ACTIVE SUBPSACES

Emerging ideas for parameter reduction in computational science and engineering

PAUL CONSTANTINE
Assistant Professor
Department of Computer Science
University of Colorado, Boulder

activesubspaces.org
@DrPaulynomial

U.S. DEPARTMENT OF ENERGY | Office of Science

DARPA

NSF

SLIDES AVAILABLE UPON REQUEST

DISCLAIMER: These slides are meant to complement the oral presentation. Use out of context at your own risk.

What kinds of problems are you trying to solve?

What kinds of models do you work with?

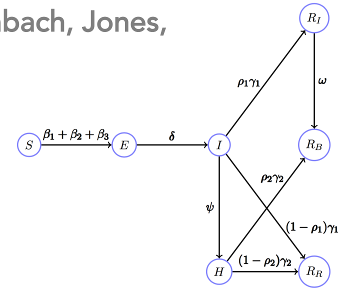
Do the models have parameters?
If so, what do the parameters do or represent?

Why do I care?

(And why I think you should, too.)

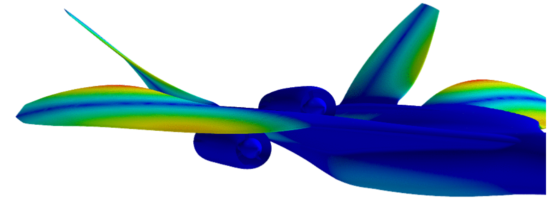
Ebola transmission models

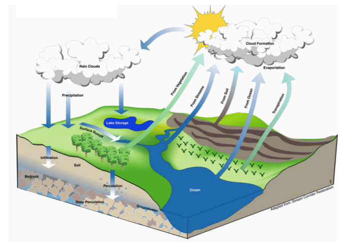
Diaz, Constantine, Kalmbach, Jones, and Pankavich (2018)



Aerospace design

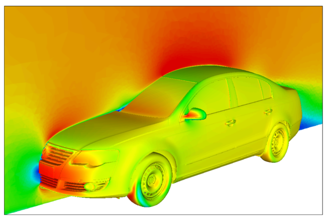
Lukaczyk, Palacios, Alonso, and Constantine (2014)





Integrated hydrologic models

Jefferson, Gilbert, Constantine, and Maxwell (2015)

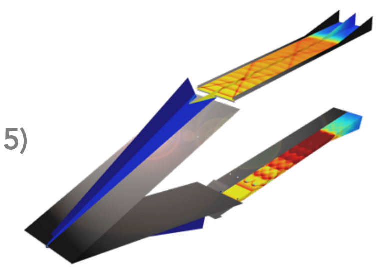


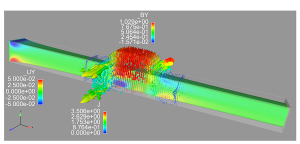
Automobile design

Othmer, Lukaczyk, Constantine, and Alonso (2016)

Hypersonic scramjet models

Constantine, Emory, Larsson, and Iaccarino (2015)





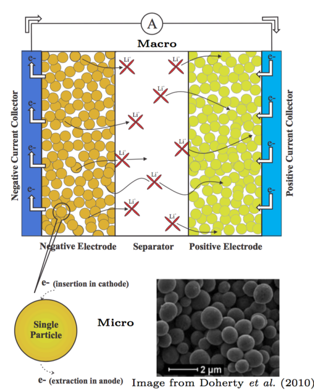
Magnetohydrodynamics models

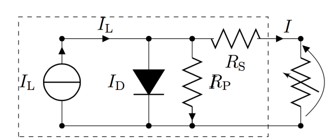
Glaws, Constantine, Shadid, and Wildey (2017)

Solar cell models

Constantine, Zaharatos, and Campanelli (2015)

$$f(\mathbf{x})$$





Lithium ion battery model

Constantine and Doostan (2017)

PROPERTIES:

Computer model of a physical system

Several independent inputs

Deterministic

Continuous inputs / outputs

"Smoothness"

$$f(\mathbf{x})$$

TO DO:

APPROXIMATION

$$\tilde{f}(\mathbf{x}) \approx f(\mathbf{x})$$

INTEGRATION

$$\int f(\mathbf{x})\,\mathrm{d}\mathbf{x}$$

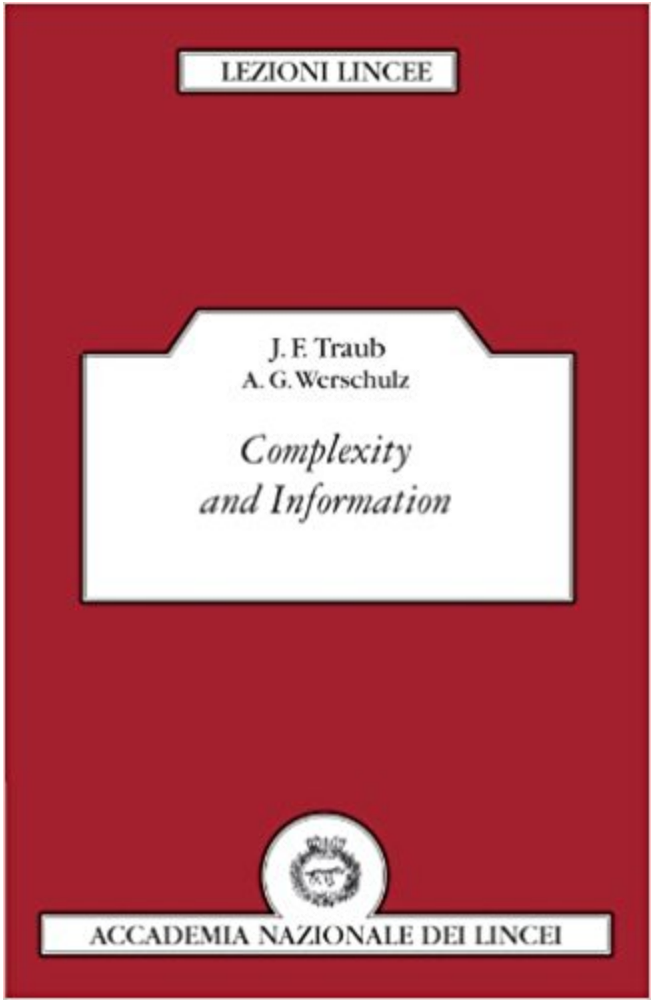
OPTIMIZATION

$$\underset{\mathbf{x}}{\text{minimize}}\; f(\mathbf{x})$$

How many dimensions is high dimensions?

Troubles in high dimensions

the information-based complexity (IBC) notion of **tractability**



REDUCED-ORDER MODELS
or
PARALLEL PROCESSING

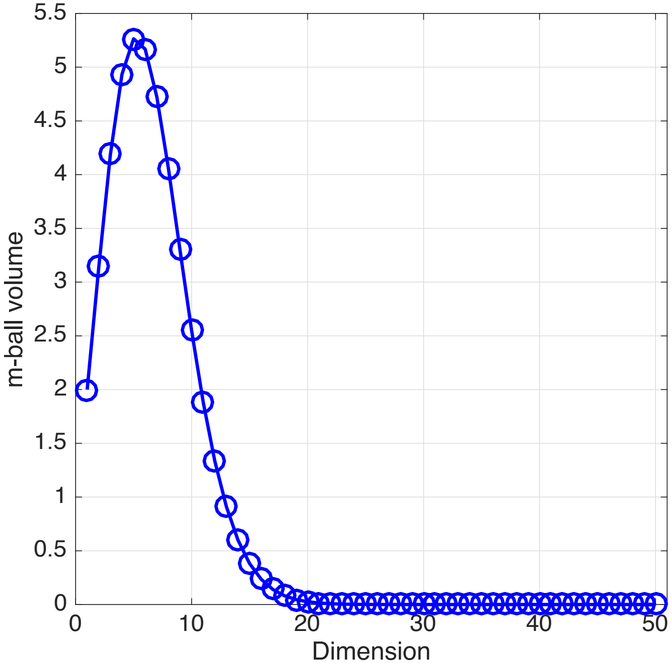
Number of parameters (the dimension)	**Number of model runs** (at 10 points per dimension)	**Time for parameter study** (at 1 second per run)
1	10	10 sec
2	100	~ 1.6 min
3	1,000	~ 16 min
4	10,000	~ 2.7 hours
5	100,000	~ 1.1 days
6	1,000,000	~ 1.6 weeks
…	…	…
20	1e20	3 trillion years (240x age of the universe)

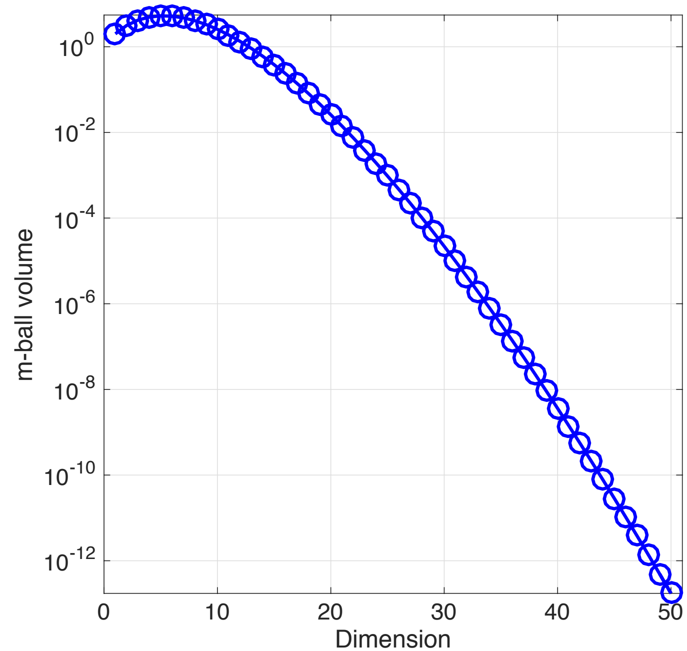
BETTER DESIGNS
or
ADAPTIVE SAMPLING

Number of parameters (the dimension)	**Number of model runs** (at 10 points per dimension)	**Time for parameter study** (at 1 second per run)
1	10	10 sec
2	100	~ 1.6 min
3	1,000	~ 16 min
4	10,000	~ 2.7 hours
5	100,000	~ 1.1 days
6	1,000,000	~ 1.6 weeks
...	...	...
20	1e20	3 trillion years (240x age of the universe)

Troubles in high dimensions

volume of a unit ball in m dimensions: $$\frac{\pi^{\frac{m}{2}}}{\Gamma(\frac{m}{2}+1)}$$





When Is "Nearest Neighbor" Meaningful?

Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft

CS Dept., University of Wisconsin-Madison
1210 W. Dayton St., Madison, WI 53706
{beyer, jgoldst, raghu, uri}@cs.wisc.edu

Database Theory --- ICDT'99, Springer (1999)

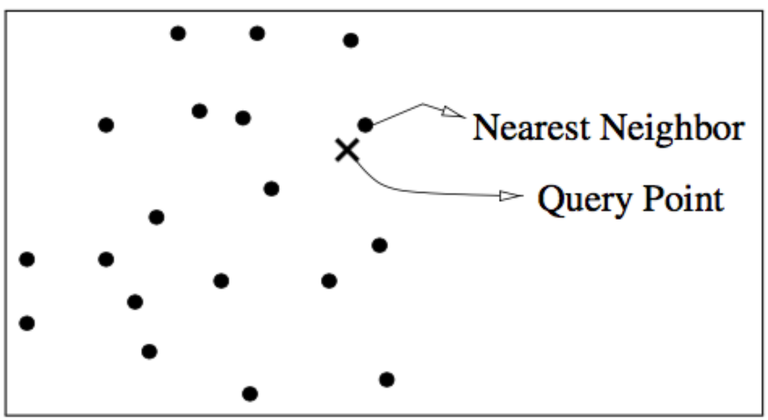


Fig. 1. Query point and its nearest neighbor.

