Modeling and Simulation of Wireless Link Quality (ETT) Through Principal Component Analysis of Trace Data

Anh Le, Prashant Krishnamurthy,
David Tipper, Konstantinos Pelechrinis

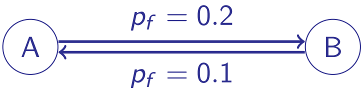
Graduate Telecommunications and Networking Program, University of Pittsburgh

PE-WASUN 2011

Outline

Introduction

- What is Expected Transmission Time (ETT)?

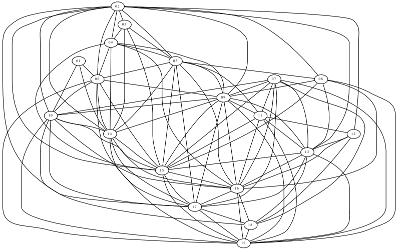


$$ETT = \frac{PacketSize}{Bandwidth} \times \frac{1}{(1 - p_f)(1 - p_r)}$$

  - Used as a metric for wireless link quality
  - No trivial way to simulate ETT values
- What is Principal Component Analysis (PCA)?
  - A matrix decomposition method, used widely in machine learning to reduce dimension of data
- Our Goal: Build a model for ETT simulation
- Our Approach: Using PCA to analyze captured ETT values from a real wireless network

Trace Data

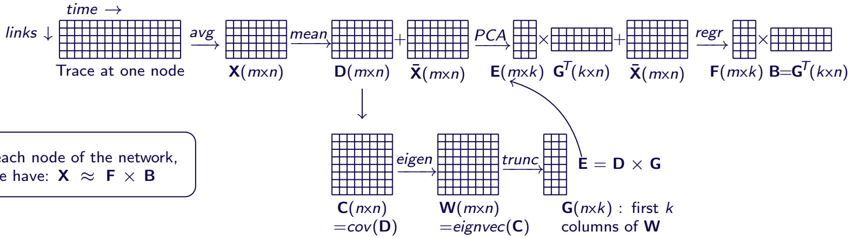
- The ETT trace was collected at UCSB wireless mesh network
  - 802.11a/b
  - 19 nodes, 192 links
  - located on 5 floors of a building



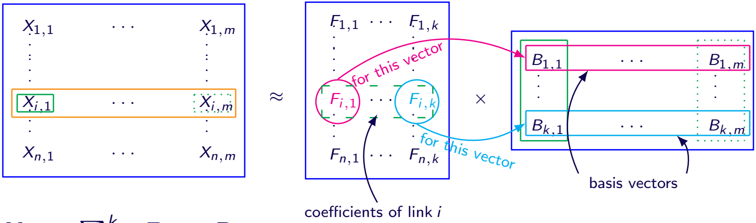
- There are 3 datasets collected at different times.
- Preprocessing:
  - Convert from raw data to a set of matrices containing ETT values
  - Get rid of loss data: use largest continuous block
  - Averaging with a window size of 10min

Outline

Apply PCA to ETT traces



At each node of the network,
we have: $\mathbf{X} \approx \mathbf{F} \times \mathbf{B}$



$$\mathbf{X}_i = \sum_{j=1}^{k} \mathbf{F}_{i,j} \times \mathbf{B}_j$$

Choosing k

We define two indicators to estimate the amount of information lost:

- Coverage α is defined as the cumulative sum of the selected normalized eigenvalues

$$\alpha = \sum_{u=1}^{k} \mathbf{V}[u] / \sum_{u=1}^{n} \mathbf{V}[u]$$

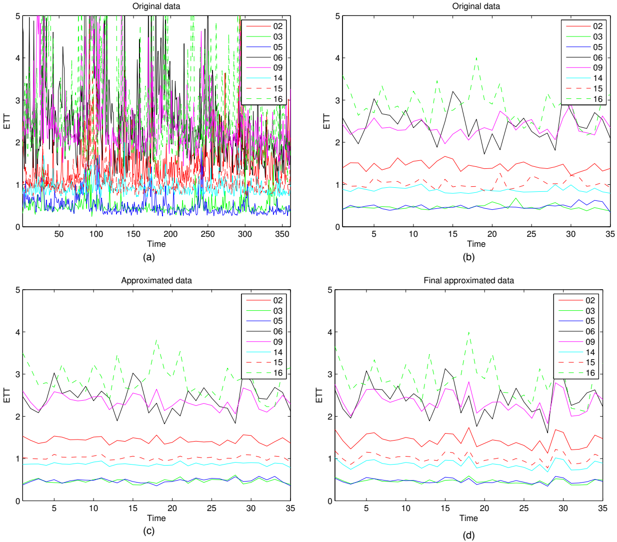
- Loss β is the significance of the last selected eigenvalue

