Comparing Two Dimension-Reduction Methods for Network Simulation Models

Kevin Mills & Jim Filliben, NIST

Winter Simulation Conference – Dec. 6, 2010

Outline

- Goal – Problem – Solution
- Dimension Reduction for Multivariate Responses
- Overview of Candidate *MesoNet* Responses
- Reduction via Correlation Analysis with Clustering
- Reduction via Principal Components Analysis
- Characterization & Comparison of Dimension Reduction Techniques
- Findings

Goal – Problem – Solution

- **Goal** – compare proposed Internet congestion control algorithms under a wide range of controlled, repeatable conditions, as simulated by selecting combinations of parameter values for *MesoNet*, a 20-parameter TCP/IP network model

- **Problem** – how to determine which *MesoNet* responses to analyze when characterizing model behavior

- **Solution** – apply and evaluate two techniques: correlation analysis with clustering & principal components analysis

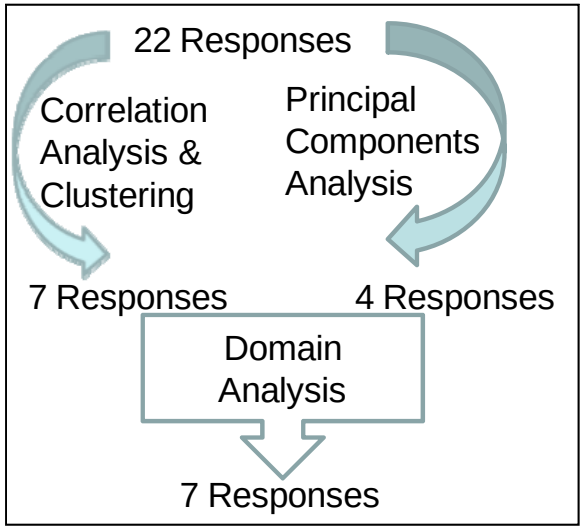
Scale Reduction: Theory & Practice

Simulating large, fast networks across many conditions and congestion control algorithms requires scale reduction in both model parameters & responses

$$y_1, \ldots, y_z = f(x_{1|[1,\ldots,\ell]} \ldots, x_{p|[1,\ldots,\ell]})$$

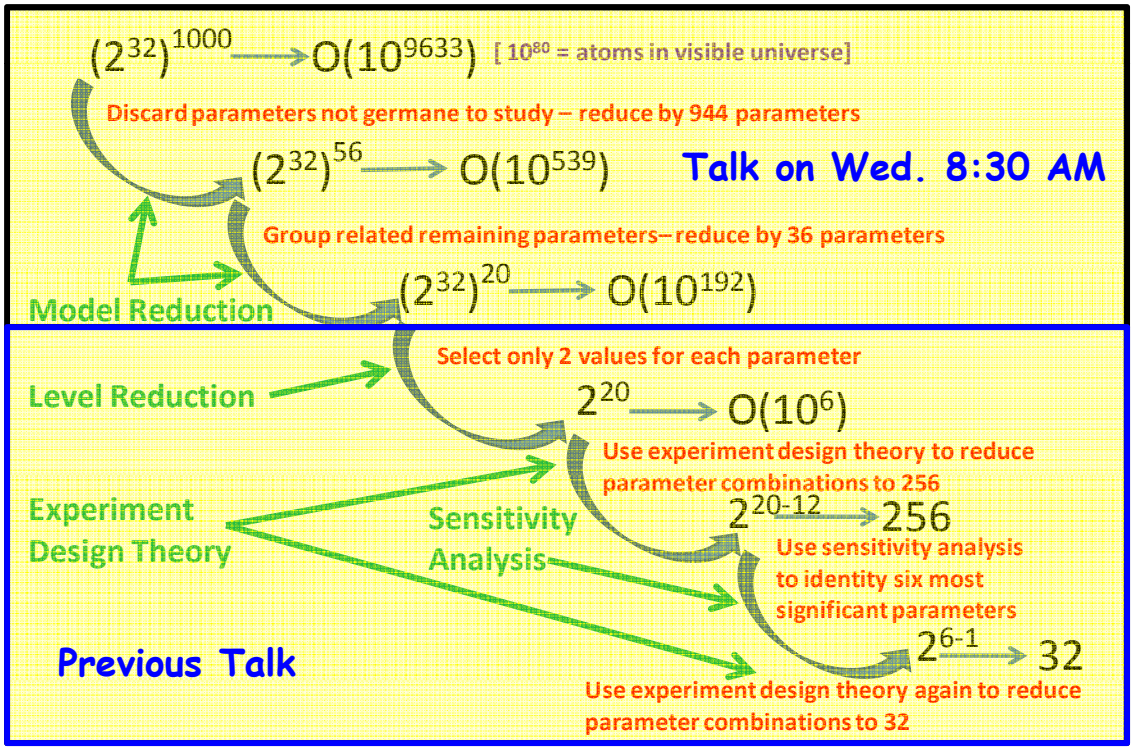
Response State-Space — Stimulus State-Space

