

**Pedro Henriques dos Santos Teixeira**

**Data Stream Anomaly Detection through Principal  
Subspace Tracking**

**Dissertação de Mestrado**

Dissertation presented to the Postgraduate Program in Informatics of the Departamento de Informática do Centro Técnico Científico PUC–Rio as partial fulfillment of the requirements for the degree of Mestre em Informática.

Advisor: Prof. Ruy Luiz Milidiú

Rio de Janeiro  
September 2009



**Pedro Henrique dos Santos Teixeira**

## **Data Stream Anomaly Detection through Principal Subspace Tracking**

Dissertation presented to the Postgraduate Program in Informatics of the Departamento de Informática do Centro Técnico Científico PUC-Rio as partial fulfillment of the requirements for the degree of Mestre em Informática. Approved by the following commission:

**Prof. Ruy Luiz Milidiú**

Advisor

Departamento de Informática — PUC-Rio

**Prof. Marco Antonio Casanova**

Departamento de Informática — PUC-Rio

**Prof. Raúl Pierre Rentería**

Departamento de Informática – PUC-Rio

**Prof. Prof. José Eugenio Leal – PUC-Rio**

Coordinator of the Centro Técnico Científico — PUC-Rio

Rio de Janeiro — September 04, 2009

All rights reserved.

### **Pedro Henriques dos Santos Teixeira**

Graduated from the University of Warwick (England) in Computer Science in 2004, and worked as a senior software developer for the last five years building complex scalable distributed systems. His academic interests are in the area of Optimization and Automatic Reasoning and throughout the program he has participated in the Spock and Netflix challenges, developed an AI video game for the XNA framework and implemented many heuristics for optimizing NP-hard problems such as TOP and TSP. He is currently working on machine learning solutions for monitoring data streams at his own start-up.

#### Bibliographic data

Teixeira, Pedro Henriques dos Santos

Data Stream Anomaly Detection through Principal Subspace Tracking / Pedro Henriques dos Santos Teixeira; advisor: Ruy Luiz Milidiú. — 2009.

95 f. : il. ; 29,7 cm

Dissertação (Mestrado em Informática) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática.

Inclui bibliografia.

1. Informática – Teses. 2. Detecção de Anomalia. 3. Rastreamento do Subespaço Principal. 4. Aprendizado não-supervisionado. 5. Fluxo de dados. 6. Séries temporais. 7. Redução de dimensão. I. Milidiú, Ruy. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

## Acknowledgments

To my parents for their unconditional support and for being role models of moral integrity and kindness throughout the years. I am forever grateful to them for giving me the freedom to pursue my own goals and for having invested in my education.

To Gabriela for the caring love and for having all the patience in the world throughout the past two years.

To my adviser Professor Ruy Milidiú, for sharing his wisdom. The incisive comments were essential to my journey, and I am especially thankful for the early guidance on conciliating my academic work with my professional life.

To the Professors in the comission, Raúl Rentería and Marco Casanova, whose reviews helped improve this dissertation.

To the colleagues at LEARN and to all the teachers and friends at the DI for providing an excellent master's degree.

To my fellow associates who missed me at the company during the last month.

To Globo.com, especially to Marco Lucio and Denis Vieira, who supported our project.

I thank Professor Peter Strobach for replying my e-mails and enlightening me with some history on subspace trackers. I also thank Professor Roland Badeau for sharing his Matlab implementation of many subspace trackers and Professor Mark Coates who answered my inquiries about KOAD's code.

## Abstract

Teixeira, Pedro Henriques dos Santos; Milidiú, Ruy. **Data Stream Anomaly Detection through Principal Subspace Tracking**. Rio de Janeiro, 2009. 95p. MScThesis — Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

