Comprehending high dimensions and dimension reduction

Andrew Glaws

Abstract

High-dimensional spaces appear frequently in problems related to computational or data science. Unfortunately, we are three-dimensional beings trapped in a three-dimensional physical world. This makes it difficult for the average person to understand these complex spaces. Mathematics provides us with the tools to work in high-dimensions but leaves us lost without any intuition into the behavior of these spaces. This project attempts to impart that intuition by explaining and visualizing high-dimensional spaces with interactive tools.

1 High-dimensions in computer experiments

For this project, we focus on high-dimensions within the context of *computer experiments* [6, 8]. Computer experiments can supplement or replace physical experimentation when the latter is expensive, impossible, or immoral. Mathematically, we represent the underlying model within a computer experiment as a deterministic mapping from a vector-valued set of inputs to a scalar-valued output,

$$y \;=\; f(\mathbf{x}), \qquad y \in \mathbb{R}, \quad \mathbf{x} \in \mathbb{R}^m. \tag{1}$$

We generally assume that (1) is accompanied by a known input density function $p(\mathbf{x})$, which encodes uncertainty in the inputs of the computer experiment. One common choice for $p(\mathbf{x})$—and the one we assume for this project—is the uniform density over the $[-1,1]$ hypercube. A *hypercube* is the extension of a standard three-dimensional cube to arbitrary dimensions. If we consider each component of $\mathbf{x}$ to come from the interval $x_i \in [-1,1]$, then the vector $\mathbf{x}$ is an element of the m-dimensional $[-1,1]$ hypercube—denoted by $[-1,1]^m$. Mathematically, we write the uniform density function over this hypercube as

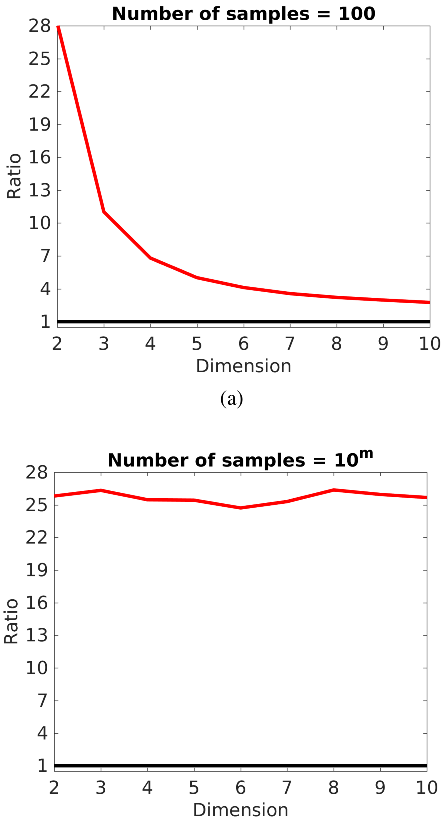
$$p(\mathbf{x}) \;=\; \begin{cases} 1/2^m & ||\mathbf{x}||_\infty \leq 1, \\ 0 & \text{otherwise}\,, \end{cases} \tag{2}$$

The length of the input vector m is referred to as the *dimension* of the problem. As the dimension grows, studying and understanding the behavior of f quickly becomes more difficult. For example, consider the k nearest neighbor (or knn) algorithm which is popular in many regression and classification machine learning techniques [4]. For a given point $\mathbf{x} \in \mathbb{R}^m$, this algorithm finds the k closest points from a set of M points $\{\mathbf{x}_i\}_{i=1}^M$. The knn algorithm is often used to approximate conditional moments and to cluster data. However, our traditional intuitions about distance fall apart as the dimension of our problem grows [1]. For a given point, consider the ratio of the distance to its furthest point over the distance to its closest point,

$$\mathscr{R}(\mathbf{x}) \;=\; \frac{\max_i ||\mathbf{x}-\mathbf{x}_i||}{\min_i ||\mathbf{x}-\mathbf{x}_i||}. \tag{3}$$

We can interpret this ratio as the relative importance of nearby points compared to far away points. When (3) is large, the closer points should be much more informative than far away points. When (3) is small (and note that $\mathscr{R}$ has a lower limit of 1), there is essentially no difference between the ”close” points and the ”far” points. In this case, the results from the k nearest neighbor algorithm is essentially meaningless.