$$\beta = \mathbf{V}[k]/\mathbf{V}[1]$$

If we can choose k to be significantly smaller than n (while we still have large coverage α and small loss β), then we can efficiently represent the matrix $\mathbf{X}$.

Outline

Approximation results for $k = 2$

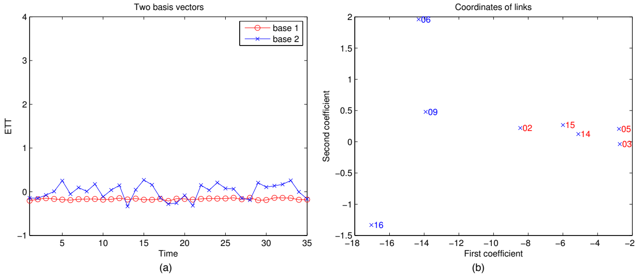


Node 4 as an example ($\alpha = 98\%$, $\beta = 2.5\%$): (a) Original ETT trace (b) After averaging (T=10 min)

(c) Approximated ETT values before regression (d) Approximated ETT values after regression

Basis vectors from one node

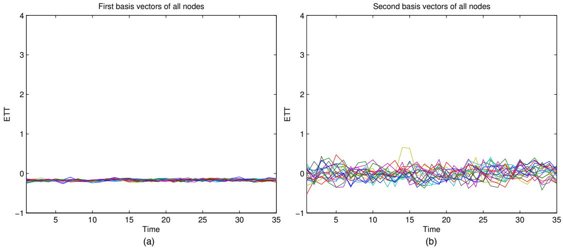
Now, *each link* from this node can be expressed as a linear combination of two *basis vectors* $\mathbf{B}_1$ and $\mathbf{B}_2$.



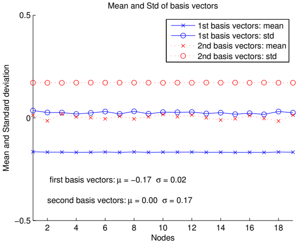
(a) Basis vectors and (b) Final coefficients for Node 4. Links with similar dynamics are highlighted with red color

The first basis vectors are quite stable, while the second basis vectors contain time-varying character.

Basis vectors from all nodes



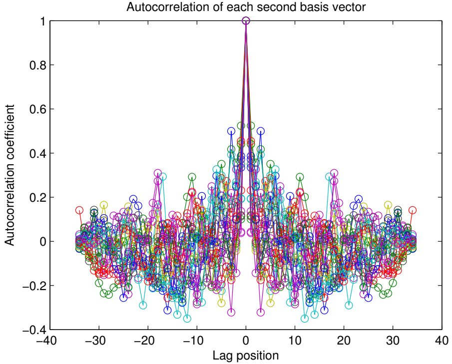
(a) First and (b) second basis vectors from all nodes

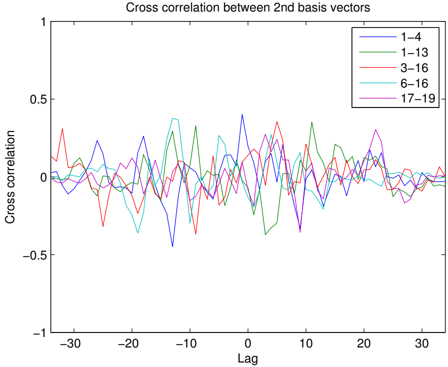


Mean and standard deviation over time

- Most time variations are expressed by second basis vectors.
- First basis vectors have almost zero variance (over time)
- Second basis vectors have almost zero mean (over time)