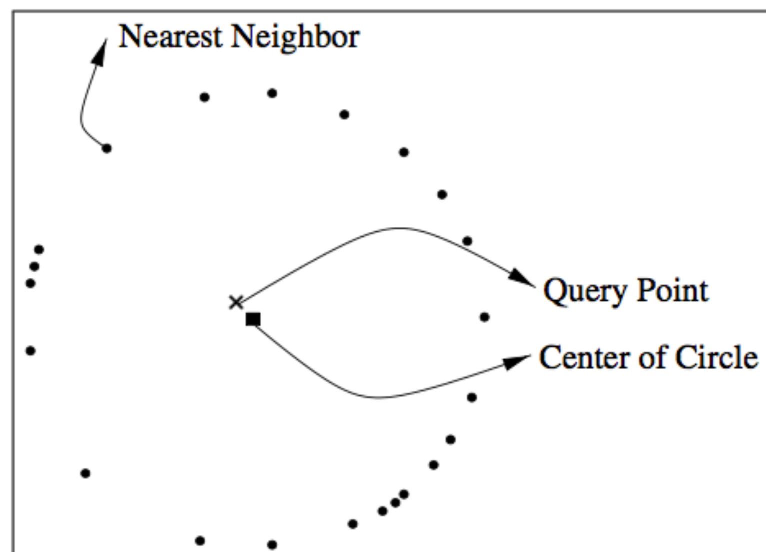
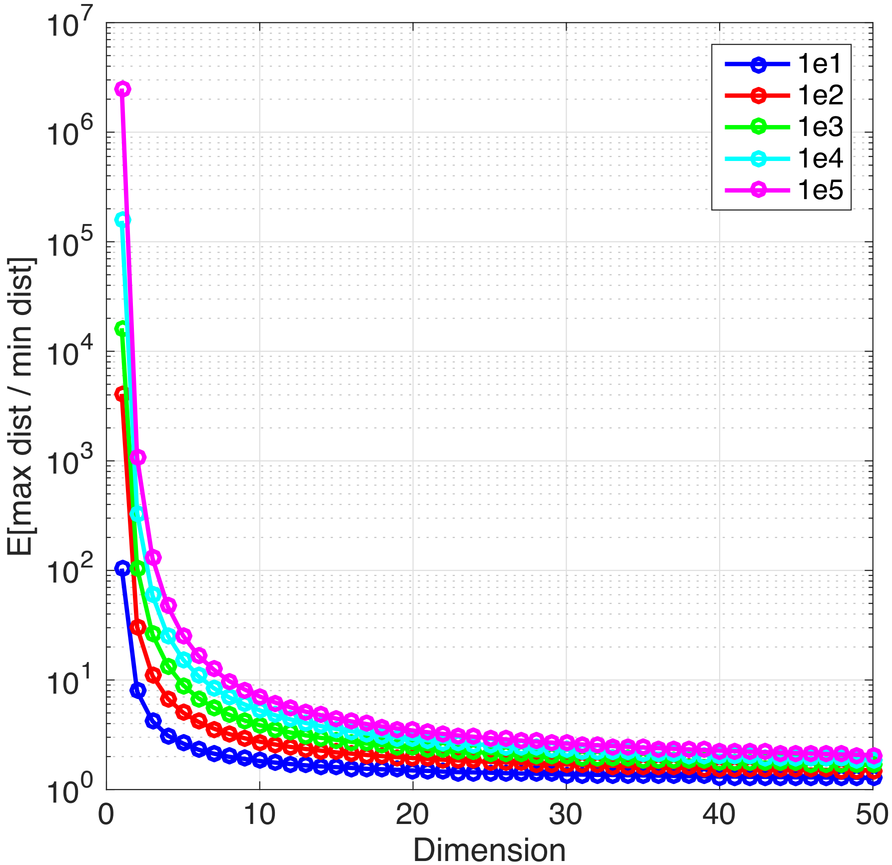
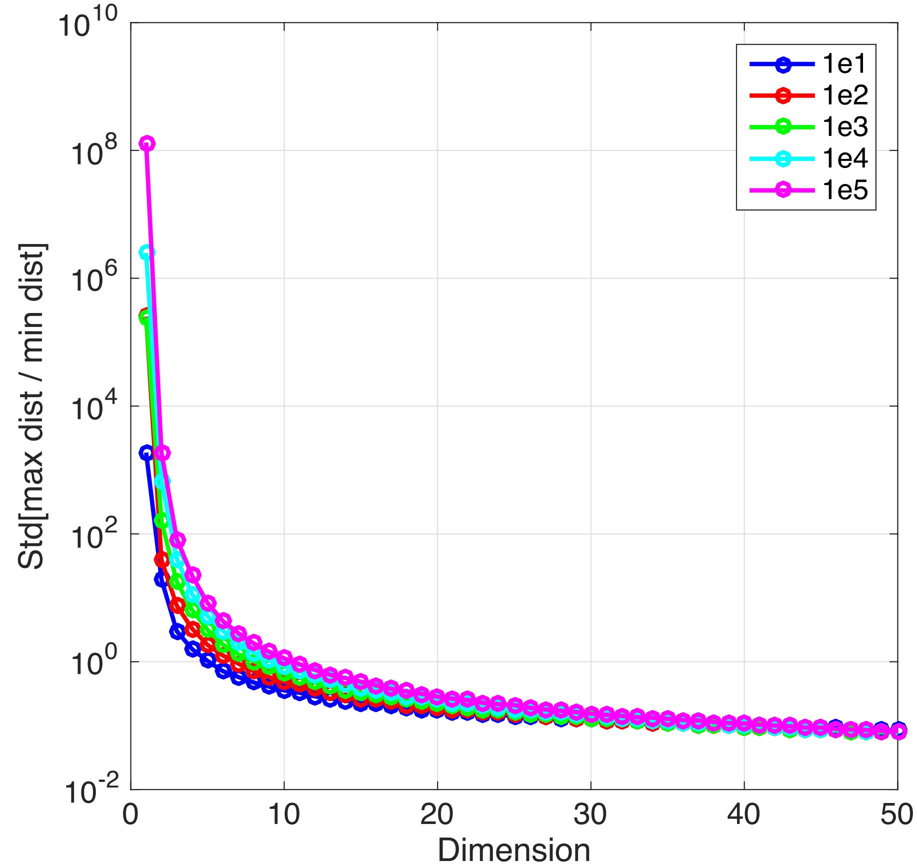


Fig. 2. Another query point and its nearest neighbor.

When Is “Nearest Neighbor” Meaningful?





Structure-exploiting methods

STRUCTURE

METHODS

$$f(\mathbf{x}) \approx f_1(x_1) + \cdots + f_m(x_m)$$

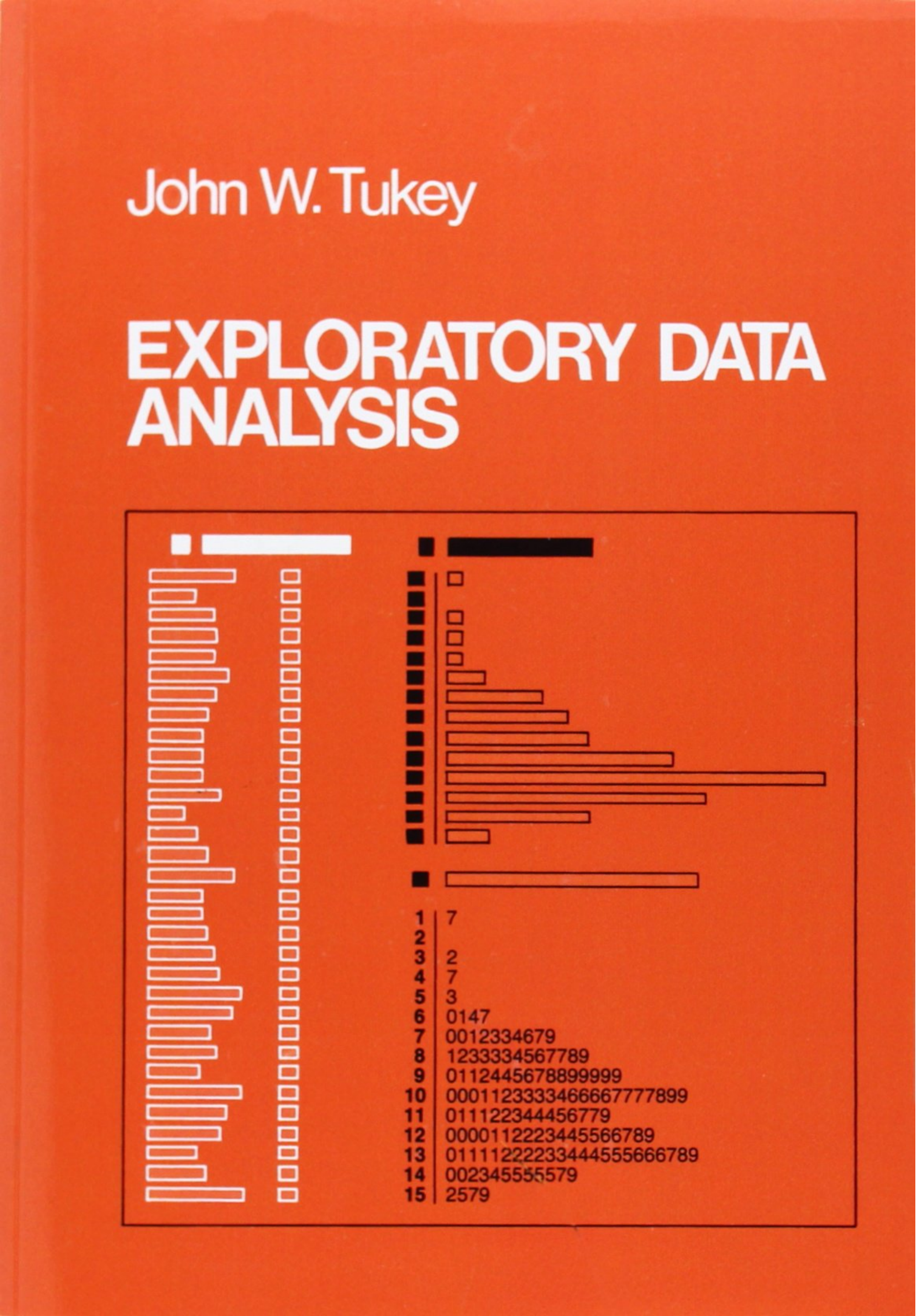
Sparse grids [Bungartz & Griebel (2004)], HDMR [Sobol (2003)], ANOVA [Hoeffding (1948)], QMC [Niederreiter (1992)], …

$$f(\mathbf{x}) \approx \sum_{k=1}^{r} f_{k,1}(x_1) \cdots f_{k,m}(x_m)$$

Separation of variables [Beylkin & Mohlenkamp (2005)], Tensor-train [Oseledets (2011)], Adaptive cross approximation [Bebendorff (2011)], Proper generalized decomposition [Chinesta et al. (2011)], …

$$f(\mathbf{x}) \approx \sum_{k=1}^{p} a_k \, \phi_k(\mathbf{x}), \quad \|\mathbf{a}\|_0 \ll p$$

Compressed sensing [Donoho (2006), Candès & Wakin (2008)], …



"Even more understanding is *lost* if we consider each thing we can do to data *only* in terms of some set of very restrictive assumptions under which that thing is best possible—assumptions we *know we CANNOT check in practice.*"

The best way to fight the curse is to reduce the dimension.

But what is *dimension reduction*?

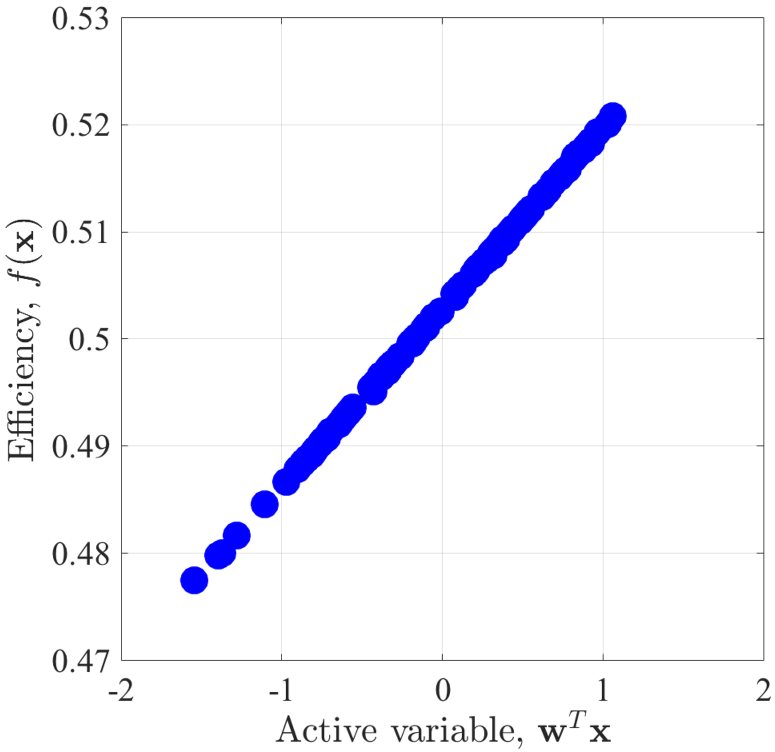
- physical reasoning
- dimensional analysis [Barrenblatt (1996)]
- correlation-based reduction [Jolliffe (2002)]
- global sensitivity analysis [Saltelli et al. (2008)]

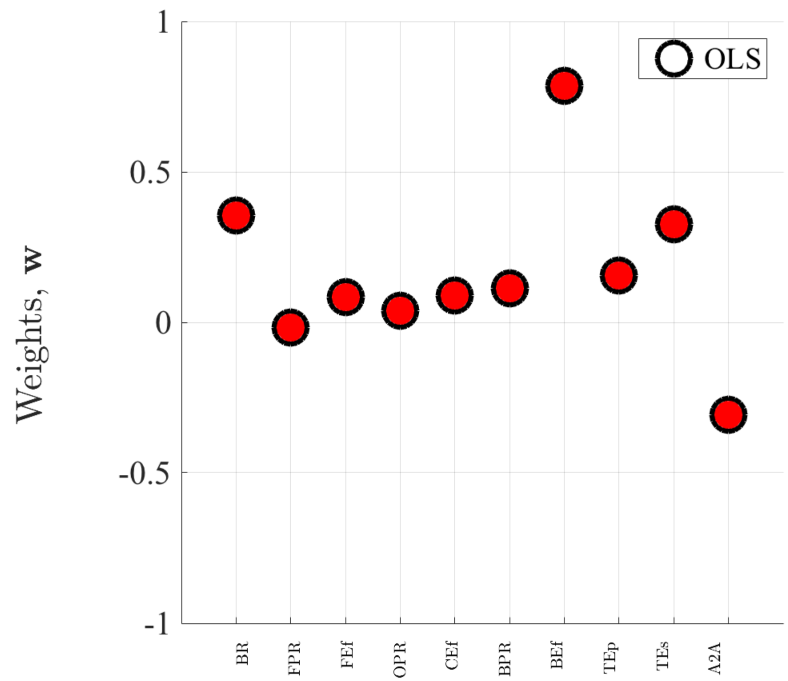
ACTIVE SUBSPACES

activesubspaces.org

www.youtube.com/watch?v=mJvKzjT6lmY

Design a jet nozzle under uncertainty
(DARPA SEQUOIA project)





