fields on surfaces. *Computer Graphics Forum 30*, 3 (2011), 851–860.

[TG09] Tricoche X., Garth C.: Topological methods for visualizing vortical flows. In *Mathematical Foundations of Scientific Visualization, Computer Graphics, and Massive Data Exploration*, Farin G., Hege H.-C., Hoffman D., Johnson C. R., Polthier K., (Eds.), Mathematics and Visualization. Springer, 2009, pp. 89–107.

[TK94] Taylor B., Kuyatt C.: *Guidelines for Expressing and Evaluating the Uncertainty of NIST Experimental Results*. NIST tech. note 1297, 1994.

[TSP01] Thong J., Sim K., Phang J.: Single-image signal-to-noise ratio estimation. *Scanning 23*, 5 (2001), 328–336.

[Win08] Winitzki S.: A handy approximation for the error function and its inverse, 2008. lecture note, https://sites.google.com/site/winitzki/sergei-winitzkis-files/erf-approx.pdf?attredirects=0&d=1.

[WMK13] Whitaker R. T., Mirzargar M., Kirby R. M.: Contour boxplots: A method for characterizing uncertainty in feature sets from simulation ensembles. *Visualization and Computer Graphics, IEEE Transactions on 19*, 12 (2013), 2713–2722.

[WPL96] Wittenbrink C. M., Pang A. T., Lodha S. K.: Glyphs for visualizing uncertainty in vector fields. *IEEE Transactions on Visualization and Computer Graphics 2*, 3 (1996), 266–279.

[WSH03] Weber G. H., Scheuermann G., Hamann B.: Detecting critical regions in scalar fields. In *VisSym* (2003), Eurographics Association, pp. 85–94.

[WYM08] Wang C., Yu H., Ma K.-L.: Importance-driven time-varying data visualization. *IEEE Transactions on Visualization and Computer Graphics 14* (2008), 1547–1554.

[YXK13] Yang C., Xiu D., Kirby R. M.: Visualization of covariance and cross-covariance fields. *International Journal for Uncertainty Quantification 3*, 1 (2013), 25–38.

[Zad65] Zadeh L.: Fuzzy sets. *Information Control 8* (1965), 338–353.

[ZDG*08] Zuk T., Downton J., Gray D., Carpendale S., Liang J.: Exploration of uncertainty in bidirectional vector fields. In *Proc. SPIE & IS&T Conf. Electronic Imaging, Vol. 6809: Visualization and Data Analysis 2008* (2008), vol. 6809, SPIE, p. 68090B.

[ZMB*03] Zhukov L., Museth K., Breen D., Whitaker R., Barr A. H.: Level set modeling and segmentation of DT-MRI brain data. *Journal of Electronic Imaging 12* (2003), 125–133.

[Zuk08] Zuk T.: *Visualizing Uncertainty*. Ph.d. thesis, Department of Computer Science, University of Calgary, April 2008.

[ZWK10] Zehner B., Watanabe N., Kolditz O.: Visualization of gridded scalar data with uncertainty in geosciences. *Computers & Geosciences 36*, 10 (2010), 1268–1275.

# Selbständigkeitserklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht. Insbesondere habe ich nicht die Hilfe eines kommerziellen Promotionsberaters in Anspruch genommen. Dritte haben von mir weder unmittelbar noch mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen. Die Arbeit wurde bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form als Dissertation eingereicht und ist als Ganzes auch noch nicht veröffentlicht.

Berlin,

Kai Pöthkow