Multidimensional Response Reduction



This Talk

Parameter Reduction



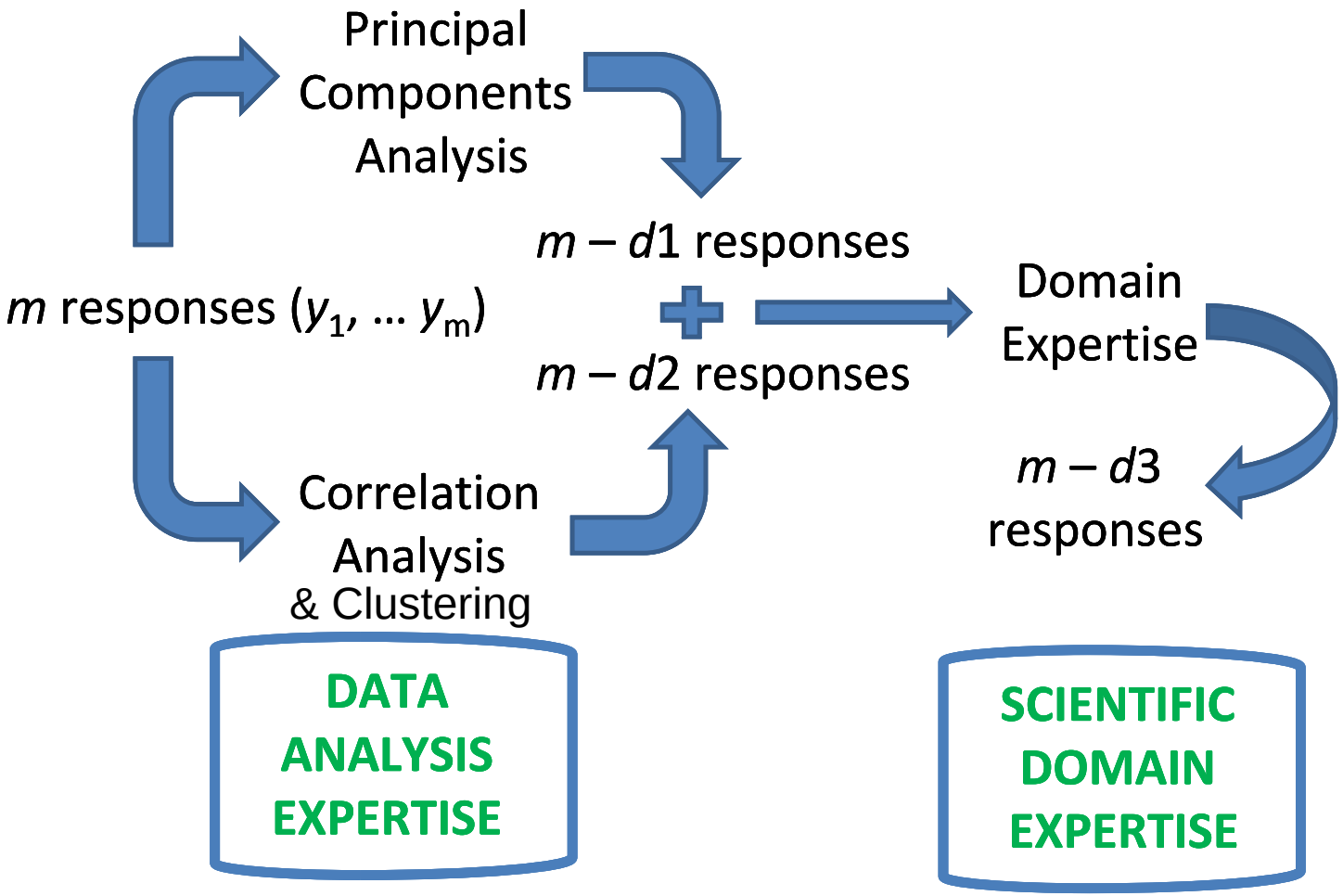
The Dimension-Reduction Problem for Multivariate Responses

Fodor, I.K. 2002. A Survey of Dimension Reduction Techniques. Lawrence Livermore National Laboratory Technical Report no. UCRL-ID-148494:

"given the r-dimensional random variable $x = (x_1, \ldots, x_r)^T$, find a lower dimensional representation, $s = (s_1, \ldots, s_k)^T$ with $k \leq r$, that captures the content in the original data, according to some criterion."

Fodor identifies principal components analysis (PCA) as the best (in terms of mean-square error) linear dimension reduction technique.