Figure 1 contains plots of the ratio in (3) from a given point in m dimensions to M randomly distributed points. Figure 1a holds M constant as we increase the dimension. In this case, we get a rapid decay in $\mathscr{R}$. This suggests that furthest and close points to a given point are nearly the same distance away. In Figure 1b, we increase M exponentially with the dimension. This allows us to maintain a relatively constant $\mathscr{R}$.



(a)

(b)

Figure 1: Plots of the distance ratio in (3) as a function of dimension using (a) a constant number of samples and (b) an exponentially increasing number of samples.

In general, the exponential increase in computational costs with dimension is referred to as the *curse of dimensionality* [3, 10]. This curse can make a variety of numerical studies (e.g., optimization, uncertainty quantification, approximation) infeasible in high dimensions. Increases in computational power or improved algorithms can help mitigate the high cost associated with high dimensions; however, only dimension reduction address the curse at its source.

We focus on linear dimension reduction of the form

$$y \;=\; f(\mathbf{x}) \;=\; g(\boldsymbol{A}^\top \mathbf{x}), \tag{4}$$

where $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ for $n < m$. When a function exhibits the type of structure in (4), we call that function a *ridge function* [7]. Figure 2 shows a two-dimensional ridge function with one-dimensional ridge structure. The matrix $\boldsymbol{A}$ defines an orthogonal rotation of the input space which uncovers the low-dimensional structure of the function.

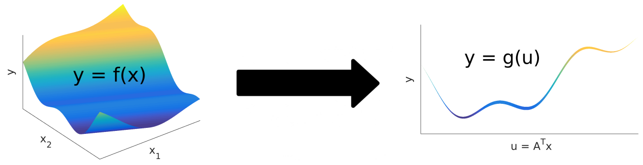


Figure 2: An example of a two-dimensional function with one-dimensional ridge structure. By applying the appropriate rotation to the input space, we see low-dimensional structure in the function.

Ridge functions appear as a component of a variety of regression and approximation techniques such as projection pursuit regression and neural networks. However, they can also be used directly for dimension reduction. If a computer experiment exhibits ridge-like structure, then it may be approximated by a ridge function,

$$y = f(\mathbf{x}) \approx g(\boldsymbol{A}^{\top}\mathbf{x}). \tag{5}$$

We can then more easily study the reduced ridge profile $g(\boldsymbol{A}^{\top}\mathbf{x})$ as an approximate surrogate of $f(\mathbf{x})$.

Figure 2 shows the full surface of a given ridge function. In general, we do not have this sort of insight into the models underlying our computer experiment. Instead, we have a sample of point evaluations $\{\mathbf{x}_i, f(\mathbf{x}_i)\}_{i=1}^{M}$. A *shadow plot* or *sufficient summary plot* is a scatter plot of this data projected along the appropriate ridge directions. Figure 3 is an example of a one-dimensional shadow plot. Variation in the plot is caused by variation in the $m-1$ directions orthogonal to $\mathbf{a}$. However, this variation is relatively small compared to the variation along the direction $\mathbf{a}$. Shadow plots play a key role in qualitatively evaluating the ridge structure within a given function.

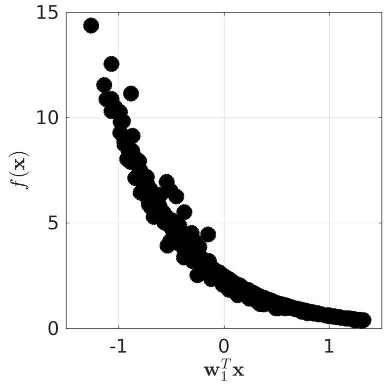


Figure 3: Example of a one-dimension shadow plot.

