

# Gappy Data Reconstruction and Applications in Archaeology

Robert P. Stephan<sup>1</sup>    Kevin T. Carlberg<sup>2</sup>

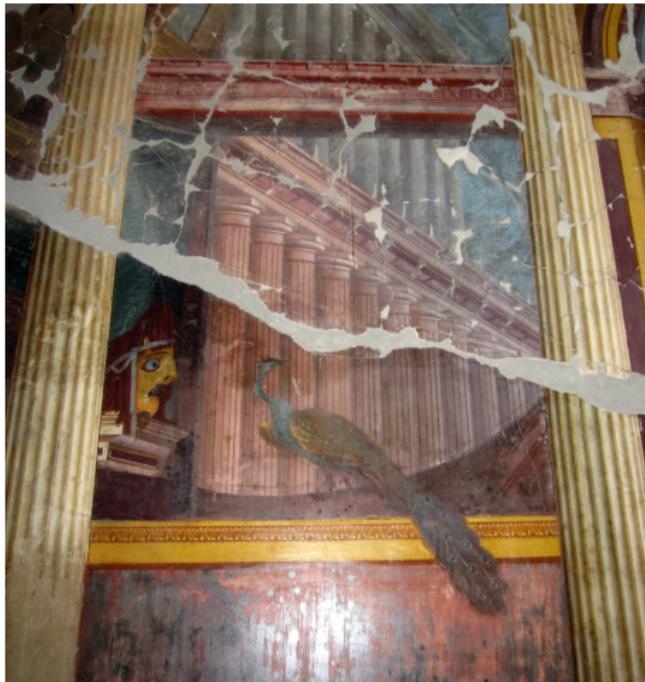
Stanford University

<sup>1</sup>Department of Classics  
rstephan@stanford.edu

<sup>2</sup>Department of Aeronautics & Astronautics  
carlberg@stanford.edu

XXXVIII Annual Conference on Computer Applications &  
Quantitative Methods in Archeology  
April 8, 2010

# Impressionistic archaeology



- 1 Goal
- 2 "Standard" regression
- 3 Gappy Proper Orthogonal Decomposition
- 4 Case Study
- 5 Conclusions

# Goal

- **Archaeology**: the study of the past through fragmentary material remains.
- Want a method to complete the fragments that is:
  - Quantitative
  - Employs as few arbitrary modeling choices as possible
  - Lends insight into the archeological record

## Definitions

- **Sample:** A sample  $A$  is represented by an  $n$ -vector:

$$x(A) = \begin{bmatrix} x_1(A) \\ \vdots \\ x_n(A) \end{bmatrix}$$

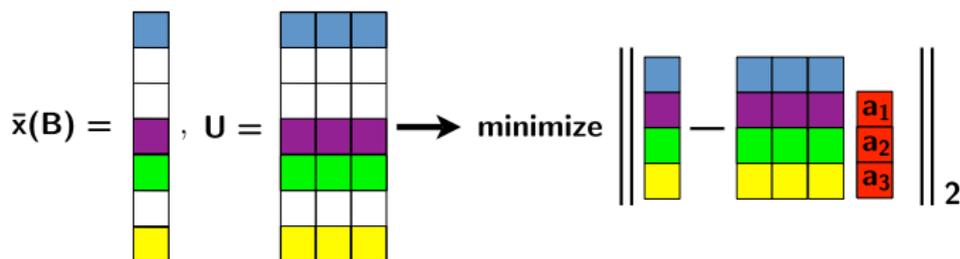
- **Attribute:** A attribute is one of the  $n$  entries of the sample (e.g.  $x_1(A)$ =number of rooms in house  $A$ )
- **Complete samples:** The set of  $m$  samples where all attributes are known (e.g. complete excavation)
- **Gappy sample:** A sample where some attributes are unknown or missing (e.g. incomplete excavation)

## "Standard" regression



- **Supervised learning** method: maps inputs to outputs
- Build a model for each combination of known (inputs) and unknown (outputs) attributes.