We Applied & Compared 2 Different Techniques



Overview of (*r* = 22) Candidate *MesoNet* Responses

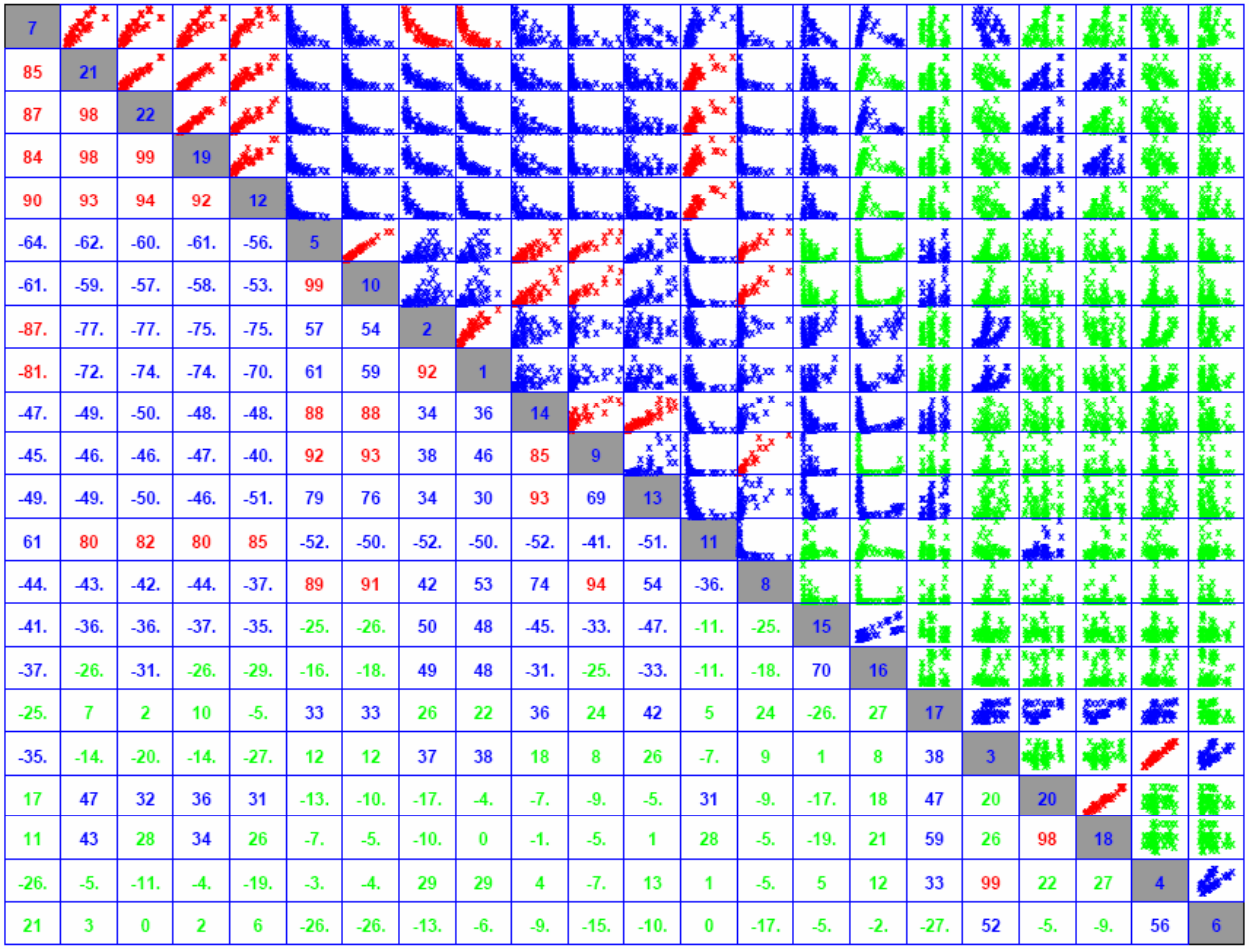
Response	Definition
y1	Active Sources – sources attempting to transfer data
y2	Proportion of total sources that are active: Active Flows/All Sources
y3	Data packets entering the network per second
y4	Data packets leaving the network per second
y5	Packet Loss Rate: y4/(y3+y4)
y6	Flows Completed per second
y7	Flow Completion Rate: y6/(y6+y1)
y8	Connection Failures per second
y9	Connection Failure Rate: y8/(y8+y1)
y10	Retransmission Rate (ratio)
y11	Average Per Flow Congestion Window (packets)
y12	Average Number of Increases in Congestion Window Per Flow Per Second
y13	Average Number of Negative Acknowledgments Per Flow Per Second
y14	Average Number of Timeouts Per Flow Per Second
y15	Average Round-trip Time (ms)
y16	Relative Queuing Delay (y15/average propagation delay)
y17	Average Throughput (packets/second) for DD Flows
y18	Average Throughput (packets/second) for DF Flows
y19	Average Throughput (packets/second) for DN Flows
y20	Average Throughput (packets/second) for FF Flows
y21	Average Throughput (packets/second) for FN Flows
y22	Average Throughput (packets/second) for NN Flows

64 x 22 Multivariate Data Set Resulting from a 2^{11-5} Orthogonal Fractional Factorial Experiment Design

Run	y1	y2	…	y21	y22
1	4680.619	0.168126	…	92.034	89.785
2	6654.512	0.239371	…	72.596	57.738
3	9431.405	0.339259	…	29.569	13.963
4	11565.81	0.415439	…	23.427	19.882
…	…	…	…	…	…
61	10319.55	0.247471	…	87.969	41.573
62	1738.469	0.093668	…	159.298	161.602
63	1783.509	0.096094	…	148.395	161.36
64	21467.6	0.514811	…	26.159	9.981

Reduction via Correlation Analysis & Clustering (CAC)

Correlation Analysis – Part (a)