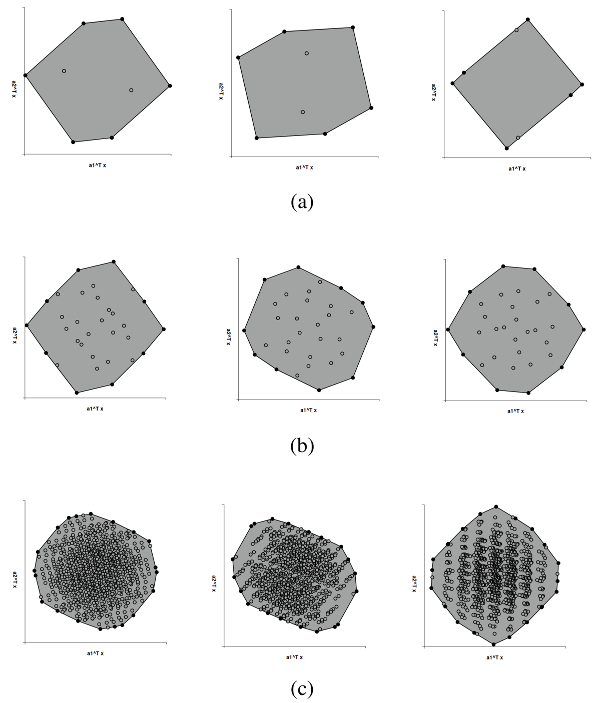
2 Zonotopes

A zonotope is the linear projection of an m-dimensional hypercube onto an n-dimensional space. Mathematically, we write this as

$$\mathscr{Z}(\boldsymbol{A}) = \left\{ \boldsymbol{A}^{\top}\mathbf{x} \,|\, \mathbf{x} \in [-1,1]^m \right\} \tag{6}$$

where $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ with $n < m$ [5]. We are interested in zonotopes within the context of dimension reduction and ridge functions. Thus, we assume that $\boldsymbol{A}$ has orthonormal columns such that the zonotope is the orthogonal rotation and projection of the hypercube. In (6), the zonotope $\mathscr{Z}(\boldsymbol{A})$ defines the domain of the ridge profile g.

The interactive page includes a two-dimensional zonotope graphic with variable dimension and adjustable rotation. Figure 4 contains several snapshots of various zonotopes. We see that the number of hypercube corners grows exponentially with dimension. However, most of these corners project down to the interior of the zonotope with just a few corners defining the boundary of $\mathscr{Z}(\boldsymbol{A})$. Working with the interactive tool, we can gain intuition into the way we can rotate high-dimensional objects.



(a)

(b)

(c)

Figure 4: Zonotopes of m-dimensional hypercubes randomly projected onto two-dimensional spaces with (a) $m = 3$, (b) $m = 5$, and (c) $m = 10$.

3 Active subspaces

One method for discovering ridge structure in high-dimensional functions is *active subspaces* [2]. Consider the matrix

$$\boldsymbol{C} = \int \nabla f(\mathbf{x}) \nabla f(\mathbf{x})^{\top} p(\mathbf{x}), d\mathbf{x}, \tag{7}$$

where $\nabla f(\mathbf{x}) \in \mathbb{R}^m$ is the gradient of f with respect to $\mathbf{x}$ and $p(\mathbf{x})$ is the input probability density function. The matrix $\boldsymbol{C}$ is symmetric and positive semi-definite. Thus, we can take the eigendecomposition

$$\boldsymbol{C} = \boldsymbol{W} \boldsymbol{\Lambda} \boldsymbol{W}^{\top} \tag{8}$$

and assume without loss of generality that the eigenvalues are in descending order $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 0$ and the eigenvalues are orthonormal.

Using the Rayleigh quotient, we get the property

$$\begin{aligned} \lambda_i &= \frac{\mathbf{w}_i^\top \boldsymbol{C} \mathbf{w}_i}{\mathbf{w}_i^\top \mathbf{w}_i} \\ &= \mathbf{w}_i^\top \left(\int \nabla f(\mathbf{x}) \nabla f(\mathbf{x})^\top \, p(\mathbf{x}), d\mathbf{x} \right) \mathbf{w}_i \\ &= \int \left(\mathbf{w}_i^\top \nabla f(\mathbf{x}) \right)^2 p(\mathbf{x}) \, d\mathbf{x}. \end{aligned} \tag{9}$$

Thus, the mean squared directional derivative of f along the direction of $\mathbf{w}_i$ is equal to the associated eigenvalue λ_i. If λ_i is large, then $\mathbf{w}_i$ is a relatively important direction in the input space with respect to changes in f. If λ_i is small, the $\mathbf{w}_i$ is a relatively unimportant direction in the input space.