## Example: $n = 4$ attributes, $m = 3$ complete samples



$$E [x_1(B_1)] = a_1 + a_2 x_2(B_1) + a_3 x_3(B_1)$$

$$E [x_4(B_1)] = b_1 + b_2 x_2(B_1) + b_3 x_3(B_1)$$

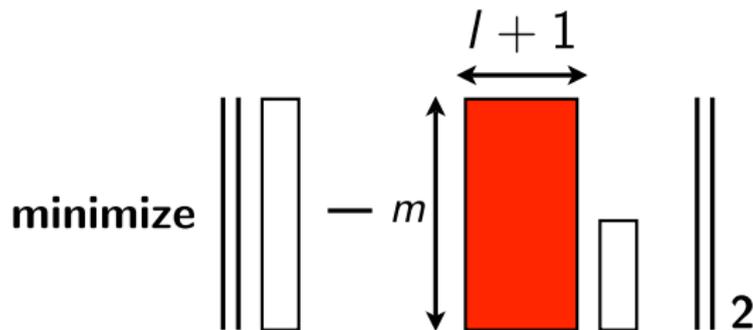
$$E [x_2(B_2)] = c_1 + c_2 x_1(B_2) + c_3 x_4(B_2)$$

$$E [x_3(B_2)] = d_1 + d_2 x_1(B_2) + d_3 x_4(B_2)$$

- × Basis functions arbitrarily chosen (linear, why not quadratic?)
- × Need  $n = 4$  different models ( $a, b, c, d$ )

## Dimension of the system

- Least-squares matrix (linear regression) is  $m \times (l + 1)$ , with  $l \leq n$  the number of known attributes



- **Matrix** must be "skinny" do avoid being underdetermined
- ✓ OK when more full samples than attributes ( $m > n$ )
- ✗ Cannot handle the case with more attributes than full samples ( $n > m$ )

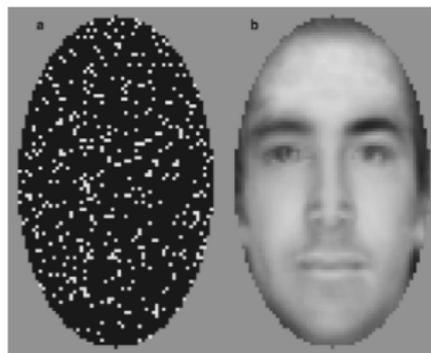
## Problems and a proposed solution

- "Standard" regression
  - 1 Arbitrarily-chosen basis functions
  - 2 Requirement: more complete samples than attributes
  - 3 New model for each gappy sample
- Gappy Proper Orthogonal Decomposition (POD): proposed method
  - 1 Basis functions are derived empirically ✓
  - 2 Can handle the case with more attributes than samples ✓
  - 3 One model for all gappy samples ✓

**Additional benefit:** "eigensamples" which lend insight into the archaeological record

## Gappy POD—previous applications

- Facial image reconstruction [Everson and Sirovich, 1995]



- Aerodynamic flow field reconstruction [Bui-Thanh et al., 2003]
- Optimization [Robinson et al., 2006]
- Real-time simulation of large-scale nonlinear systems [Carlberg et al., 2010]

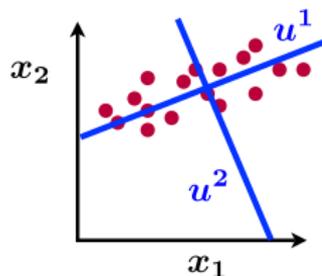
## Method overview

A **supervised** learning method based on an **unsupervised** technique:

- 1 Compute **eigensamples**: vectors which best represent the complete sample set
- 2 Represent the gappy sample using the eigensamples by trying to **match the known attributes**
- 3 Use this representation to **estimate the missing attributes** of the gappy sample

## Step 1: compute eigensamples

- **Eigensamples:** the principal components of the  $m$  complete samples  $\{A_1, \dots, A_m\}$



- **Efficiently obtain** them via the singular value decomposition of the scaled full sample matrix:

$$[\bar{x}(A_1) \ \cdots \ \bar{x}(A_m)] = [u^1 \ \cdots \ u^m] \Sigma V^T$$

- ✓ **Empirically derived** basis functions (unsupervised learning method): not arbitrary