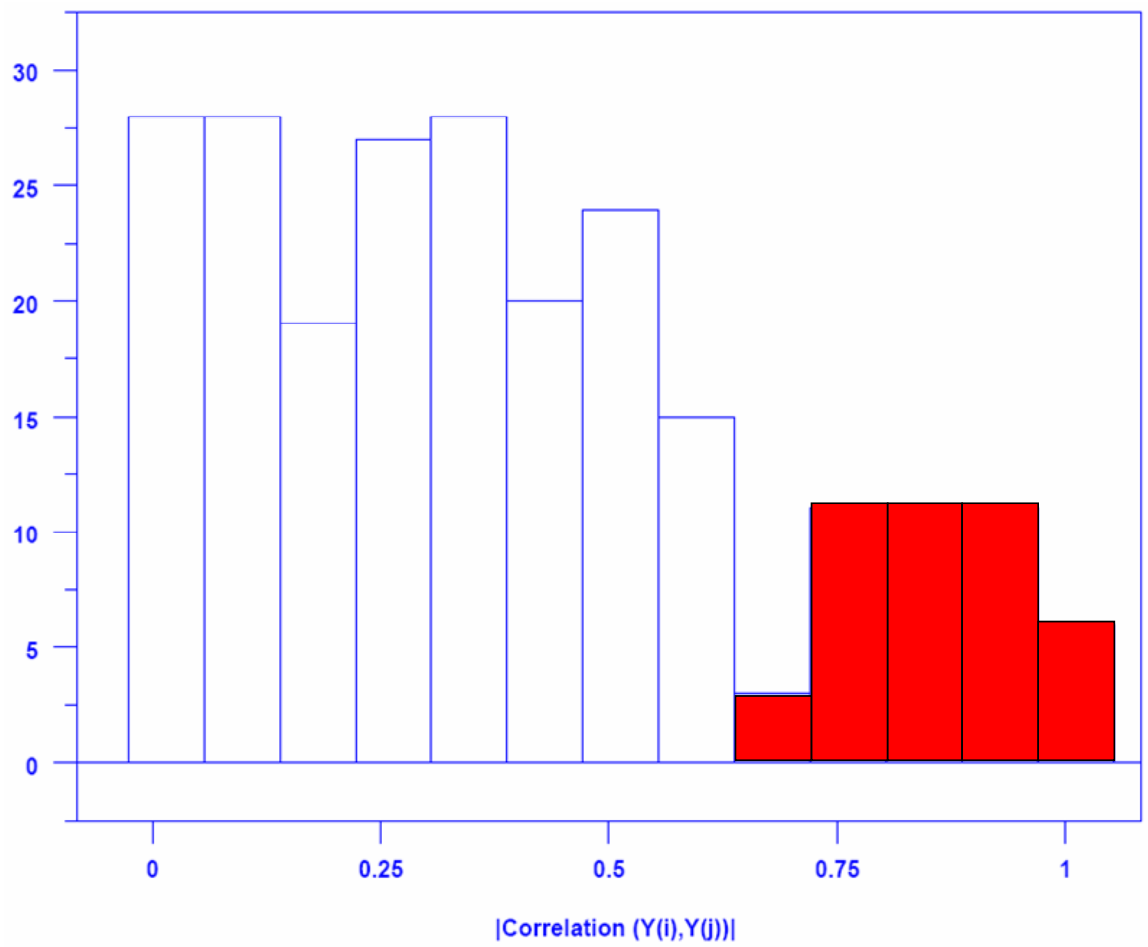


7																			
85	21																		
87	98	22																	
84	98	99	19																
90	93	94	92	12															
-64.	-62.	-60.	-61.	-56.	5														
-61.	-59.	-57.	-58.	-53.	99	10													
-87.	-77.	-77.	-75.	-75.	57	54	2												
-81.	-72.	-74.	-74.	-70.	61	59	92	1											
-47.	-49.	-50.	-48.	-48.	88	88	34	36	14										
-45.	-46.	-46.	-47.	-40.	92	93	38	46	85	9									
-49.	-49.	-50.	-46.	-51.	79	76	34	30	93	69	13								
61	80	82	80	85	-52.	-50.	-52.	-50.	-52.	-41.	-51.	11							
-44.	-43.	-42.	-44.	-37.	89	91	42	53	74	94	54	-36.	8						
-41.	-36.	-36.	-37.	-35.	-25.	-26.	50	48	-45.	-33.	-47.	-11.	-25.	15					
-37.	-26.	-31.	-26.	-29.	-16.	-18.	49	48	-31.	-25.	-33.	-11.	-18.	70	16				
-25.	7	2	10	-5.	33	33	26	22	36	24	42	5	24	-26.	27	17			
-35.	-14.	-20.	-14.	-27.	12	12	37	38	18	8	26	-7.	9	1	8	38	3		
17	47	32	36	31	-13.	-10.	-17.	-4.	-7.	-9.	-5.	31	-9.	-17.	18	47	20	20	
11	43	28	34	26	-7.	-5.	-10.	0	-1.	-5.	1	28	-5.	-19.	21	59	26	98	18
-26.	-5.	-11.	-4.	-19.	-3.	-4.	29	29	4	-7.	13	1	-5.	5	12	33	99	22	27
21	3	0	2	6	-26.	-26.	-13.	-6.	-9.	-15.	-10.	0	-17.	-5.	-2.	-27.	52	-5.	-9.