Autocorreleation and cross-correlation

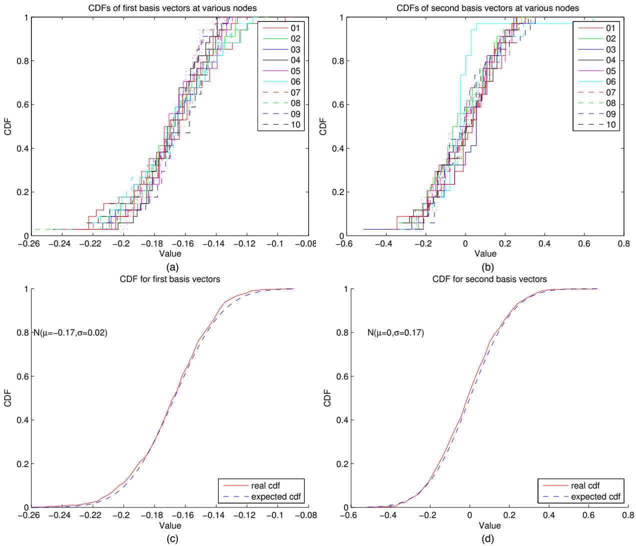




Auto-correlation and cross-correlation between some of second basis vectors

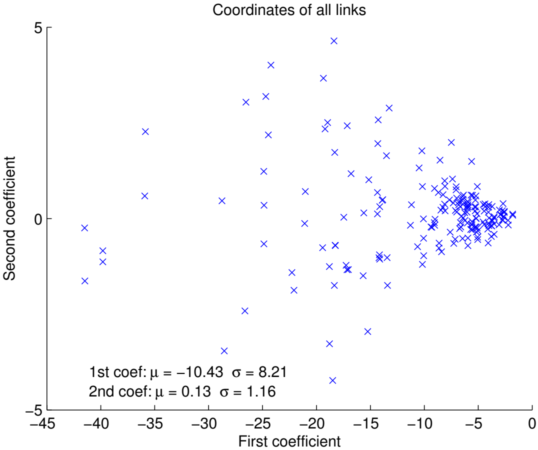
The autocorrelation has small sidelobes and cross-correlation values are small. This supports our assumption about independence.

Distribution of Basis vector components

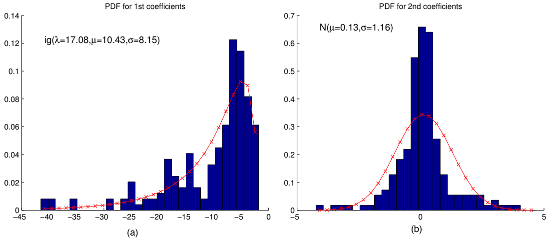


(a) CDFs of first basis vectors from different nodes (b) CDFs of second basis vectors from different nodes
(c) Combined CDF of first basis vectors & fit (d) Combined CDF of second basis vectors & fit

