Visual Analysis of Large Multivariate Data using Clustering and Probabilistic Summaries Supplementary Material

Tobias Rapp, Christoph Peters, and Carsten Dachsbacher

1 3D Gaussian Ray Integration

As motivated in the paper, we integrate a trivariate Gaussian distribution along a ray $o + xd$ starting at $o \in \mathbb{R}^3$ in normalized direction $d \in \mathbb{R}^3$ with $x \in \mathbb{R}$. The Gaussian is given by its mean $\mu \in \mathbb{R}^3$ and covariance $\Sigma \in \mathbb{R}^{3\times 3}$. To derive a general solution, we integrate over $[a, b]$ by substituting the ray equation into the trivariate Gaussian distribution:

$$I(a,b) := \int_a^b \frac{1}{\sqrt{|2\pi\Sigma|}} \exp\left(-\frac{(o + xd - \mu)^\intercal \Sigma^{-1} (o + xd - \mu)}{2}\right) \mathrm{d}x. \tag{1}$$

Note that $|2\pi\Sigma| = (2\pi)^3|\Sigma|$ for trivariate Gaussians, which we prefer due to its compactness. We start by simplifying the equation:

$$\begin{aligned} I(a,b) &= \int_a^b \frac{1}{\sqrt{|2\pi\Sigma|}} \exp\left(-\frac{((o-\mu) + xd)^\intercal \Sigma^{-1} ((o-\mu) + xd)}{2}\right) \mathrm{d}x \\ &= \int_a^b \frac{1}{\sqrt{|2\pi\Sigma|}} \exp\left(-(o-\mu)^\intercal \Sigma^{-1}(o-\mu)\right) \exp\left(-2x \underbrace{\frac{1}{2}(o-\mu)^\intercal \Sigma^{-1} d}_{c_{o,d}}\right) \exp\left(-x^2 \underbrace{\frac{1}{2} d^\intercal \Sigma^{-1} d}_{c_{d,d}}\right) \mathrm{d}x \\ &= \frac{1}{\sqrt{|2\pi\Sigma|}} \exp\left(-(o-\mu)^\intercal \Sigma^{-1}(o-\mu)\right) \int_a^b \exp\left(-2x c_{o,d} - x^2 c_{d,d}\right) \mathrm{d}x. \end{aligned} \tag{2}$$

We substitute $r := x + \frac{c_{o,d}}{c_{d,d}}$ and thus rewrite and simplify the integrand as follows:

$$\begin{aligned} I(a,b) &= \frac{1}{\sqrt{|2\pi\Sigma|}} \exp\left(-(o-\mu)^\intercal \Sigma^{-1}(o-\mu)\right) \int_{a+\frac{c_{o,d}}{c_{d,d}}}^{b+\frac{c_{o,d}}{c_{d,d}}} \exp\left(-2\left(r - \frac{c_{o,d}}{c_{d,d}}\right) c_{o,d} - \left(r - \frac{c_{o,d}}{c_{d,d}}\right)^2 c_{d,d}\right) \mathrm{d}r \\ &= \frac{1}{\sqrt{|2\pi\Sigma|}} \exp\left(-(o-\mu)^\intercal \Sigma^{-1}(o-\mu)\right) \int_{a+\frac{c_{o,d}}{c_{d,d}}}^{b+\frac{c_{o,d}}{c_{d,d}}} \exp\left(-2r c_{o,d} + 2\frac{c_{o,d}^2}{c_{d,d}} - r^2 c_{d,d} + 2r c_{o,d} - \frac{c_{o,d}^2}{c_{d,d}}\right) \mathrm{d}r \\ &= \underbrace{\frac{1}{\sqrt{|2\pi\Sigma|}} \exp\left(-(o-\mu)^\intercal \Sigma^{-1}(o-\mu)\right) \exp\left(\frac{c_{o,d}^2}{c_{d,d}}\right)}_{c} \int_{a+\frac{c_{o,d}}{c_{d,d}}}^{b+\frac{c_{o,d}}{c_{d,d}}} \exp\left(-r^2 c_{d,d}\right) \mathrm{d}r. \end{aligned} \tag{3}$$