To illustrate the use of active subspaces in uncovering ridge structure, we examine a standard test problem in uncertainty quantification—the piston cycle problem [9]. This function models the cycle time of a piston in a cylinder with respect to seven physical inputs,

$$y = 2\pi \sqrt{\frac{M}{k + S^2 \frac{P_0 V_0}{T_0} \frac{T_a}{V^2}}}, \tag{10}$$

where

$$\begin{aligned} V &= \frac{S}{2k} \left(\sqrt{A^2 + 4k \frac{P_0 V_0}{T_0} T_a} - A \right), \\ A &= P_0 S + 19.62 M - \frac{k V_0}{S}. \end{aligned} \tag{11}$$

The inputs are given in Table 1. We assume a uniform distribution over the given input ranges. For the active subspace analysis, we apply linear transformations to center and scale each input to be on the $[-1, 1]$ interval.

Symbol	Range	Description
M	$[30, 60]$	piston weight (kg)
S	$[0.005, 0.020]$	piston surface area (m^2)
V	$[0.002, 0.010]$	initial gas volume (m^3)
k	$[1\,000, 5\,000]$	spring coefficient (N/m)
P_0	$[90\,000, 110\,000]$	atmospheric pressure (N/m^2)
T_a	$[290, 296]$	ambient temperature (K)
t_0	$[340, 360]$	filling gas temperature (K)

Table 1: The seven physical inputs to the piston cycle model in (10).

Figure 5 contains one- and two-dimensional shadow plots of the 500 randomly sampled data points projected into the dominant eigenspace of $\boldsymbol{C}$ for the piston model. The interactive page allows the user to vary components of $\mathbf{a}_1$ and $\mathbf{a}_2$ that define the projects in Figure 5; however, they are initially set at their optimal value for displaying ridge structure. By changing $\mathbf{a}_1$ and $\mathbf{a}_2$, the user can gain insight into the structure of high-dimensional space in relation to ridge functions.

4 Conclusion

The goal of this project was to create static, animated, and interactive visualizations that help the user gain an intuition for high-dimensional spaces and dimension reduction. Zonotopes provides a basic sense of intuition to high-dimensional spaces by projecting hypercubes down to spaces we can envision. The interactive tool enables the user to rotate an m-dimensional hypercube for m as large as 10. In the context of dimension reduction, we provided a seven-dimensional function that exhibits strong one- and two-dimensional ridge structure. Again, rotating this space enables the user to view different directions within the high-dimensional spaces.

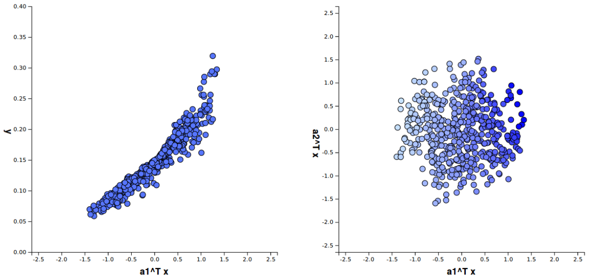


Figure 5: One- and two-dimensional shadow plots of the seven-dimensional piston model. The piston cycle time is plotted along the vertical axis on the one-dimensional shadow plot and using color in the two-dimensional shadow plot.

References

[1] K. Breyer, J. Goldstein, R. Ramakrishnan, and Shaft. U. When is ”nearest neighbor” meaningful? In *International Conference on Database Theory*, pages 217–235, 1999.

[2] P. G. Constantine. *Active Subspaces: Emerging Ideas for Dimension Reduction in Parameter Studies*. SIAM, Philadelphia, 2015.

[3] D. L. Donoho. High-dimensional data analysis: The curses and blessings of dimensionality. In *AMS Conference on Math Challenges of the 21st Century*, 2000.

[4] J. Friedman, T. Hastie, and R. Tibshirani. *The Elements of Statistical Learning*. Springer, New York, 2001.

[5] K. Fukuda. From the zonotope construction to the minkowski addition of convex polytopes. *Journal of Symbolic Computation*, 38:1261–1272, 2004.

[6] J. R. Koehler and A. B. Owen. Computer experiments. *Handbook of Statistics*, 13(9):261–308, 1996.

[7] A. Pinkus. *Ridge Functions*. Cambridge University Press, 2015.

[8] J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn. Design and analysis of computer experiments. *Statistical Science*, 4(4):409–423, 1989.

[9] S. Surjanovic and D. Bingham. Virtual library of simulation experiments: Test functions and datasets, 2015.

[10] J. F. Traub and A. G. Werschulz. *Complexity and Information*. Cambridge University Press, Cambridge, 1998.