Red $80 \geq |r| \times 100 \leq 100$ **Blue** $30 \geq |r| \times 100 < 80$ **Green** $|r| \times 100 < 30$

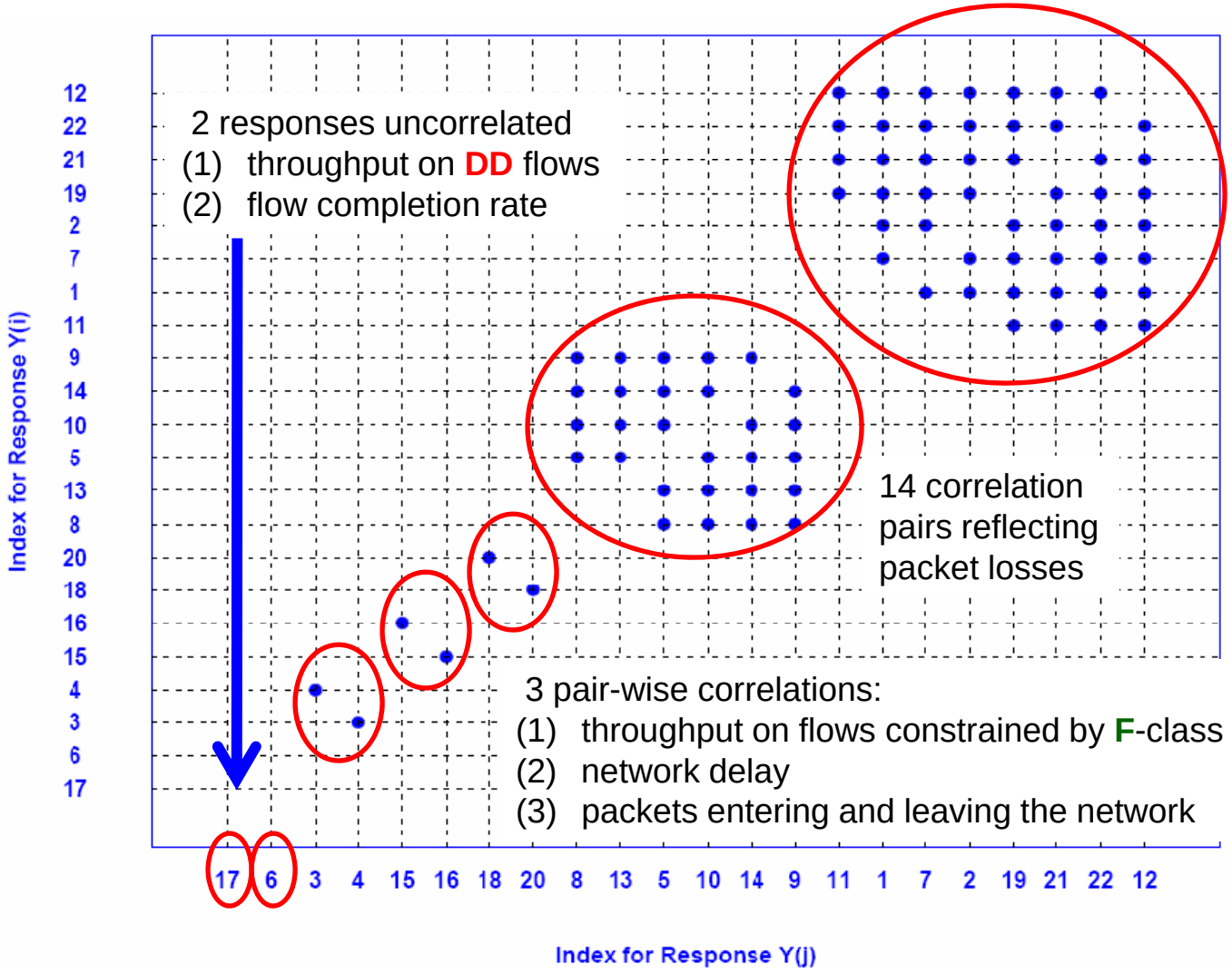
Pair-wise Correlation Matrix

Correlation Analysis – Part (b)



Histogram: bins where |r| > 0.65 highlighted in red

Clustering: Response Index-Index Plot where $|r_{i,j}| > 0.65$ Clustered into Mutual Correlations