## Eigensample interpretation

- Mathematically characterize dominant attributes in the data set
- Most members of the data set can be well-represented using the first few eigensamples



Figure: "Eigenfaces" [Delac et al., 2005]

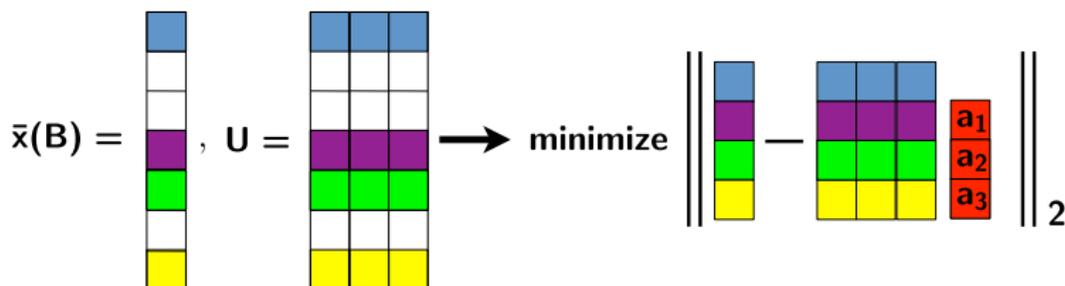
- ✓ Lends additional insight into archaeological record

## Step 2: match known attributes

- Represent the gappy sample with the first  $p \leq m$  eigensamples

$$E[\bar{x}(B)] \approx \sum_{j=1}^p u^j a_j(B)$$

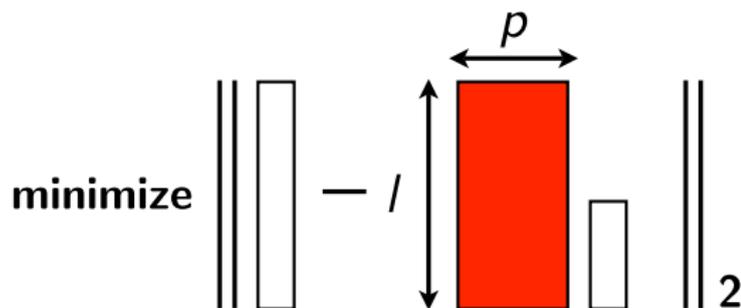
- ✓ Same model for all gappy samples
- Compute coefficients  $a_j(B)$  to match known attributes



- Equivalent to **least-squares regression** in one discrete variable with eigensamples as basis functions

## Dimension of the system

- Least-squares matrix is  $l \times p$ , with  $l \leq n$ ,  $p \leq m$



- ✓ Can handle the case with more attributes than full samples ( $n > m$ )

## Step 3: compute expectation and variance

- The  $i^{\text{th}}$  attribute of the gappy sample has expectation

$$E[x_i(B)] = \mu_i + s_i \sum u_i^j a_j(B)$$

and variance

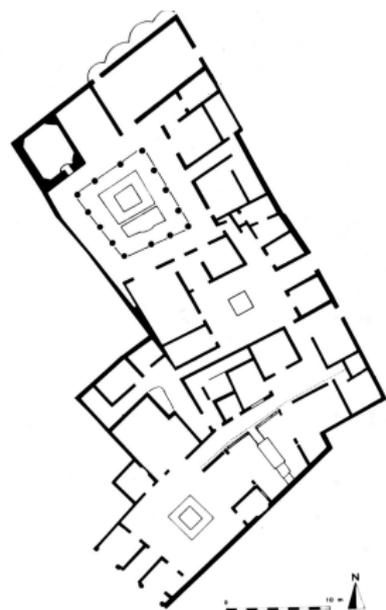
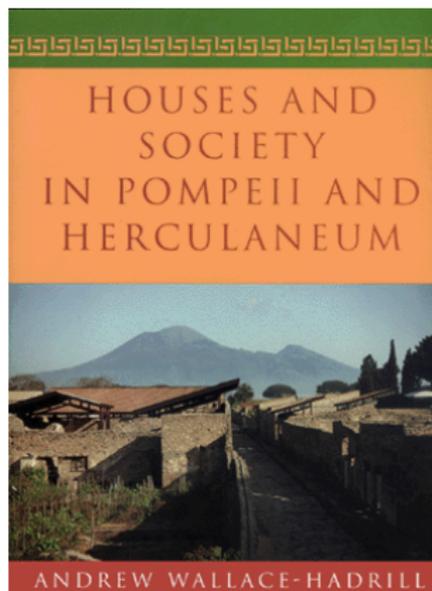
$$\sigma_i^2 = \frac{s_i^2}{I} \sum_{k=1}^I \left( \bar{x}_k(B) - \sum_{j=1}^p u_k^j a_j \right)^2$$