Now, we perform another substitution using $p := r\sqrt{c_{d,d}}$:

$$I(a,b) = c\frac{1}{\sqrt{c_{d,d}}} \int_{\sqrt{c_{d,d}}\left(a+\frac{c_{o,d}}{c_{d,d}}\right)}^{\sqrt{c_{d,d}}\left(b+\frac{c_{o,d}}{c_{d,d}}\right)} \exp(-p^2)\, \mathrm{d}p. \tag{4}$$

The integrand can now be expressed by its antiderivative, which leads us to the following solution:

$$I(a,b) = c\frac{1}{\sqrt{c_{d,d}}}\left[\frac{\sqrt{\pi}}{2}\operatorname{erf}(p)\right]_{\sqrt{c_{d,d}}\left(a+\frac{c_{o,d}}{c_{d,d}}\right)}^{\sqrt{c_{d,d}}\left(b+\frac{c_{o,d}}{c_{d,d}}\right)}. \tag{5}$$

Although no closed-form solution exists for the error function, fast and accurate numerical approximations of this well known function are available. Lastly, when integrating over $(-\infty, \infty)$, the error function disappears:

$$\left[\frac{\sqrt{\pi}}{2}\operatorname{erf}(p)\right]_{-\infty}^{\infty} = \sqrt{\pi},$$

which leads to

$$I(-\infty,\infty) = c\frac{\sqrt{\pi}}{\sqrt{c_{d,d}}}. \tag{6}$$

2 Fast Selection of GMM Components

In the main paper, we propose a fast selection of GMM components by formulating lower und upper bounds and using subsampling. In Figure 1, we illustrate the impact of this approximation on the Spray Nozzle (a) and the Hurricane Isabel (b) dataset compared to a brute-force selection of GMM components. Using only the bounds does not introduce a measurable error. However, the subsampling leads to a slightly increased error. Note that for subsampling we always take a fixed amount of 200 samples from each cluster. By increasing this fixed number of samples, we lower this error, at the cost of additional computational effort.

3 Additional Results

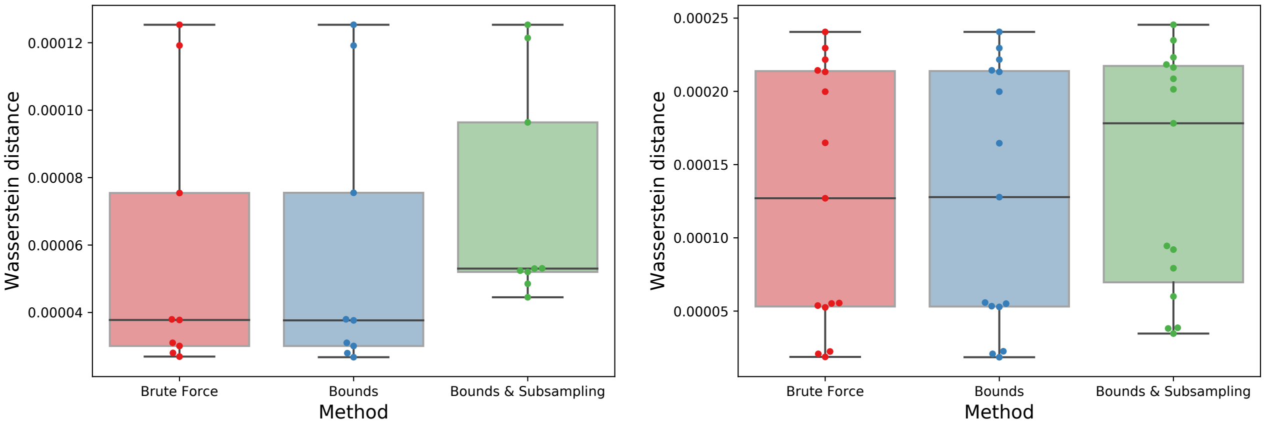
In this section, we show additional results for the datasets presented in the paper.