10-parameter engine performance model
(See animation at https://youtu.be/Fek2HstkFVc)

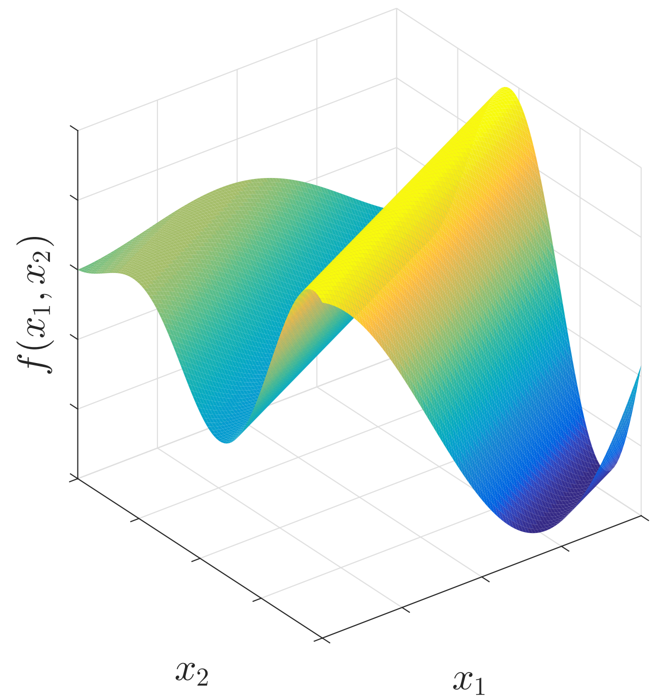
Ridge approximations

$$f(\mathbf{x}) \approx g(\boldsymbol{U}^T\mathbf{x})$$

where

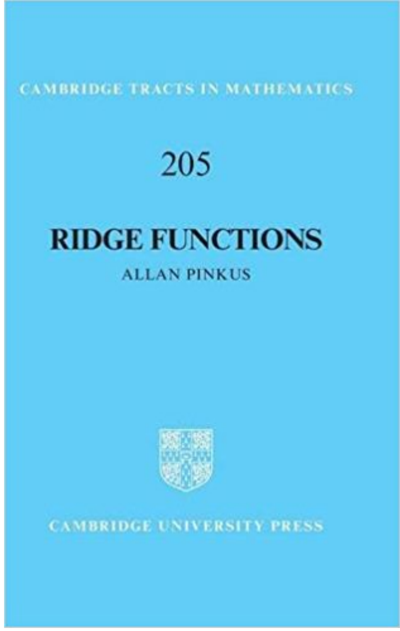
$$\boldsymbol{U}^T : \underbrace{\mathbb{R}^m}_{\text{“big”}} \rightarrow \underbrace{\mathbb{R}^n}_{\text{“small”}}$$

$$g : \mathbb{R}^n \rightarrow \mathbb{R}$$



Constantine, Eftekhari, Hokanson, and Ward (2017)

Ridge approximations



$$f(\mathbf{x}) \;\approx\; g(\boldsymbol{U}^T\mathbf{x})$$

A subset of related literature

Approximation theory: Mayer et al. (2015), Pinkus (2015), Diaconis and Shahshahani (1984), Donoho and Johnstone (1989)

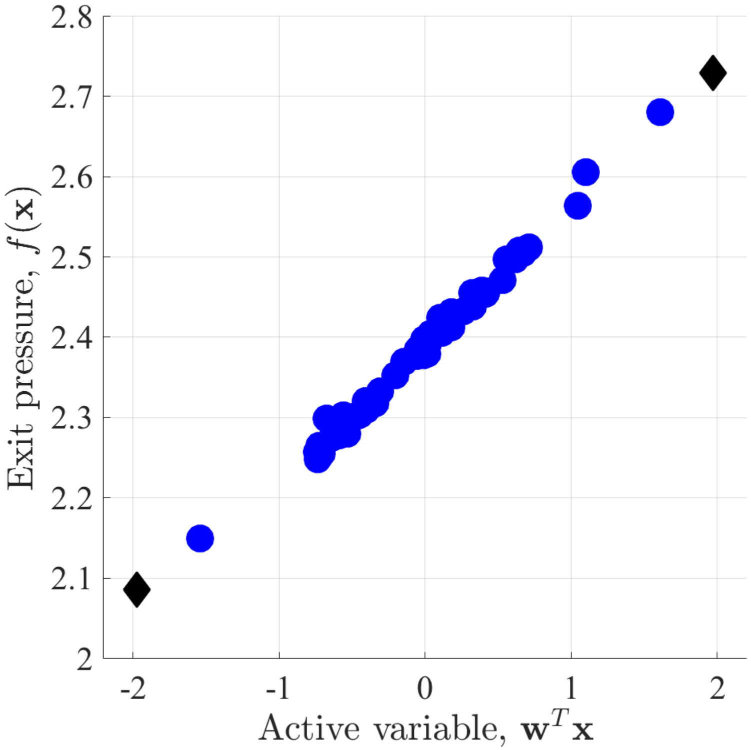
Compressed sensing: Fornasier et al. (2012), Cohen et al. (2012), Tyagi and Cevher (2014)

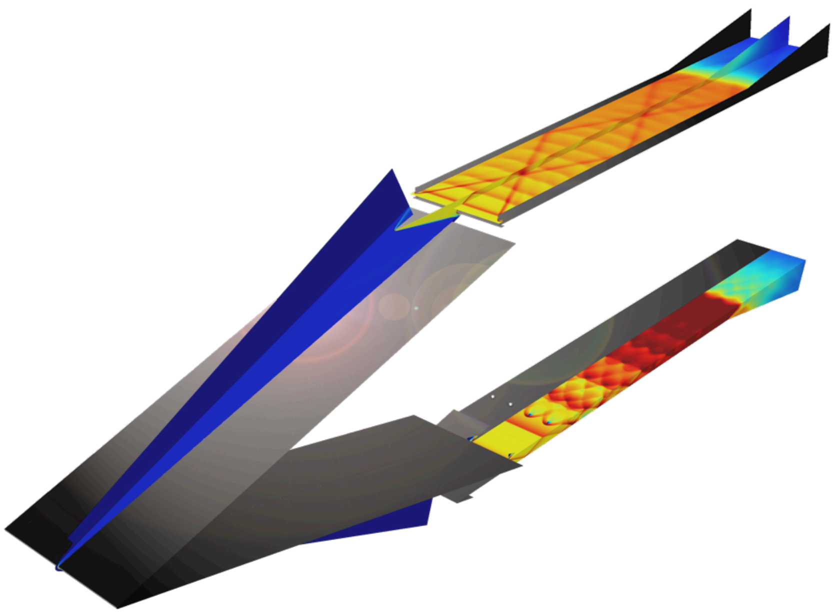
Statistical regression: Friedman and Stuetzle (1981), Ichimura (1993), Hristache et al. (2001), Xia et al. (2002)

Uncertainty quantification & computational science: Tipireddy and Ghanem (2014); Lei et al. (2015); Stoyanov and Webster (2015); Tripathy, Bilionis, and Gonzalez (2016); Li, Lin, and Li (2016); …

Do these structures arise in real models?

Evidence of 1d ridge structures across science and engineering models

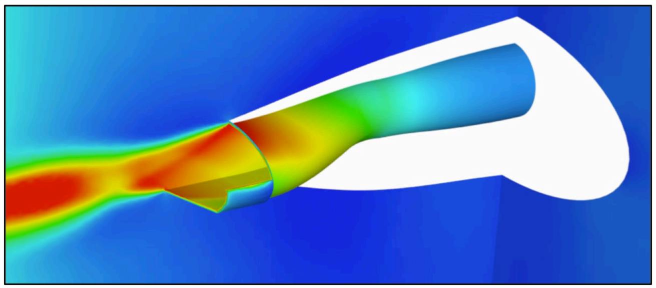




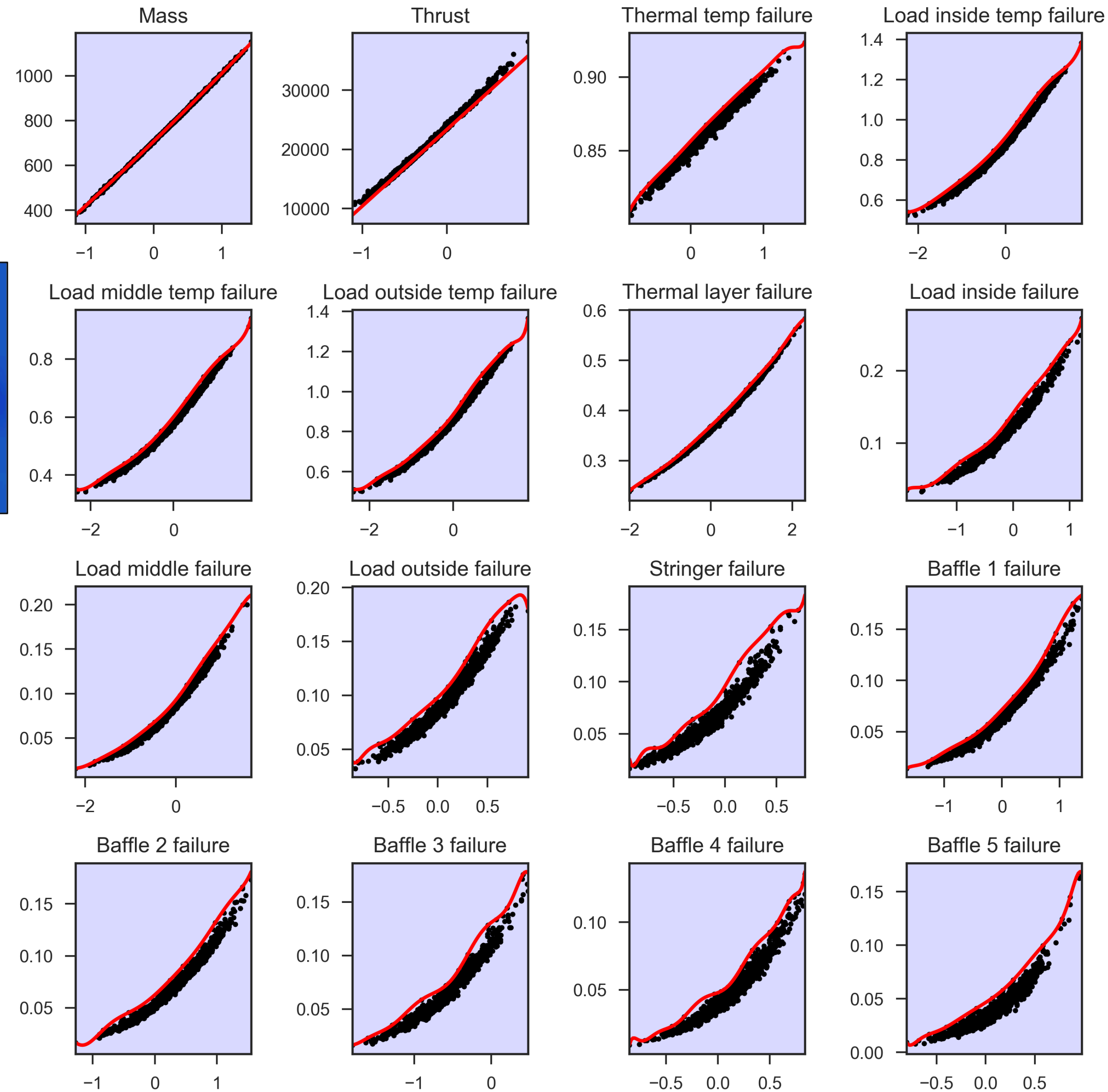
Hypersonic scramjet models

Constantine, Emory, Larsson, and Iaccarino (2015)