- Can construct **confidence intervals** since least-square regression assumes a Gaussian distribution

## Case study: housing in Campania



# Data set: Pompeii & Herculaneum

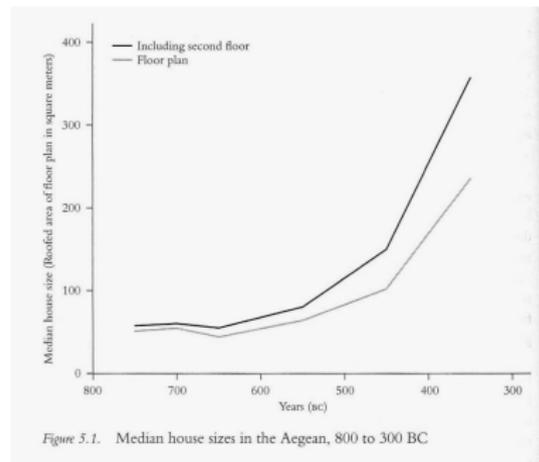


# Attribute selection

$n = 4$  attributes

[Wallace-Hadrill, 1996]:

- 1 House size
- 2 Number of rooms
- 3 Number of decorated rooms
- 4 Distance to the forum



# Methodology

## ■ Insert artificial gaps

	House Nai Size (m2)	# of Room	Number o	Proximity to Forum (m)	
1					
2	1.6.2/16	1200	16	1	116
3	1.6.4	400	14	8	110
4	1.6.11	500	11	6	101
5	1.6.13/14	275	13	0	121
6	1.6.15	300	10	9	125
7	1.7.1	790	16	7	119
8	1.7.2/3	125	8	7	120
9	1.7.7	230	11	6	127
10	1.7.10-12	660	19	4	145
11	1.7.19	330	14	9	137
12	1.8.1-3	550	9	2	119
13	1.8.4-6	390	15	3	138
14	1.8.14	190	8	3	164
15	1.8.17/11	500	17	11	150
16	1.8.18	225	9	1	147
17	1.9.1/2	475	12	5	145
18	1.9.3/4	490	10	4	148
19	1.9.5-7	425	12	7	150
20	1.9.9	120	5	0	169
21	1.9.10	220	9	0	176
22	1.9.13/14	500	12	8	173
23	1.10.4	1700	28	19	129
24	1.10.7	310	9	6	125
25	1.10.8	270	10	3	122
26	1.10.10-11	470	17	9	136

	House Nai Size (m2)	# of Room	Number o	Proximity to Forum (m)	
1					
2	1.6.2/16	1200	16	1	116
3	1.6.4	400	14	8	110
4	1.6.11	████████	11	6	101
5	1.6.13/14	275	13	0	121
6	1.6.15	300	10	9	125
7	1.7.1	790	████████	7	119
8	1.7.2/3	125	8	7	120
9	1.7.7	230	11	6	127
10	1.7.10-12	660	19	████████	145
11	1.7.19	330	14	9	137
12	1.8.1-3	550	9	2	119
13	1.8.4-6	390	████████	3	████████
14	1.8.14	190	8	3	164
15	1.8.17/11	500	17	11	150
16	1.8.18	225	9	1	147
17	1.9.1/2	████████	12	5	145
18	1.9.3/4	490	10	4	148
19	1.9.5-7	425	12	7	150
20	1.9.9	120	5	0	169
21	1.9.10	220	████████	████████	176
22	1.9.13/14	500	12	8	173
23	1.10.4	1700	28	19	129
24	1.10.7	310	9	6	████████
25	1.10.8	270	10	3	122
26	1.10.10-11	470	17	9	136