Plot suggests *MesoNet* exhibits 7 distinct behavioral dimensions

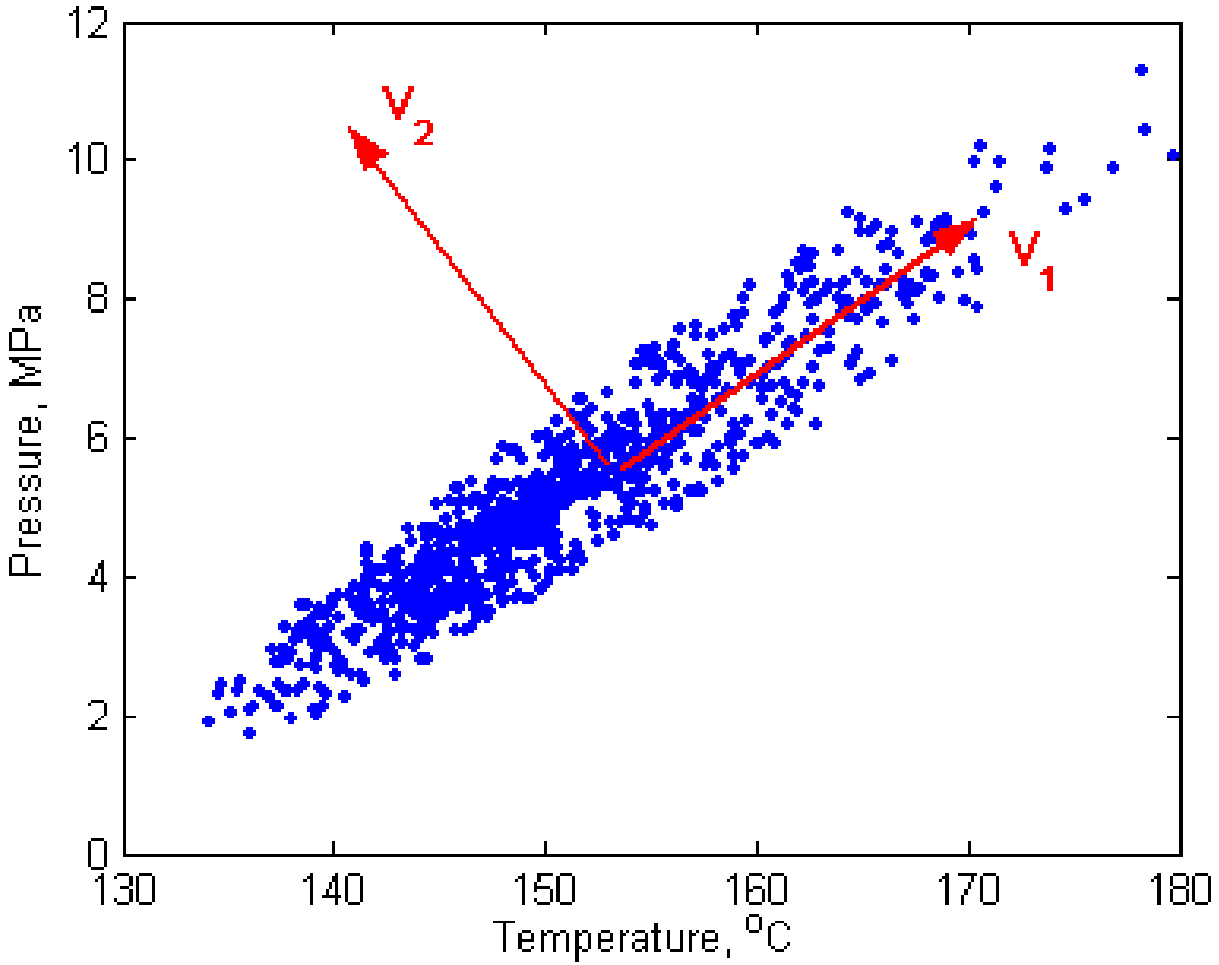
Correlation Analysis & Clustering Suggests *MesoNet* Behavior Can be Represented using only 7 Responses

CAC

Dimension	**Responses**
Congestion	y1, y2, y7, y11, y12, y19, y21, **y22**
Losses	y5, y8, y9, **y10**, y13, y14
Delay	**y15**, y16
F-class Throughput	y18, **y20**
D-class Throughput	**y17**
Packet Throughput	y3, **y4**
Flows Per Second	**y6**

Using subjective, domain-specific reasoning, a domain analyst selects one response to represent each cluster