3.1 Synthetic Dataset

Figure 2 compares our data summaries to high-dimensional Gaussian mixture models. In (a), the mean Wasserstein distance of all 1D dimensions is shown. Although the error is comparable, the high-dimensional GMMs lead to a higher error for the more complex dimensions, such as the exponentially distributed dimension shown in (b) and (c).

3.2 Spray Nozzle

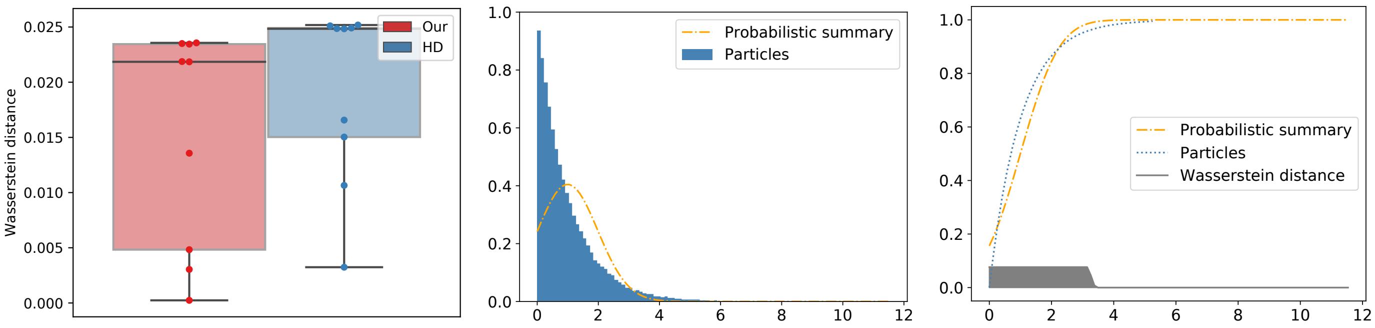
In Figure 3, we compare our data summaries to those of high-dimensional GMMs. In (a), the error of our and the high-dimensional data model is shown. The error is comparable between both approaches and low in absolute size. The table in (b) shows that the high-dimensional data model is smaller in size. However, the high-dimensional model requires extremely expensive preprocessing and the large amount of used GMM components affects the performance of all visualizations.



(a) Spray Nozzle (k=2000)

(b) Hurricane Isabel (k=1000)

Figure 1: Comparison of our fast selection of GMM components with the brute-force approach on the Spray Nozzle (a) and the Hurricane Isabel (b) dataset.

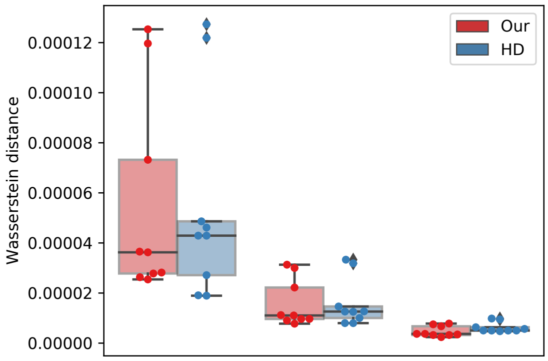


(a) Error of 1D dimensions

(b) Probability density function

(c) Cumulative distribution function

Figure 2: For the synthetic dataset, we compare our data summaries to high-dimensional GMMs (a). In (b), the exponentially distributed dimension from the high-dimensional GMMs is shown. The corresponding CDF in (c) indicates a high error.



Model	# Clusters	Size	GMM comp.	Wasserstein dist.
Our	2,000	6.7 MB	1.7 ± 1.19	5.54×10^{-5}
HD	2,000	3.7 MB	21.6 ± 3.8	5.50×10^{-5}
Our	8,000	19.9 MB	1.4 ± 0.85	1.57×10^{-5}
HD	8,000	9.2 MB	13.5 ± 3.5	1.59×10^{-5}
Our	32,000	47.5 MB	1.2 ± 0.61	4.64×10^{-6}
HD	32,000	13.6 MB	4.9 ± 2.86	6.24×10^{-6}

(a) Error of 1D dimensions

(b) Overview of the data models

Figure 3: We compare our data summaries to modeling high-dimensional GMMs.