Evidence of 1d ridge structures across science and engineering models

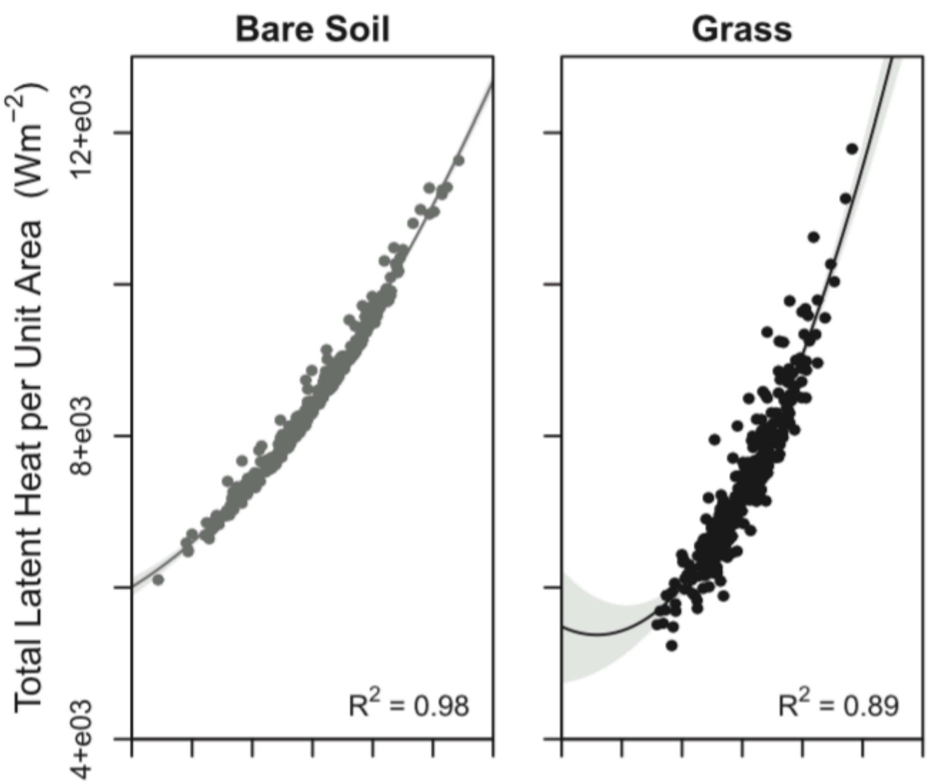


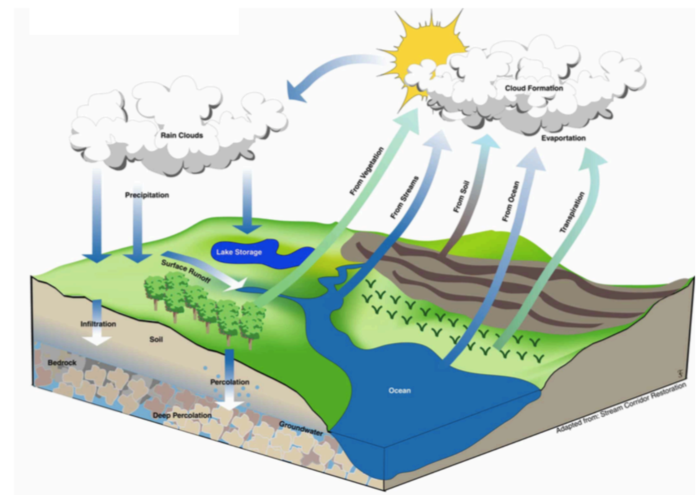
Integrated jet nozzle models



Alonso, Eldred, Constantine, Duraisamy, Farhat, Iaccarino, and Jakeman (2017)

Evidence of 1d ridge structures across science and engineering models

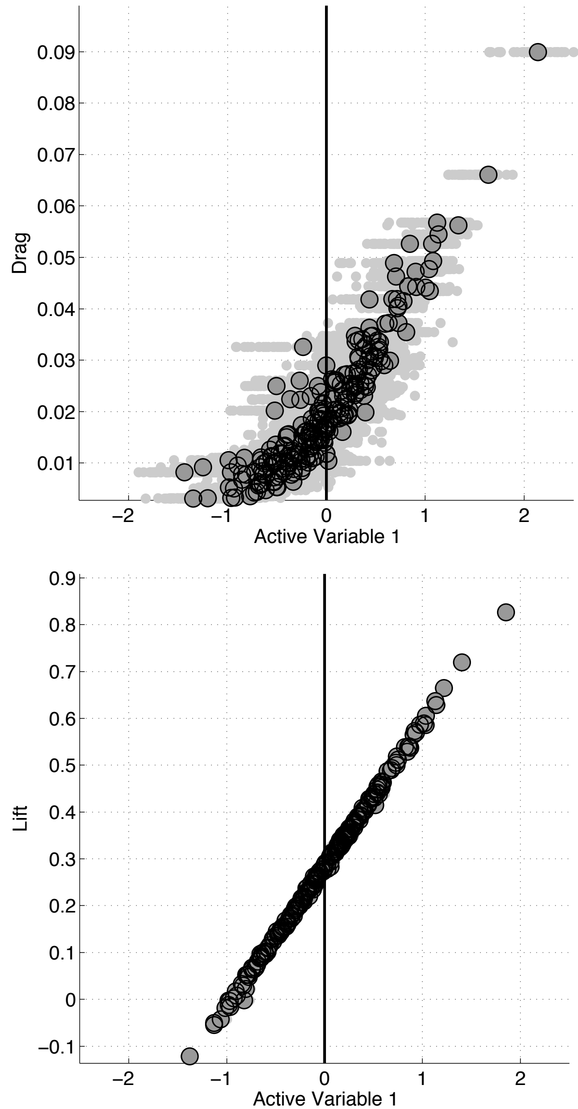


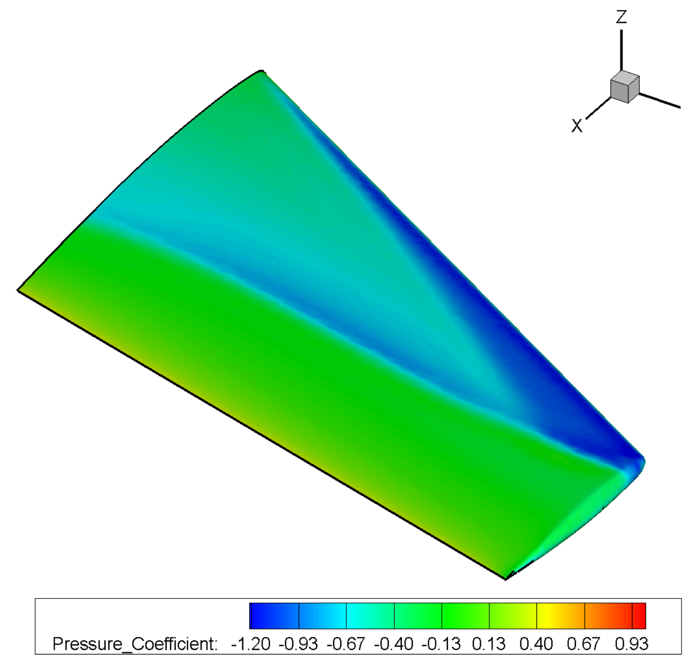


Integrated hydrologic models

Jefferson, Gilbert, Constantine, and Maxwell (2015)

Evidence of 1d ridge structures across science and engineering models

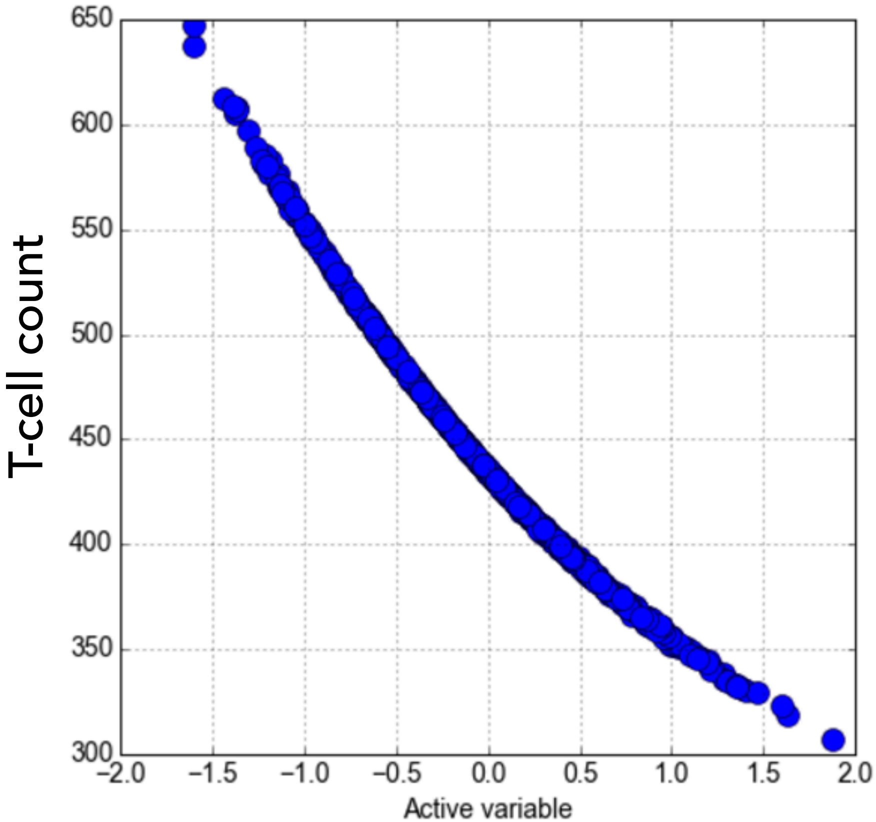


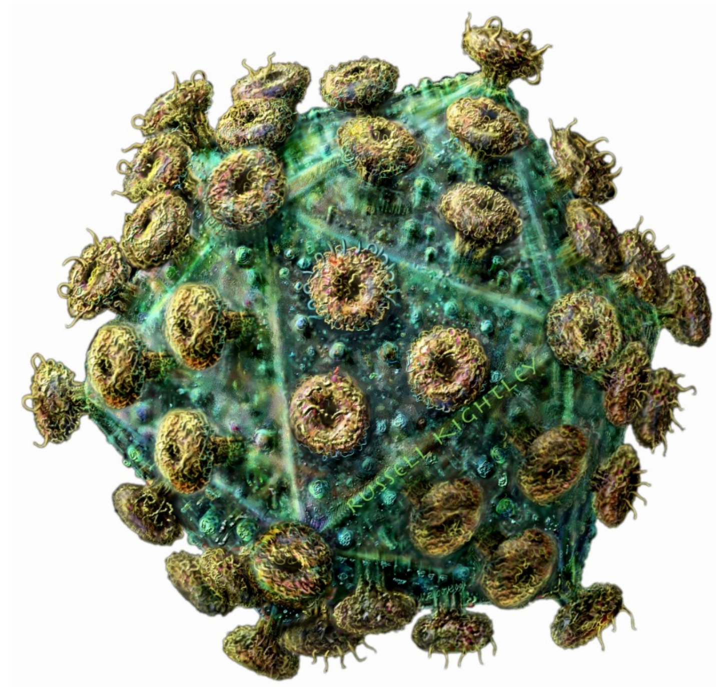


Aerospace vehicle geometries

Lukaczyk, Constantine, Palacios, and Alonso (2014)

Evidence of 1d ridge structures across science and engineering models

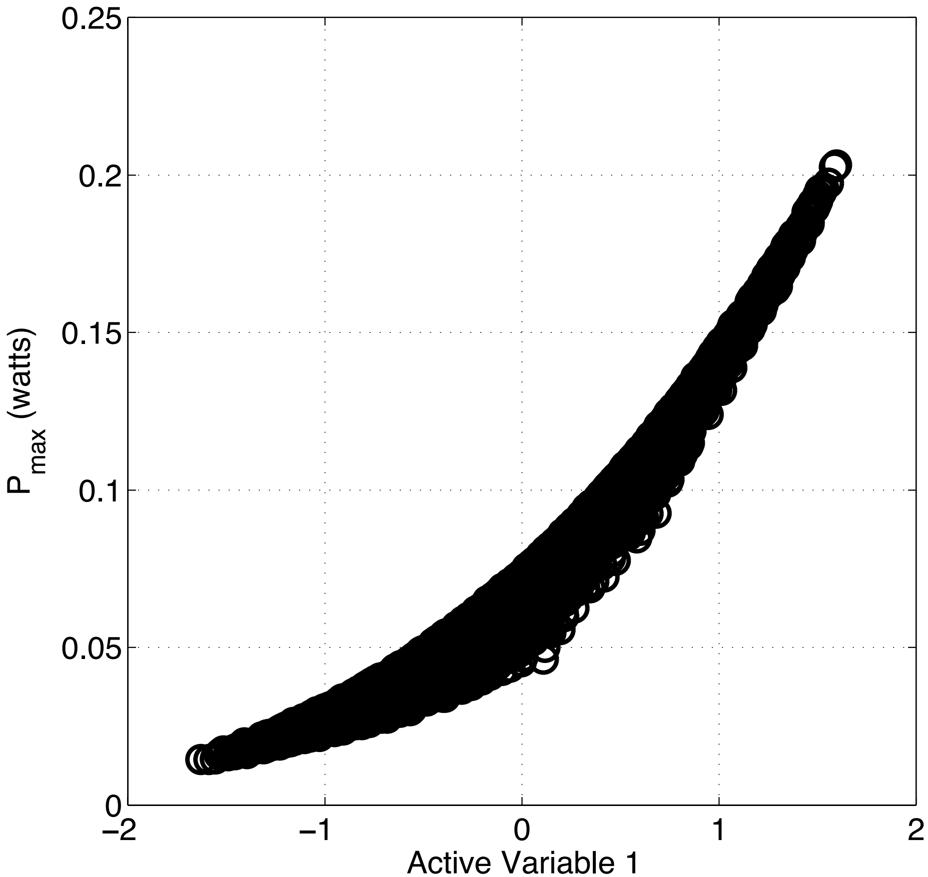


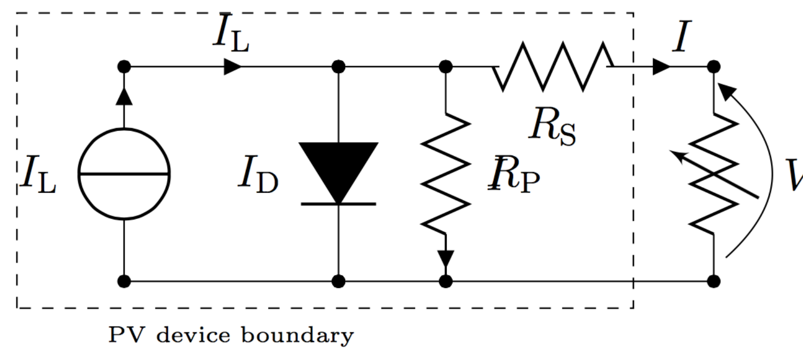


In-host HIV dynamical models

Loudon and Pankavich (2016)