Principal Components Analysis



Principal Components Analysis of 22 *MesoNet* Responses

(a) Distribution of Variance

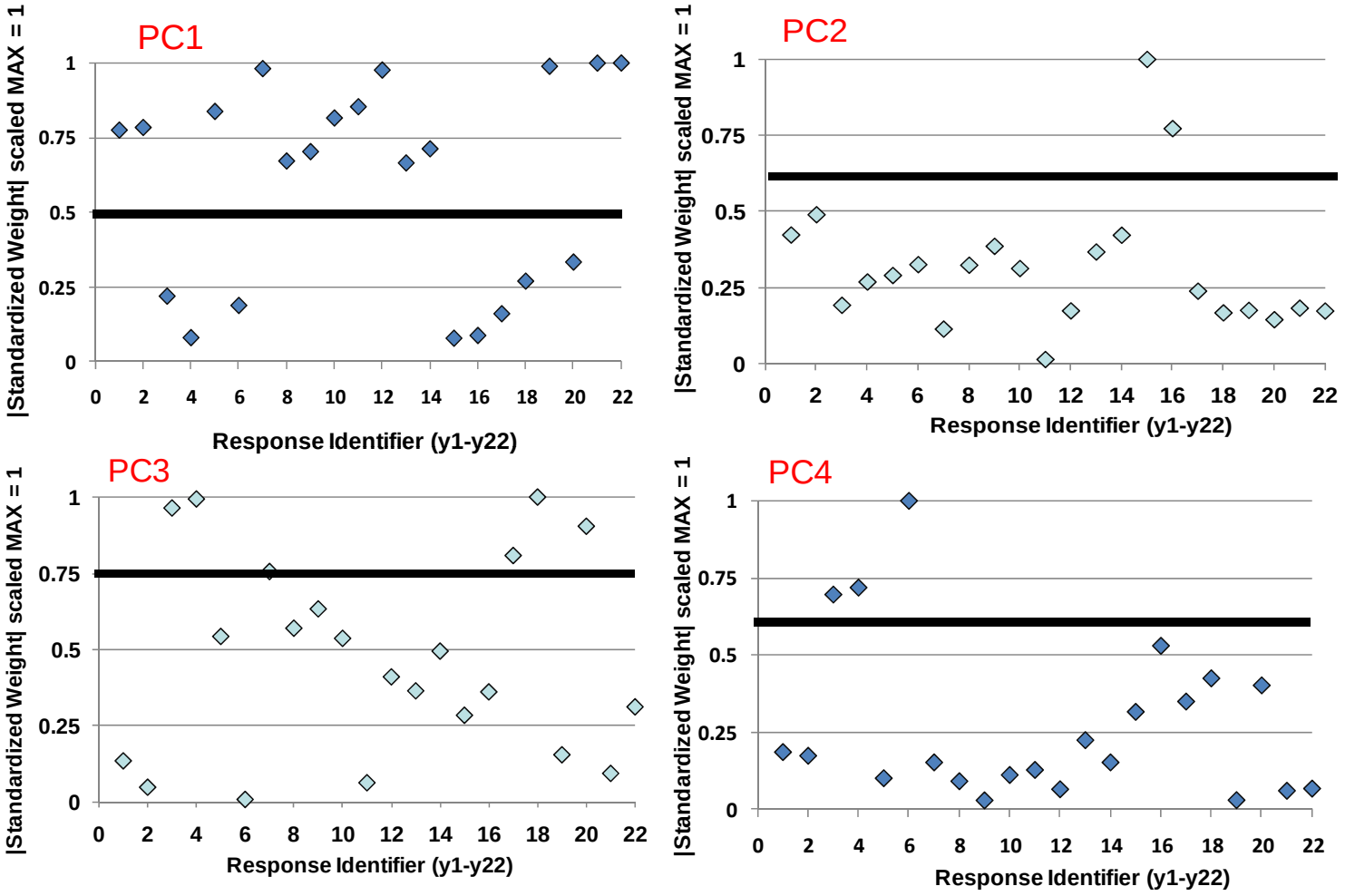
PC	Std. Dev.	Cumulative %
PC1	9.7091	0.441325
PC2	4.0161	0.62388
PC3	3.2322	0.77079
PC4	2.0630	0.86457
PC5	0.9716	0.90873
PC6	0.7585	0.94321
PC7	0.4537	0.96383
PC8	0.2569	0.97551
PC9	0.1835	0.98385
PC10	0.1254	0.98955
PC11	0.0588	0.99222
PC12	0.0504	0.99451
PC13	0.0360	0.99615
PC14	0.0286	0.99745
PC15	0.0225	0.99847
PC16	0.0113	0.99899
PC17	0.0099	0.99944
PC18	0.0060	0.99971
PC19	0.0032	0.99985
PC20	0.0019	0.99994
PC21	0.0011	0.99999
PC22	0.0002	1.00000

(b) Weight Vector Matrix

Response	PC1	PC2	PC3	PC4	…	PC22
y1	0.2581	0.1421	-0.1667	0.1126	…	-0.0236
y2	0.2608	0.1768	-0.1407	0.1058	…	-0.0002
y3	0.0864	0.0218	-0.4163	-0.3909	…	0.7017
y4	0.0435	0.0617	-0.4252	-0.4037	…	-0.6919
y5	0.2774	-0.2276	0.0363	0.0644	…	-0.0601
y6	-0.039	0.0914	-0.1246	-0.5636	…	0.0078
y7	-0.284	-0.1358	0.1004	-0.0815	…	-0.0072
y8	0.2261	-0.2452	0.0444	0.0589	…	-0.0472
y9	0.2358	-0.2775	0.0635	0.0236	…	0.0128
y10	0.2706	-0.2395	0.0344	0.0704	…	0.0401
y11	-0.245	-0.0838	-0.1081	0.0797	…	0.0026
y12	-0.283	-0.167	-0.0036	0.0441	…	-0.0061
y13	0.2241	-0.2678	-0.0174	-0.1225	…	-0.0005
y14	0.2387	-0.2965	0.0218	-0.0815	…	-0.0154
y15	0.0429	0.4426	-0.0415	0.1868	…	0.01
y16	0.0457	0.3242	-0.2348	0.3085	…	0.0236
y17	0.0682	-0.2008	-0.3694	0.2058	…	-0.0279
y18	-0.064	-0.1636	-0.4271	0.2485	…	0.0531
y19	-0.287	-0.1679	-0.0805	0.024	…	0.0433
y20	-0.084	-0.1519	-0.3983	0.2356	…	-0.062
y21	-0.29	-0.1716	-0.0989	0.0413	…	0.0704
y22	0.2581	-0.1667	-0.0332	0.0308	…	-0.074

Most response variance appears to be accounted for by the first 4 or 5 principal components (depending on the threshold selected), we highlight components 6 and 7 to align with the number of responses suggested by the *CAC* analysis

One Approach to Identify Variables to Include in Each PC



Principal Components Analysis Suggests *MesoNet* Behavior Can be Represented using only 4 Responses

PCA

Dimension (PC)	Responses
Congestion	y1, y2, y5, y7, y8, y9, y10, y11, y12, y13, y14, y19, y21, y22
Delay	y15, y16
D-class & F-class Throughput	y3, y4, y17, y18, y20
Flows Per Second	y3, y4, y6

How many responses should be used to account for the PCs?

