Correlation-Aware Uncertainty Visualization of Multivariate Data With Many-Core Acceleration

Alex Gorczowski
University of Illinois Urbana-Champaign
Urbana, USA
agorc@illinois.edu

Timbwaoga A. J. Ouermi
Oak Ridge National Laboratory
Oak Ridge, USA
ouermita@ornl.gov

Tushar M. Athawale
Oak Ridge National Laboratory
Oak Ridge, USA
athawaletm@ornl.gov

***Abstract*—Probabilistic feature extraction propagates data uncertainty into the multivariate features scientists inspect, but existing formulations either assume the variables are independent or rely on Monte Carlo integration too slow for production use. We present the first closed-form, correlation-aware estimators of trait probability for rectangular and arbitrary polygonal traits, together with a kernel density noise model for ensembles that parametric distributions cannot capture, implemented as portable many-core filters in Viskores. On a Red Sea ensemble of 2.5 million points evaluated on Frontier, they match Monte Carlo within its sampling noise while running up to 477 times faster.**

***Index Terms*—uncertainty visualization, multivariate data, correlation, kernel density estimation, many-core**

I. Introduction

Uncertainty is present in all simulation, observational, and experimental data, and neglecting it yields analysis that reports features with unearned confidence. Multivariate feature extraction is one such case. Given several scalar fields on a shared grid and a user-specified region of value space called a trait, fiber methods extract its spatial preimage, for example water that is simultaneously warm and salty. The probabilistic form of that query asks how likely a grid point's uncertain values lie inside the trait, replacing a hard classification with a probability field.

Two limitations restrict current formulations. Most assume the variables are independent, since independence factors the trait probability into a product of one-dimensional integrals. Real fields are coupled by physical laws and shared error sources, and discarding that coupling leaves estimates overconfident or underconfident. At one grid point of a temperature and salinity ensemble, assuming independence gives a trait probability of 0.22, while the measured correlation of -0.82 gives 0.37. Methods that do model correlation, and those using nonparametric noise models, fall back on Monte Carlo integration, whose error falls only as the inverse square root of the sample count and is prohibitive on million-point grids.

We contribute (1) a closed form for the correlated Gaussian trait probability over rectangular traits; (2) an eigen decomposition and Green's theorem framework that extends it to arbitrary polygonal traits; (3) a kernel density noise model evaluated through the same machinery; and (4) an implementation as many-core Viskores filters evaluated on Frontier.

This project was supported by the U.S. Department of Energy Office of Science, Advanced Scientific Computing Research, Early Career Research Program, and used resources of the Oak Ridge Leadership Computing Facility, under contract number DE-AC05-00OR22725.

II. Correlation-Aware Closed Forms

We model the uncertain values at each grid point as a correlated Gaussian with mean and covariance estimated from the ensemble, and define the trait probability as the integral of that density over the trait. For rectangular traits it reduces to differences of the bivariate normal cumulative distribution function, a standard special function we evaluate on device at fixed cost per point.

Polygonal traits are harder, since correlation places a cross term in the exponent that prevents the density from factoring, and factorization is what makes a closed-form antiderivative available. We transform the data and trait into the covariance eigenbasis, which removes that term, and use Green's theorem to reduce the resulting area integral to closed-form boundary computations. The result costs time linear in the vertex count per grid point and requires no discretization of value space.

III. Nonparametric Noise Model

Parametric Gaussians cannot represent multimodal ensembles or outlier members. For those we use a kernel density estimate (KDE), placing one Gaussian kernel on each member, with bandwidths obtained from the data by Silverman's rule, retaining the correlation carried by the member locations. Integration is linear, so the trait probability is the average of the per-kernel trait masses and every closed form above applies unchanged. On a synthetic trivariate ensemble with bimodal noise it reaches a root mean square error of 0.1792 against ground truth, versus 0.2144 for an independent Gaussian and 0.2523 for the statistical mean.

IV. Parallelization and Results

Grid points are independent, so the estimators map onto the Viskores worklet model and run unmodified across CUDA, HIP, SYCL, OpenMP, and TBB. We evaluate on Frontier with 20 Red Sea members on a $250 \times 200 \times 50$ grid and a polygonal trait in the curlZ and vorticity plane, against a Monte Carlo reference of 1000 samples per point.

TABLE I
FULL-FIELD EVALUATION TIME ON FRONTIER FOR THE CORRELATED MODELS

Device, model	Monte Carlo	Closed form	Speedup
CPU, Gaussian	3.95 s	8.3 ms	477×
CPU, KDE	4.66 s	306 ms	15.2×
GPU, Gaussian	128 ms	1.0 ms	124×
GPU, KDE	307 ms	44 ms	7.0×

The CPU is one 56-core AMD EPYC 7A53 and the GPU is one graphics compute die of an AMD MI250X.

Table I reports both correlated models. The closed forms are faster on every configuration and agree with the reference at every point to within its own sampling noise, which validates the derivations on production hardware.

V. CONCLUSION

Correlation changes the reported confidence of multivariate features, and these closed forms make accounting for it substantially cheaper than Monte Carlo integration. Delivering them as Viskores filters gives ParaView a path to correlation-aware uncertainty visualization on CPUs and GPUs. Future work extends the polygonal formulation beyond two variables and to topological features.