Evidence of 1d ridge structures across science and engineering models

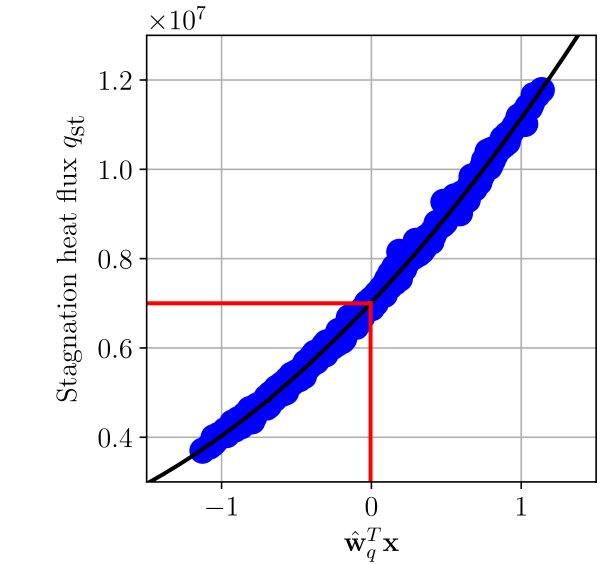


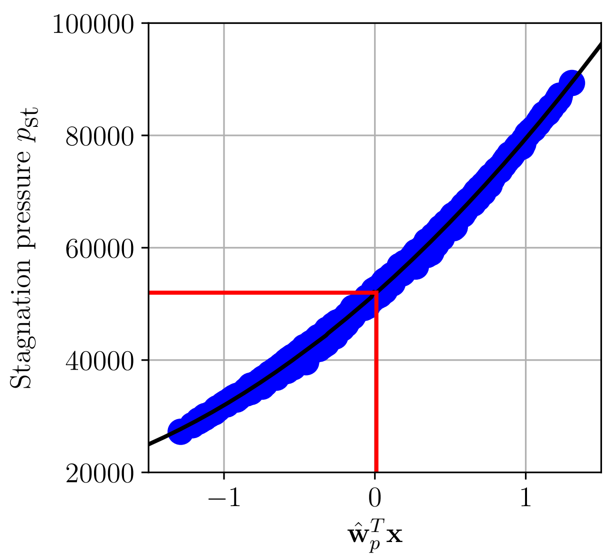


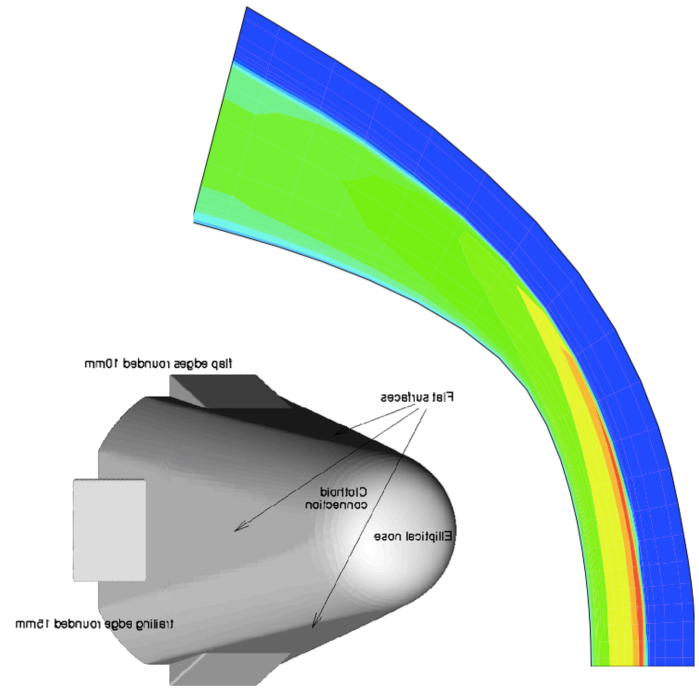
Solar cell circuit models

Constantine, Zaharatos, and Campanelli (2015)

Evidence of 1d ridge structures across science and engineering models



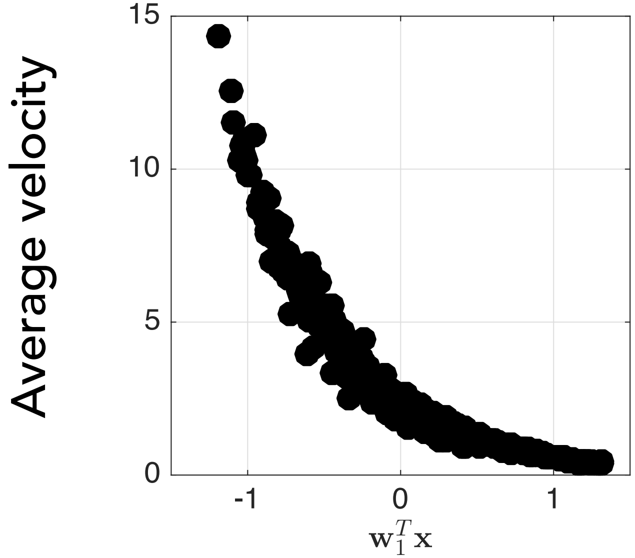


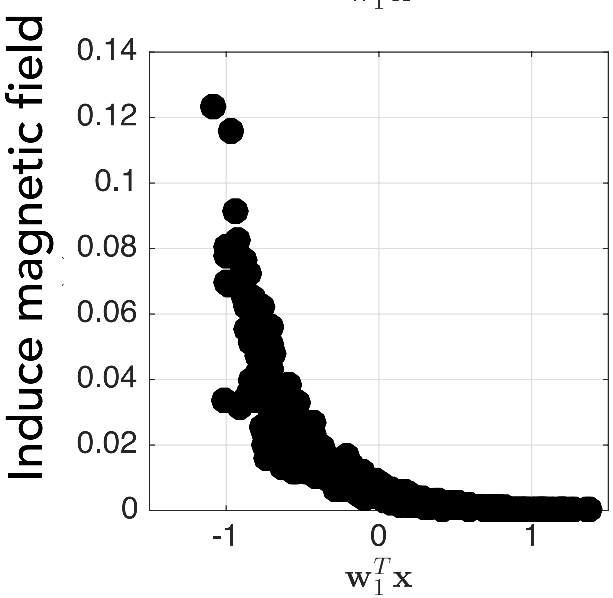


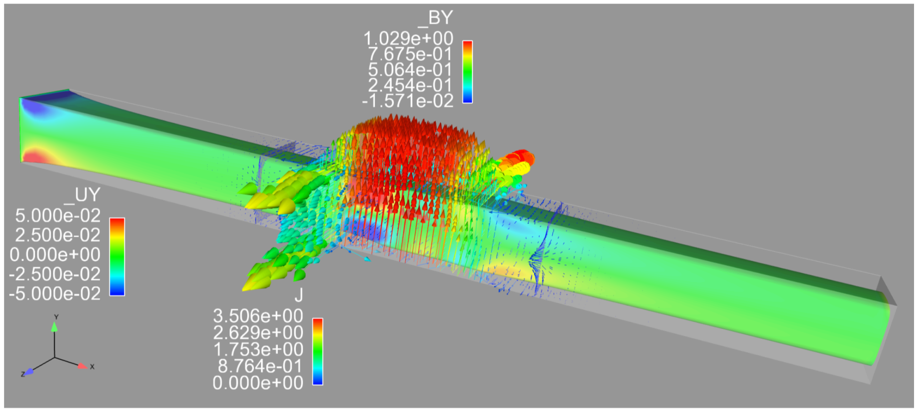
Atmospheric reentry vehicle model

Cortesi, Constantine, Magin, and Congedo (hal, 2017)

Evidence of 1d ridge structures across science and engineering models



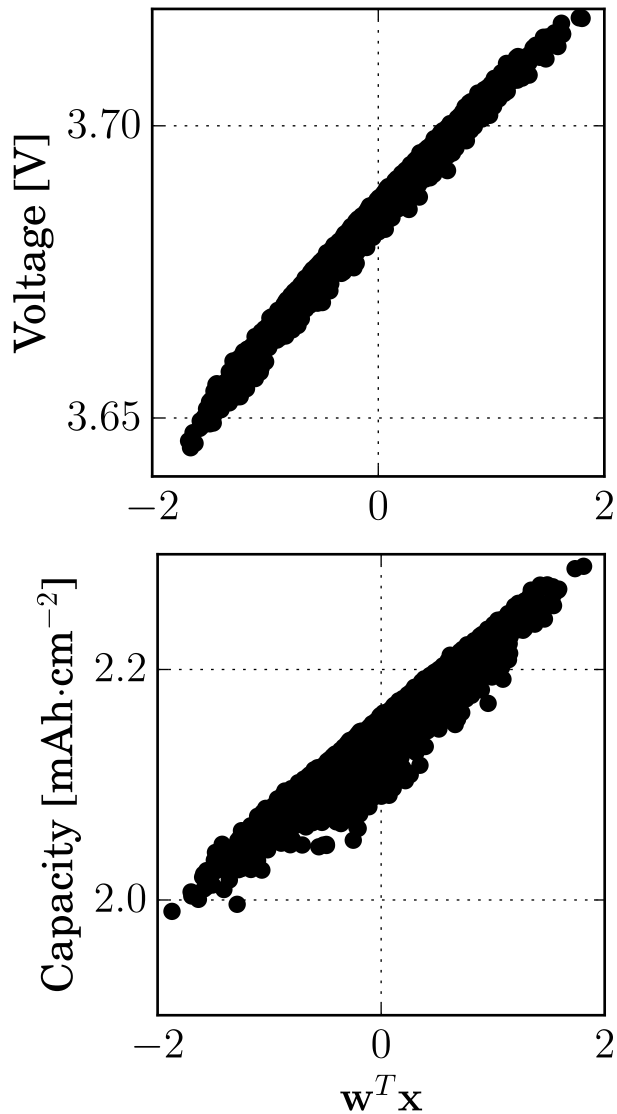




Magnetohydrodynamics generator model

Glaws, Constantine, Shadid, and Wildey (2017)

Evidence of 1d ridge structures across science and engineering models



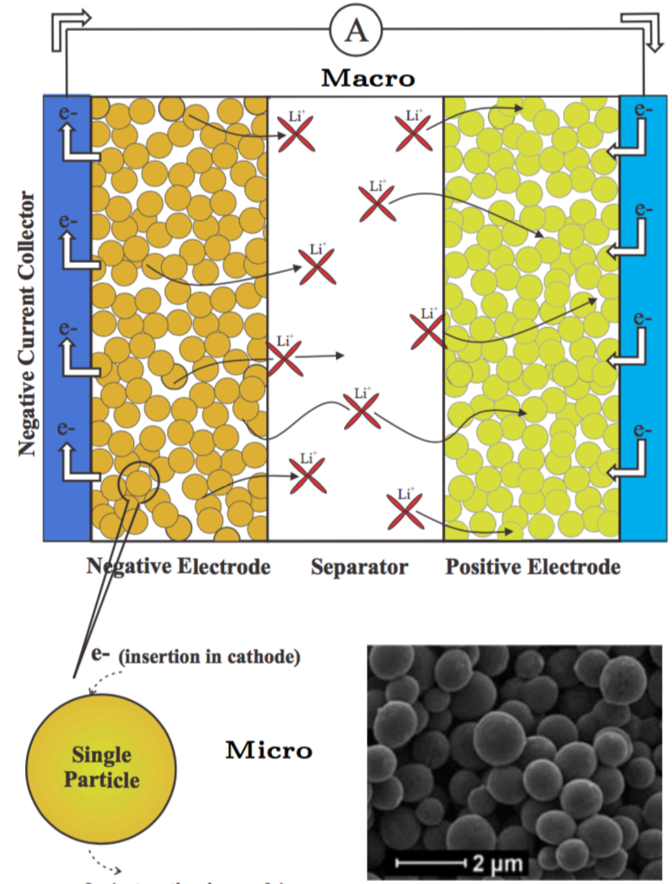
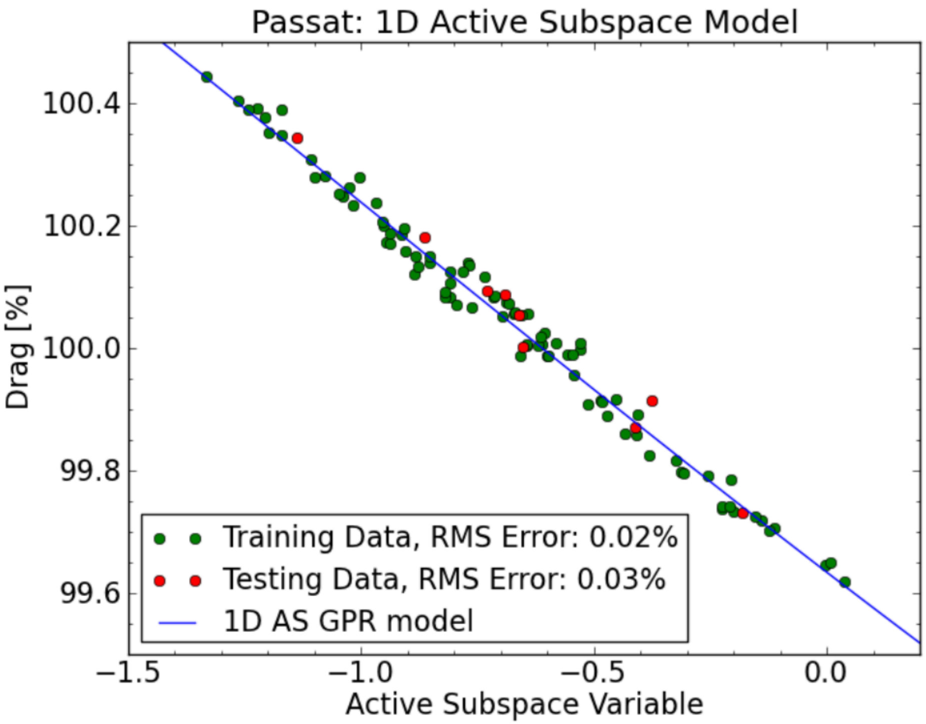