Jolliffe (2002) suggests that the number of variables to select should equal the number of PCs that account for the most variance (4-5 here), or perhaps a few more (6 or 7 here).

How might an analysis select responses to represent PCs?

(1) MKB heuristic [Mardia, Kent and Bibby,1995]: iterates over the weight vectors from the least significant (PC22 here) to the most significant (PC1). When examining each weight vector, the response corresponding to the weight with the largest magnitude is discarded from further consideration. This process continues until the remaining number of responses corresponds to the number (7 here) of variables sought.

(2) JJF heuristic inverts the MKB heuristic: iterates over the weight vectors from the most significant to the least significant. When examining each weight vector, the variable corresponding to the weight with the largest magnitude is withdrawn from further consideration. This process continues until the cardinality of the set of withdrawn responses equals the number of variables sought. The set of withdrawn responses constitutes the responses to be used in subsequent analyses.

Comparing Correlation & PCA Results

PC	JJF Heuristic	MKB Heuristic	CAC Domain Analysis
PC1	Y22 (NN flow TP)	Y4 (packets output)	Y22 (NN flow TP)
PC2	Y15 (SRTT)	Y19 (DN flow TP)	Y15 (SRTT)
PC3	Y18 (DF flow TP)	Y18 (DF flow TP)	Y4 (packets output)
PC4	Y6 (flows completed TP)	Y6 (flows completed TP)	Y6 (flows completed TP)
PC5	Y11 (CWND)	Y9 (connection failure rate)	Y10 (retransmission rate)
PC6	Y17 (DD flow TP)	Y17 (DD flow TP)	Y20 (FF flow TP)
PC7	Y16 (queuing delay)	Y16 (queuing delay)	Y17 (DD flow TP)
PC8	Y13 (NAKs)	Y13 (NAKs)	N/A
PC9	Y1 (active sources)	Y11 (CWND)	N/A
PC10	Y14 (timeouts)	Y15 (SRTT)	N/A
PC11	Y2 (% sources active)	Y7 (flow-completion rate)	N/A
PC12	Y9 (connection failure rate)	Y2 (% sources active)	N/A
PC13	Y10 (retransmission rate)	Y10 (retransmission rate)	N/A
PC14	Y8 (connection failures)	Y8 (connection failures)	N/A
PC15	Y7 (flow-completion rate)	Y14 (timeouts)	N/A
PC16	Y12 (CWND increases)	Y12 (CWND increases)	N/A
PC17	Y21 (FN flow TP)	Y1 (active sources)	N/A
PC18	Y20 (FF flow TP)	Y21 (FN flow TP)	N/A
PC19	Y19 (DN flow TP)	Y20 (FF flow TP)	N/A
PC20	Y4 (packets output)	Y22 (NN flow TP)	N/A
PC21	Y5 (loss rate)	Y5 (loss rate)	N/A
PC22	Y3 (packets input)	Y3 (packets input)	N/A

Characterization & Comparison of Dimension Reduction Techniques

Correlation Analysis & Clustering

Pros

- Provided effective dimension reduction (22 → 7) through correlations that could be vetted by a domain expert
- Examining response correlations helped to validate *MesoNet*
- Uncovered nuanced differences between flow and packet throughput rates in a network

Cons

- A second 2^{11-5} OFF experiment with different level settings revealed some (valid) differences in correlations – thus separate correlation analyses must be conducted for different level settings

Principal Components Analysis

Pros

- Provided greater dimension reduction (22 → 4 or 5) than correlation analysis & clustering

Cons

- There is no specific domain interpretation of even the top 2 or 3 principal components – in the case shown here we were able to arrive at a reasonable interpretation; in other cases, we were not
- Principal components analysis depends on seemingly arbitrary threshold selections to determine which PCs to include and which responses to use to represent PCs
- Principal components proved coarser than corresponding groupings generated by clustering mutual correlations
- A second 2^{11-5} OFF experiment with different level settings revealed some differences in principal components – such differences are difficult to understand without assistance from other analyses

Correlation Analysis or PCA?

- If limited to one technique, correlation analysis provides results easier for a domain analyst to comprehend
- Principal components take on + and – values, which present domain analysts with difficulty assigning meaning – we had to infer meaning by comparing main effects plots of principal components with main effects plots from responses chosen from groupings established by correlation analysis
- Principal components proved coarser than corresponding groupings generated by clustering mutual correlations
- PCA provides a reasonable complement to correlation analysis by giving a separate view of the data, which should be consistent with correlation results, thus helping to validate a model

Findings

- We investigated correlation and PC analyses as two techniques to reduce the dimension of responses from *MesoNet*, a TCP/IP network simulator
- We demonstrated that both techniques can significantly reduce the dimension of response data
- We found that both techniques could be used to validate a model, but that PCA is more suited to complement correlation analysis
- We found that PCA results are difficult for a domain analyst to interpret without comparison to analyses of individual responses
- We also found that results from correlation and PC analyses with one set of parameter values cannot necessarily be extrapolated to a different set of values generated from different parameter settings