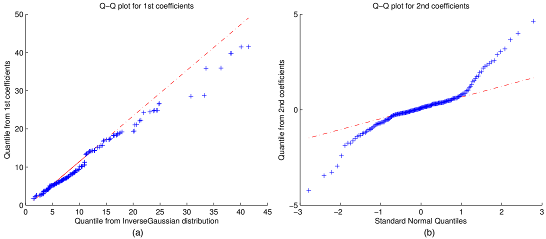
Coefficients of all links



Distributrion of Coefficients



Histogram of (a) all first coefficients (b) all second coefficients

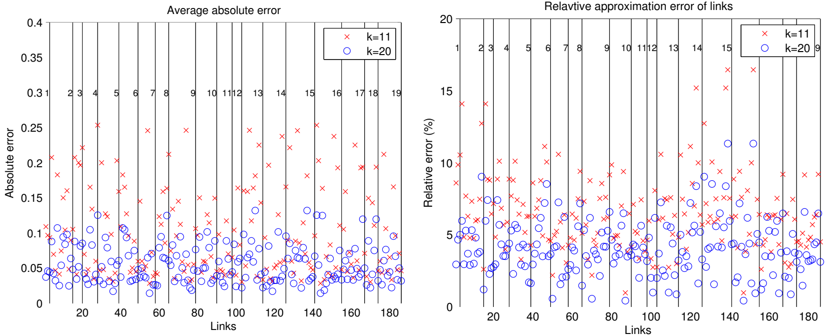


Q–Q plot for (a) first coefficients (b) second coefficients

Outline

Average error of all links

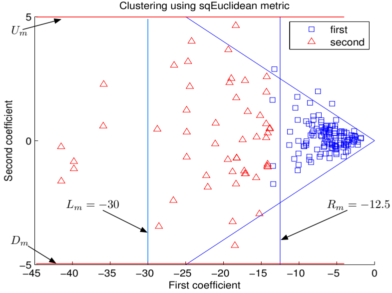
We tried to apply the same PCA method to the entire set of ETT traces for all links (instead of only those associated with one node). However, with the same value of $k = 2$, we get only $\alpha = 93\%$ and $\beta = 1.6\%$



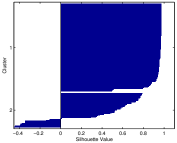
Average absolute and relative approximation error for links (grouped by node)

Outline

Simulation procedure



Clustering coefficients



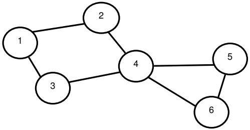
Silhouette diagram

- First cluster contains 70% of pairs and forms a triangular region
- Second cluster contains 30% of pairs, which scatter between boundaries U_m, D_m, L_m, R_m

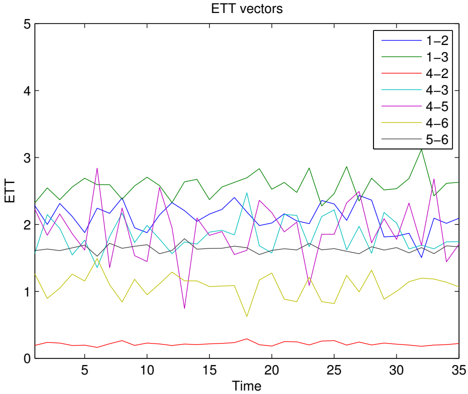
Simulation procedure for each link

- Let $\rho = 0.7$ be the fraction of coefficients from the first group, $s = 0.2$ be the slope of the lines that form the triangular region; $L_m = -30$, $R_m = -12.5$ be the boundary for the first coefficient; $U_m = 5$, $D_m = -5$ be the boundary for the second coefficient.
- Generate a uniformly distributed random number x in [0,1] for each link. If $x < \rho$ then we generate coefficients in the triangular region. Otherwise we generate coefficients in the rectangular region as follows.
- Case 1 – triangular region: Generate f_1 uniformly distributed in $(R_m, 0)$. Calculate the range for f_2 as the segment of the vertical line at f_1 truncated by two lines. Let $f_{2D} = s * f_1$, $f_{2U} = -s * f_1$. Then generate f_2 that is uniformly distributed in (f_{2D}, f_{2U}).
- Case 2 – rectangular region: Generate f_1 that is uniformly distributed in (L_m, R_m) and f_2 that is uniformly distributed in (D_m, U_m).

Simulated ETT values for a simple network



A simple network



Simulated ETT values

Outline

Conclusion

We have shown that:

- PCA is very useful to reduce the size of ETT trace;
- We can efficiently approximate ETT data of all links at any node using only two basis vectors and two coefficients for each link;
- The first basis vector can be considered as a constant and the second as one derived from a normal distribution with a zero mean;
- The marginal distributions of coefficients corresponding to first basis vectors have an inverse Gaussian distribution, while those corresponding to second basis vectors have a nearly Gaussian distribution;
- It is possible to generate the ETT traces for a given network using our observations or with a combination of existing ETT trace data from that network using only a few parameters.

Open Issues and Future Work

- There are several assumptions that have not been tested carefully (e.g., independence)
- Analyzing the trace data without taking the average
- Try different datasets
- Compare with alternative approaches for analyzing and modeling the ETT traces
- Exploit spatial correlation between ETT values

THANK YOU

Q & A