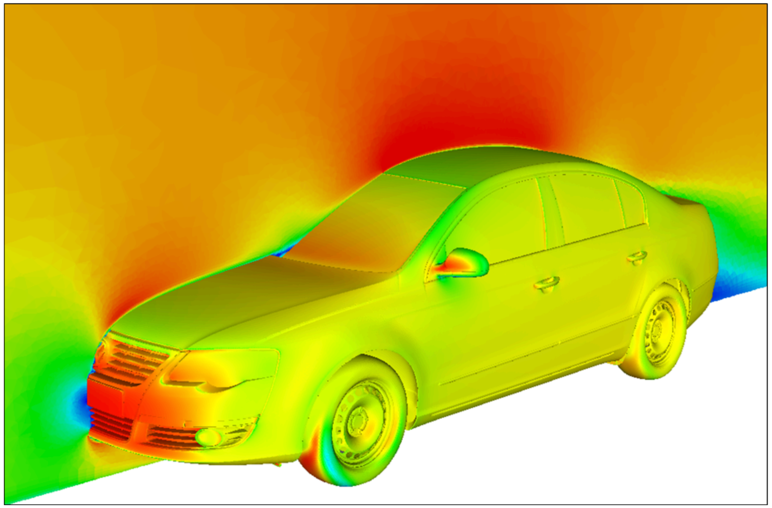
Image from Doherty *et al.* (2010)

Lithium ion battery model

Constantine and Doostan (2017)

Evidence of 1d ridge structures across science and engineering models



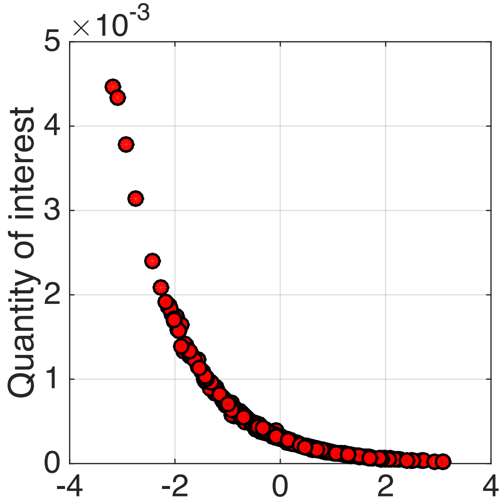


Automobile geometries

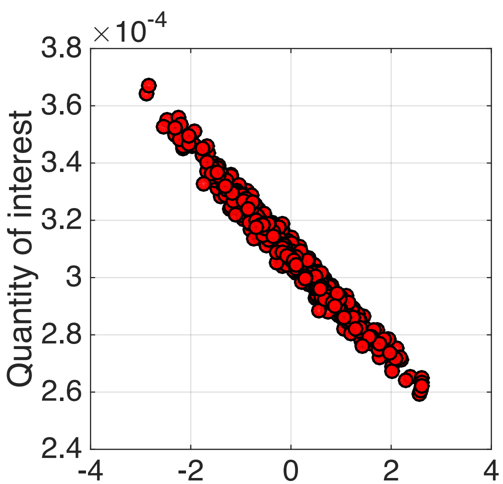
Othmer, Lukaczyk, Constantine, and Alonso (2016)

Evidence of 1d ridge structures across science and engineering models

Long length scale

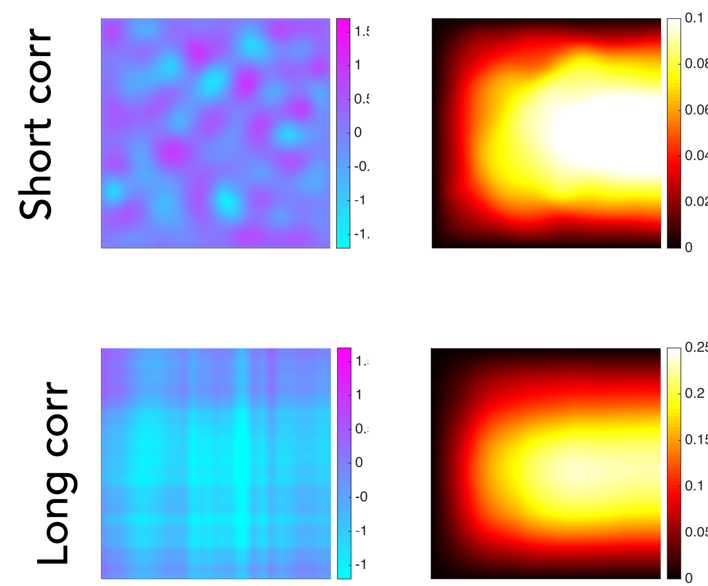


Short length scale



$$-\nabla \cdot (a \nabla u) = 1, \quad \mathbf{s} \in \mathcal{D}$$
$$u = 0, \quad \mathbf{s} \in \Gamma_1$$
$$-\mathbf{n} \cdot a \nabla u = 0, \quad \mathbf{s} \in \Gamma_2$$

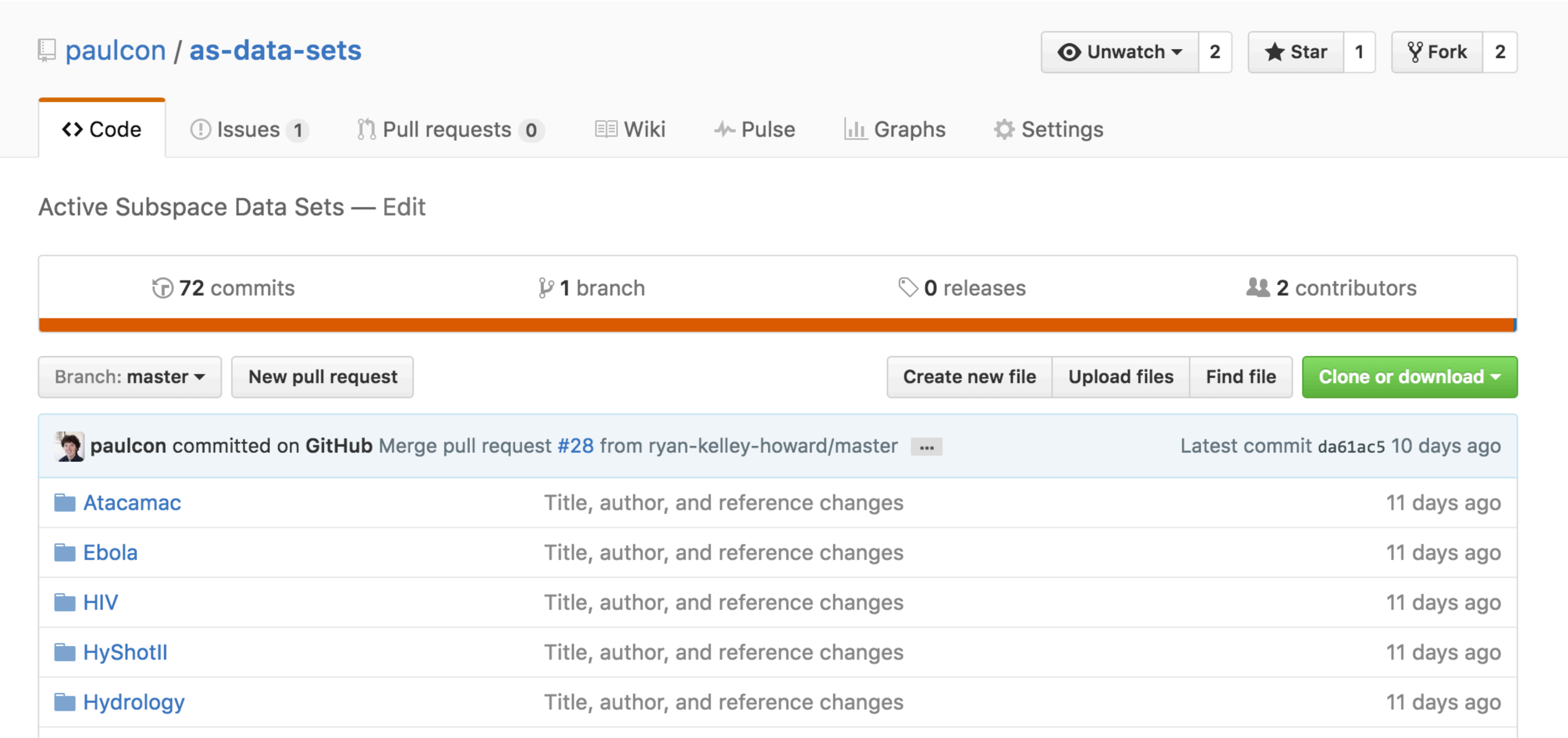
Input field Solution



Constantine, Dow, and Wang (2014)

$$f(\mathbf{x})$$

Jupyter notebooks:
github.com/paulcon/as-data-sets



What about the math?

Ridge approximations

What is the
approximation error?

What is U?

$$f(\mathbf{x}) \boxed{\approx} \boxed{g}(\boxed{\boldsymbol{U}}^T\mathbf{x})$$

What is g?

Constantine, Eftekhari, Hokanson, and Ward (2017)

Ridge approximations

$$f(\mathbf{x}) \approx \boxed{g}(\boldsymbol{U}^T\mathbf{x})$$

What is g?

Use the conditional average:

conditional density

$$\mu(\mathbf{y}) = \int f(\boldsymbol{U}\boxed{\mathbf{y}} + \boxed{\boldsymbol{V}\mathbf{z}})\,\boxed{\pi(\mathbf{z}|\mathbf{y})}\,\mathrm{d}\mathbf{z}$$

subspace coordinates

complement
subspace and
coordinates

$\mu(\boldsymbol{U}^T\mathbf{x})$ is the **best** L_2 **approximation** [Pinkus (2015)]

Constantine, Dow, and Wang (2014); Constantine, Eftekhari, Hokanson, and Ward (2017)

Ridge approximations

What is U?

$$f(\mathbf{x}) \approx g(\boxed{\boldsymbol{U}}^T\mathbf{x})$$

Constantine, Eftekhari, Hokanson, and Ward (2017)

Define the active subspace

The function, its gradient vector, and a given weight function:

$$f = f(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^m, \quad \nabla f(\mathbf{x}) \in \mathbb{R}^m, \quad \rho : \mathbb{R}^m \to \mathbb{R}_+$$

The average outer product of the gradient and its eigendecomposition,

$$\boldsymbol{C} = \int \nabla f(\mathbf{x})\, \nabla f(\mathbf{x})^T \rho(\mathbf{x})\, \mathrm{d}\mathbf{x} = \boldsymbol{W} \Lambda \boldsymbol{W}^T$$

Some relevant literature

Statistical regression:	Samarov (1993), Hristache et al. (2001)
Machine learning:	Mukerjee, Wu, and Xiao (2010); Fukumizu and Leng (2014)
Detection and estimation theory:	van Trees (2001)

Define the active subspace

The function, its gradient vector, and a given weight function:

$$f = f(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^m, \quad \nabla f(\mathbf{x}) \in \mathbb{R}^m, \quad \rho : \mathbb{R}^m \to \mathbb{R}_+$$

The average outer product of the gradient and its eigendecomposition:

$$\boldsymbol{C} = \int \nabla f(\mathbf{x})\, \nabla f(\mathbf{x})^T \rho(\mathbf{x})\, \mathrm{d}\mathbf{x} = \boldsymbol{W} \Lambda \boldsymbol{W}^T$$

Eigenvalues measure ridge structure with eigenvectors:

$$\lambda_i = \underbrace{\int (\mathbf{w}_i^T \nabla f(\mathbf{x}))^2 \rho(\mathbf{x})\, \mathrm{d}\mathbf{x},}_{\text{average, squared, directional derivative along eigenvector}} \qquad i = 1, \dots, m$$

eigenvalue

Constantine, Dow, and Wang (2014)