## ■ Check accuracy of Gappy POD prediction

## Complete set

- Herculaneum houses treated as complete (training) set
- $m = 29$  complete houses
- $p = 1$  eigensample retained (only 4 attributes)
- 99% confidence intervals

## Gappy set 1: Herculaneum

- Gappy attribute: number of rooms
- Correct predictions for 62.1% of houses (18/29)
- Example: *Casa del Papirio dipinto*
  - Correct number of rooms: 8
  - ✓ Confidence interval: [7.4, 10] rooms

## Gappy set 2: Pompeii

- Gappy attribute: area of house
- Correct predictions for 47.3% of houses (44/93)
- Example: I.6.4
  - Correct area: 400  $m^2$
  - ✓ Confidence interval: [397, 436] area ( $m^2$ )

## "Eigenhouses"

- First eigenhouse: most houses in the set have a value of  $c$  that closely describes it

- Pompeii

$$\begin{bmatrix} 409.6 \\ 12.2 \\ 5.0 \\ 111.0 \end{bmatrix} + c \begin{bmatrix} -0.9982 \\ -0.0184 \\ -0.0139 \\ 0.0553 \end{bmatrix}$$

- Herculaneum

$$\begin{bmatrix} 312.9 \\ 12.1 \\ 5.4 \\ 92.9 \end{bmatrix} + c \begin{bmatrix} -0.9984 \\ -0.0180 \\ -0.0146 \\ -0.0506 \end{bmatrix}$$

## "Eigenhouse" representation

- VI.15.1/27 (Pompeii)

$$\text{true : } \begin{bmatrix} 1100 \\ 24 \\ 15 \\ 79 \end{bmatrix}, \text{ with } c = -676.2 : \begin{bmatrix} 1070.4 \\ 24.6 \\ 14.6 \\ 73.4 \end{bmatrix}$$

- 3.11,8 (Herculaneum)

$$\text{true : } \begin{bmatrix} 520 \\ 21 \\ 14 \\ 119 \end{bmatrix}, \text{ with } c = -2.3357 : \begin{bmatrix} 673.0 \\ 19.9 \\ 12.1 \\ 129.3 \end{bmatrix}$$

# Conclusions

- New method for filling in lacunose data
- Basis functions are empirical
- Handles cases where “standard” regression breaks down (more attributes than samples)
- A uniform model for all samples
- Eigensamples lend valuable insight into the archaeological record

## Future directions

- Expand the current study
- Different scales
- More attributes than samples (better-suited for the method, "standard" regression breaks down)



# References I



Bui-Thanh, T., Murali, D., and Willcox, K. (2003).

Proper orthogonal decomposition extensions for parametric applications in compressible aerodynamics.

*AIAA Paper 2003-4213, 21st Applied Aerodynamics Conference.*



Carlberg, K., Farhat, C., and Bou-Mosleh, C. (2010).

Discrete Nonlinear Model Reduction via Least-Squares Petrov-Galerkin Projection and Compressive Tensor Approximation.

*in preparation.*



Delac, K., Grgic, M., and Liatsis, P. (2005).

Appearance-based statistical methods for face recognition.

*In 47th International Symposium ELMAR, Zadar, Croatia, pages 8–10. Citeseer.*

## References II



Everson, R. and Sirovich, L. (1995).

Karhunen-Loeve procedure for gappy data.

*Journal of the Optical Society of America A*, 12(8):1657–1664.



Robinson, T., Eldred, M., Willcox, K., and Haines, R. (2006).

Strategies for multifidelity optimization with variable dimensional hierarchical models.

*AIAA Paper 2006-1819, 47th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference.*



Wallace-Hadrill, A. (1996).

*Houses and society in Pompeii and Herculaneum.*

Princeton Univ Pr.

# Questions?