Eigenvalues control the approximation error

conditional
average

Poincaré
constant

$$\left\| f(\mathbf{x}) - \mu(\boldsymbol{W}_1^T\mathbf{x}) \right\|_{L^2(\rho)} \leq C \left(\underbrace{\lambda_{n+1} + \cdots + \lambda_m}_{\text{eigenvalues associated with inactive subspace}} \right)^{\frac{1}{2}}$$

active
subspace

Constantine, Dow, and Wang (2014)

Estimate the active subspace with Monte Carlo

(1) Draw samples: $\mathbf{x}_j \sim \rho(\mathbf{x})$

(2) Compute: $f_j = f(\mathbf{x}_j)$ and $\nabla f_j = \nabla f(\mathbf{x}_j)$

(3) Approximate with Monte Carlo, and compute eigendecomposition

$$\boldsymbol{C} \approx \frac{1}{N}\sum_{j=1}^{N} \nabla f_j \, \nabla f_j^T = \hat{\boldsymbol{W}} \hat{\Lambda} \hat{\boldsymbol{W}}^T$$

Equivalent to SVD of samples of the gradient

$$\frac{1}{\sqrt{N}} \begin{bmatrix} \nabla f_1 & \cdots & \nabla f_N \end{bmatrix} = \hat{\boldsymbol{W}} \sqrt{\hat{\Lambda}} \hat{\boldsymbol{V}}^T$$

Called an *active subspace method* in T. Russi's 2010 Ph.D. thesis,
Uncertainty Quantification with Experimental Data in Complex System Models

Constantine, Dow, and Wang (2014), Constantine and Gleich (2015, arXiv)

Remember the problem to solve

Low-rank approximation of the collection of gradients:

$$\frac{1}{\sqrt{N}} \begin{bmatrix} \nabla f_1 & \cdots & \nabla f_N \end{bmatrix} \approx \hat{\boldsymbol{W}}_1 \sqrt{\hat{\Lambda}_1} \hat{\boldsymbol{V}}_1^T$$

Low-dimensional linear approximation of the gradient:

$$\operatorname{span}\left(\hat{\boldsymbol{W}}_1\right) \approx \left\{ \nabla f(\mathbf{x}) \; : \; \mathbf{x} \in \operatorname{supp} \rho(\mathbf{x}) \right\}$$

Approximate a function of many variables by a function of a few linear combinations of the variables:

$$f(\mathbf{x}) \approx g\left(\hat{\boldsymbol{W}}_1^T \mathbf{x}\right)$$

How many gradient samples?

bound on gradient

number of samples

$$N = \Omega\left(\frac{L^2 \lambda_1}{\lambda_k^2 \varepsilon^2}\,\log(m)\right) \implies \overbrace{|\lambda_k - \hat{\lambda}_k| \leq \lambda_k\,\varepsilon}^{\text{eigenvalue error (w.h.p.)}}$$

dimension

bound on gradient

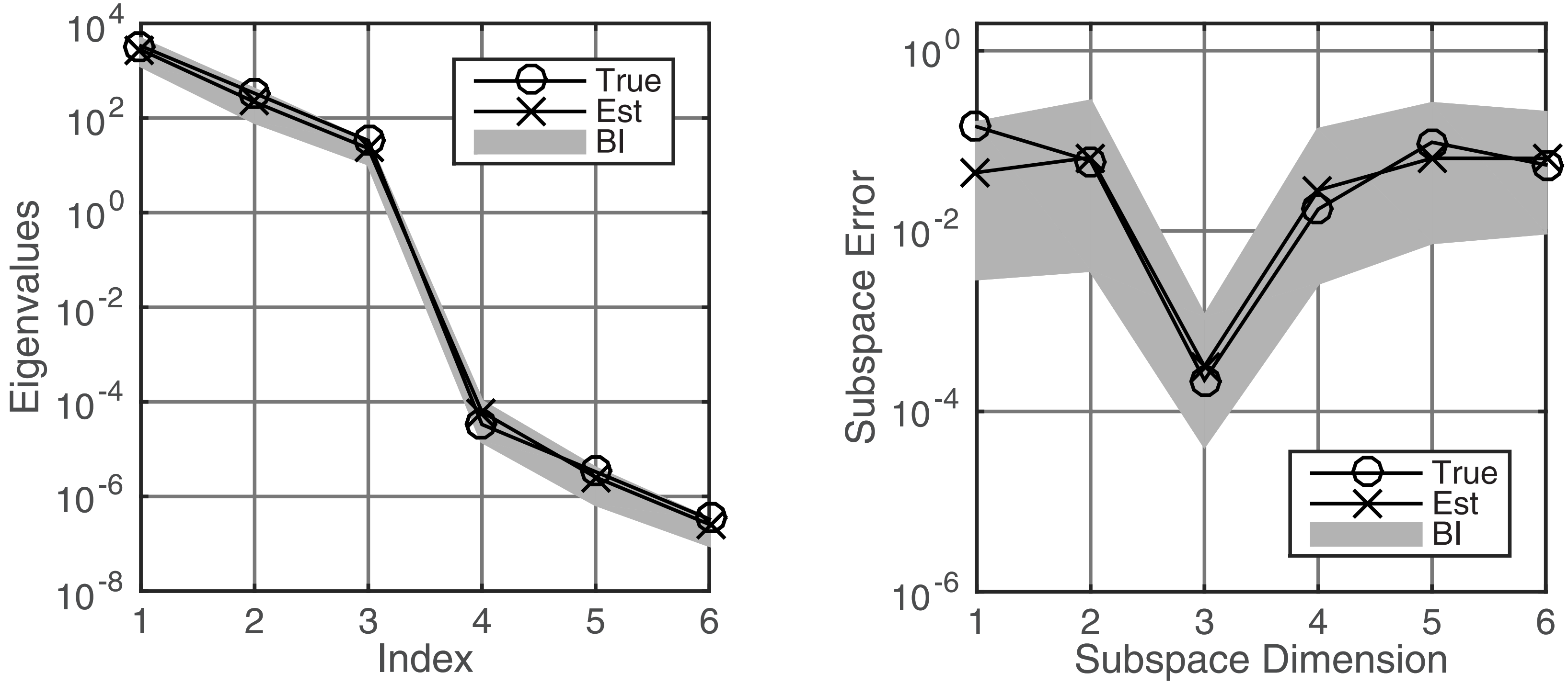
number of samples

$$N = \Omega\left(\frac{L^2}{\lambda_1 \varepsilon^2}\,\log(m)\right) \implies \underbrace{\operatorname{dist}(\boldsymbol{W}_1, \hat{\boldsymbol{W}}_1) \leq \frac{4\lambda_1 \varepsilon}{\lambda_n - \lambda_{n+1}}}_{\text{subspace error (w.h.p.)}}$$

dimension

Constantine and Gleich (2015) via Gittens and Tropp (2011), Stewart (1973)

In practice, bootstrap



Eigenvalue estimates and subspace error estimates with bootstrap intervals from quadratic function of 10 variables

Constantine and Gleich (2015, arXiv)

Effect of estimated eigenvectors?

Recall the subspace error:

$$\varepsilon \;=\; \mathrm{dist}(\boldsymbol{W}_1, \hat{\boldsymbol{W}}_1) \;\propto\; \frac{\mathcal{O}(\text{eigval. error})}{\lambda_n - \lambda_{n+1}}$$

Eigenvalues for inactive variables

$$\left\| f(\mathbf{x}) - \mu(\hat{\boldsymbol{W}}_1^T \mathbf{x}) \right\|_{L^2(\rho)} \leq C \left(\varepsilon \, (\lambda_1 + \cdots + \lambda_n)^{\frac{1}{2}} + (\lambda_{n+1} + \cdots + \lambda_m)^{\frac{1}{2}} \right)$$

Subspace error

Eigenvalues for active variables

Constantine, Dow, and Wang (2014)

Is the active subspace optimal?

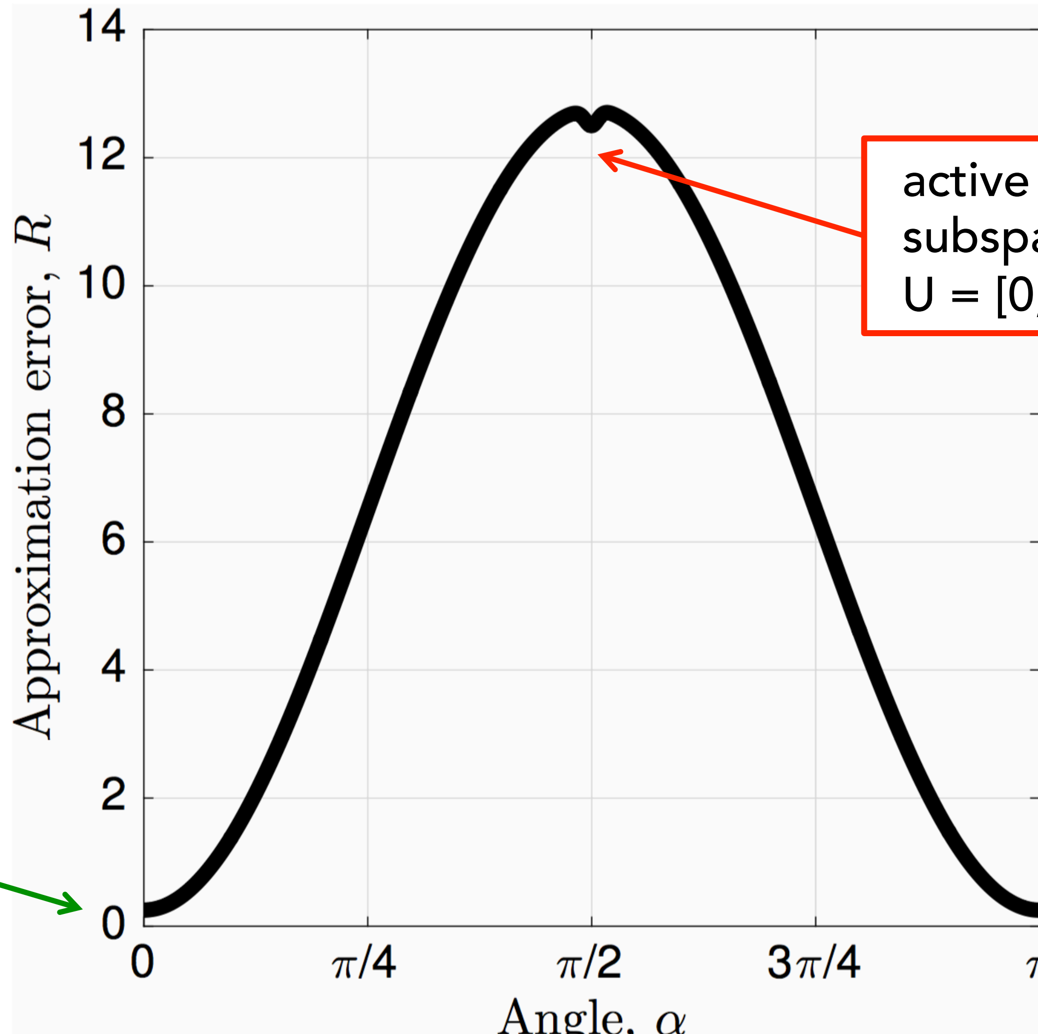
(No.)

An example where it doesn't work

$$f(x_1, x_2) = 5x_1 + \sin(10\pi x_2)$$

$$\boldsymbol{C} = \text{``}\int \nabla f \, \nabla f^T \text{''}$$

$$= \begin{bmatrix} 25 & 0 \\ 0 & 526 \end{bmatrix}$$



active subspace
U = [0; 1]

inactive subspace
U = [1; 0]

Constantine, Eftekhari, Hokanson, and Ward (2017)

Ridge approximations

What is U?

$$f(\mathbf{x}) \approx g(\boxed{\boldsymbol{U}}^T\mathbf{x})$$

Define the error function:

best approximation

$$R(\boldsymbol{U}) = \frac{1}{2}\int (f(\mathbf{x}) - \boxed{\mu(\boldsymbol{U}^T\mathbf{x})})^2\,\rho(\mathbf{x})\,d\mathbf{x}$$

Minimize the error:

$$\underset{\boldsymbol{U}}{\text{minimize}} \;\; R(\boldsymbol{U}) \quad \text{subject to} \;\; \boldsymbol{U} \in \boxed{\mathbb{G}(n,m)}$$

Grassmann manifold

Constantine, Eftekhari, Hokanson, and Ward (2017)

The active subspace is nearly stationary

Assume (1) Lipschitz continuous function
(2) Gaussian density function

gradient on the Grassmann manifold

Lipschitz constant

$$\|\bar{\nabla} R(\boldsymbol{W}_1)\|_F \;\leq\; L \overbrace{\left(2m^{\frac{1}{2}} + (m-n)^{\frac{1}{2}}\right)}^{\text{dimensions}} \underbrace{(\lambda_{n+1} + \cdots + \lambda_m)^{\frac{1}{2}}}_{\text{eigenvalues associated with inactive subspace}}$$

active subspace

Frobenius norm

Constantine, Eftekhari, Hokanson, and Ward (2017)

Estimate the optimal subspace with discrete least squares

(1) Choose points: $\mathbf{x}_j \sim \rho(\mathbf{x})$

(2) Compute: $f_j = f(\mathbf{x}_j)$

(3) Minimize the misfit

polynomial

polynomials

$$\underset{\substack{g \in \mathbb{P}_p(\mathbb{R}^n) \\ \boldsymbol{U} \in \mathbb{G}(n,m)}}{\text{minimize}} \quad \sum_{j=1}^{N} \left(f_j - g(\boldsymbol{U}^T \mathbf{x}_j) \right)^2$$

subspaces

subspace

Constantine, Eftekhari, Hokanson, and Ward (2017), Hokanson and Constantine (2018)

Two contenders for the least squares problem

$$\underset{\substack{g \in \mathbb{P}_p(\mathbb{R}^n) \\ \boldsymbol{U} \in \mathbb{G}(n,m)}}{\text{minimize}} \quad \sum_{j=1}^{N} \left(f_j - g(\boldsymbol{U}^T \mathbf{x}_j)\right)^2$$

Variable projection

Use pseudoinverse of Vandermonde matrix to express optimal polynomial coefficients

Compute the derivative of the pseudoinverse of the Vandermonde matrix [Golub & Pereyra (1973)] on the Grassmann manifold [Edelman et al. (1998)]

Run Newton on loss function

Alternating minimization

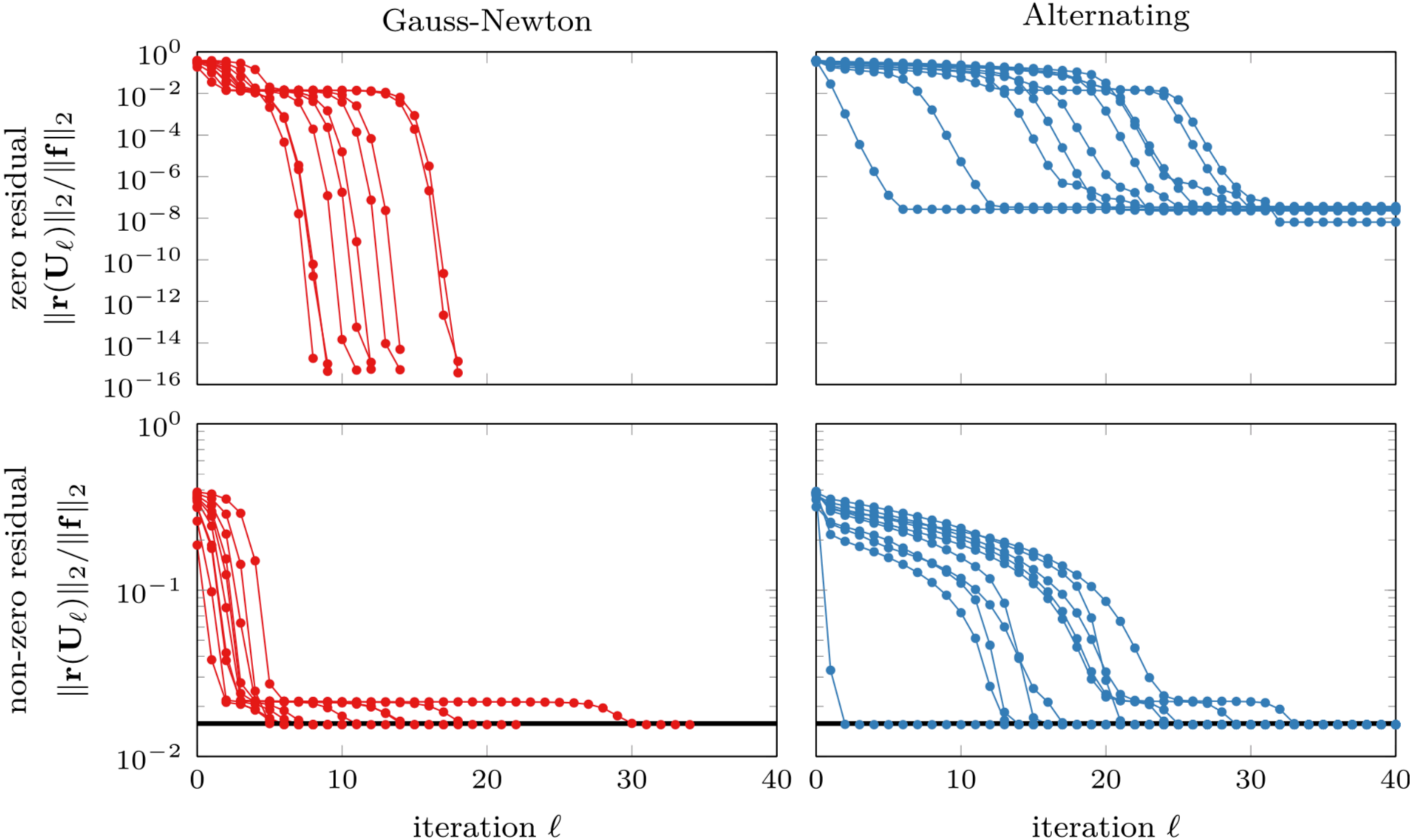
Given subspace, fit polynomial

Given polynomial coefficients, minimize over subspace

Repeat

Constantine, Eftekhari, Hokanson, and Ward (2017), Hokanson and Constantine (2018)

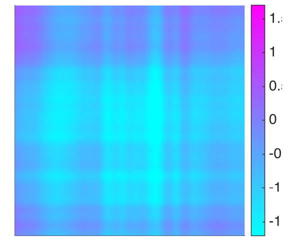
$$f(\mathbf{x}) = (\mathbf{e}_1^\top \mathbf{x})^2 + (\mathbf{1}^\top \mathbf{x}/10)^3 + 1; \quad \mathcal{D} = [-1,1]^{10}$$



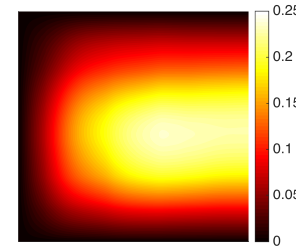
Hokanson and Constantine (2018)

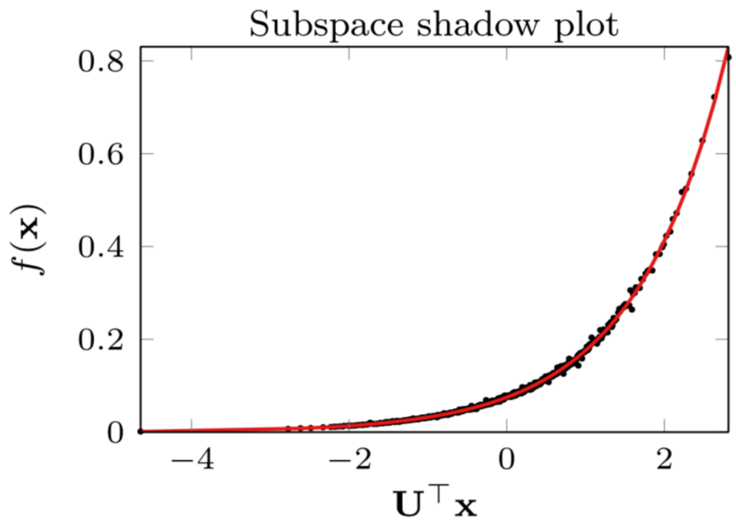
$$
\begin{aligned}
-\nabla \cdot (a \nabla u) = 1, &\quad \mathbf{s} \in \mathcal{D} \\
u = 0, &\quad \mathbf{s} \in \Gamma_1 \\
-\mathbf{n} \cdot a \nabla u = 0, &\quad \mathbf{s} \in \Gamma_2
\end{aligned}
$$

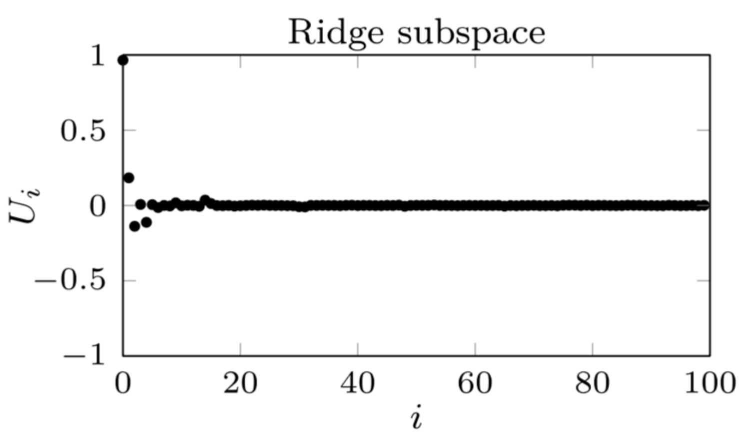
Input field

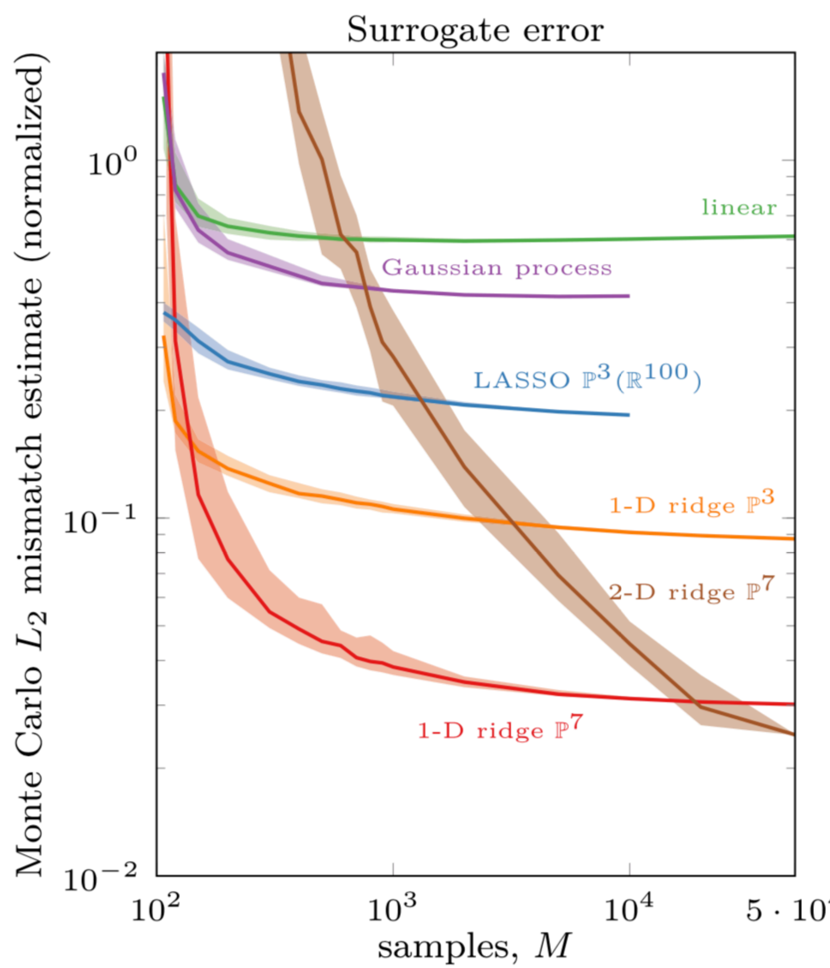


Solution









Hokanson and Constantine (2018)

SUMMARY :: Why I like ridge structure

(1) Exploitable

+ for dimension reduction, not just cheap surrogate

(2) Insights

+ which variables are *important*

(3) Discoverable / checkable

+ eigenvalues

+ non-residual metrics: $\mathbb{E}[\,\mathrm{Var}[\,f\,|\,\boldsymbol{U}^T\mathbf{x}\,]\,]$

+ **plots** in 1 and 2d

TAKE HOMES

The best way to fight the curse of dimensionality is to reduce the dimension!

There are many notions of *important subspaces*; they arise in several applications

Important subspaces are discoverable and exploitable for answering science questions

My group is busy!

Jeff Hokanson
(postdoc)

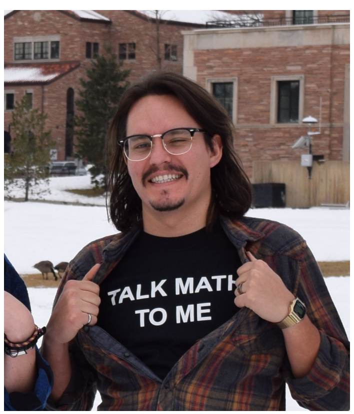
Ridge approximations in DUU

Lipschitz matrix for dimension reduction

#jeffneedsajob

Izzy Aguiar
(MS 2018)

Dynamic active subspaces for parameterized ODEs



Zach Grey
(PhD 2019)

Manifold extensions of active subspaces

Shape design

Andrew Glaws
(PhD 2018)

Sufficient dimension reduction for CS&E

Energy applications

QUESTIONS?

Are there other options for important directions?

What is the **trade-off** between discovering the low-dimensional structure vs. solving the original problem?

Why are these structures so pervasive?

What if my model doesn't fit your setup?

(no gradients, multiple outputs, correlated inputs, …)

PAUL CONSTANTINE
Assistant Professor
University of Colorado Boulder
`activesubspaces.org`
`@DrPaulynomial`

Active Subspaces
SIAM (2015)