The complexity of data centers and the high volume of generated monitoring data poses many challenges to system administrators. Current tools rely on experts to configure fixed thresholds for each data stream, which is not efficient nor appropriated for dynamic systems. We study an unsupervised learning technique based on spectral analysis proposed for anomaly detection. In this technique, just one pass over the data is allowed. It tracks the principal subspace of  $r$  dimensions from  $N$  numerical data streams. An anomaly is considered to be a sudden change in system behaviour and is indicated by a change in the number of latent variables. A previous approach [Papadimitriou et al., 2005] relies on a PAST-type subspace tracker, which is based on an inverse matrix update and is known to be unstable. It requires an extra normalization step in order to guarantee the expected reconstruction error, which really adds to  $\mathcal{O}(Nr^2)$  in time complexity per update instead of the  $\mathcal{O}(Nr)$  advertised. In this work, we present FRAHST, the first rank-adaptive principal subspace tracker based on the new recursive row-Householder algorithm, which is the state-of-the-art for an algorithm of this kind. It is stable, provides orthonormal basis and has a true dominant complexity of only  $\mathcal{O}(Nr)$  flops. Our technique reduces in 75% the number of false positives when compared against the previous subspace technique in the public ABILENE PACKETS dataset while still doubling the number of detections. By embedding lagged values to the input vector to explore temporal correlations, FRAHST successfully detects subtle anomalous patterns in the data and when compared against four other anomaly detection techniques, it is the only one with a consistent  $F_1$  score above 80% in all four datasets. As part of this work, a real-time system is successfully implemented to monitor the infrastructure of a ISP, and is a good use case for unsupervised anomaly detection in the industry.

## Keywords

Anomaly Detection. Principal Subspace Tracking. Unsupervised Learning. Data Streams. Time Series. Dimensionality reduction.

## Resumo

Teixeira, Pedro Henriques dos Santos; Milidiú, Ruy. **Deteção de Anomalia em Fluxo de Dados através de Rastreamento do Subespaço Principal**. Rio de Janeiro, 2009. 95p. Dissertação de Mestrado — Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

A complexidade de um data center combinada com o alto volume de dados gerados para monitoração traz diversos desafios para administradores de sistemas. As ferramentas atuais requerem especialistas que configurem limites constantes para cada fluxo de dado, o que é ineficiente e inapropriado para estes sistemas dinâmicos. Estudamos uma técnica não supervisionada baseada em teoria espectral que foi proposta para deteção de anomalia, na qual é permitida uma única passada sobre os dados e monitora o subespaço principal de dimensão  $r$  a partir de  $N$  fluxos de dados numéricos. Uma anomalia é considerada uma mudança brusca no comportamento do sistema e pode ser indicada por uma mudança no número de variáveis latentes nos dados. A abordagem anterior depende de um algoritmo do tipo PAST, que é baseado em atualizações sobre uma matrix invertida e que é conhecido por ser instável. Além disso, é necessário esforço adicional de ortonormalização para garantir o erro de reconstrução condizente com os parâmetros, o que adiciona  $\mathcal{O}(Nr^2)$  em complexidade de tempo por atualização. Neste trabalho, apresentamos FRAHST, o primeiro algoritmo de rank adaptativo baseado no método de row-Householder recursivo para aproximar o subespaço principal, que representa o estado-da-arte para um algoritmo deste tipo. FRAHST é estável, fornece estimativas orthonormais e tem complexidade dominante de  $\mathcal{O}(Nr)$  flops. Nossa técnica reduz em 75% o número de falsos positivos quando comparado com a técnica anterior no dataset público ABILENE PACKETS e ao mesmo tempo dobra o número de deteções. Ao incorporar valores passados no vector de entrada para explorar correlações temporais, FRAHST deteta sutis padrões anômalos nos dados e quando comparado com outras quatro técnicas para deteção de anomalia, é o único que obtem de forma consistente um valor de  $F_1$  acima de 80% em todos datasets. Como parte deste trabalho, um sistema de tempo-real foi implementado para monitorar a infraestrutura de um provedor de internet e é um caso de sucesso para deteção de anomalia não-supervisionada na indústria.

## Palavras-chave

Deteção de Anomalia. Rastreamento do Subespaço Principal. Aprendizado não-supervisionado. Fluxo de dados. Séries temporais. Redução de dimensão.

# Contents

1	Introduction	<b>11</b>
1.1	Context and Motivation	11
1.2	Objectives	15
1.3	Notation	16
1.4	Contributions	16
1.5	Dissertation structure	19
2	Related work	<b>20</b>
3	Background	<b>25</b>
3.1	Dimensionality reduction	25
3.2	Principal component analysis	25
3.3	Principal subspace tracking	26
4	Fast rank and subspace tracking	<b>36</b>
4.1	Tracking the principal subspace basis	36
4.2	Tracking the principal subspace rank	37
4.3	Exception handling	40
4.4	Achieving lower asymptotic computational complexity	40
4.5	Real-time Anomaly Detection System	46
5	Experiments	<b>51</b>
5.1	Considerations	51
5.2	Describing the datasets	51
5.3	Subspace tracker evaluation	55
5.4	Anomaly detection evaluation	61
5.5	Results for anomaly detection	64
6	Conclusion	<b>78</b>
6.1	Contributions	78
6.2	Future work	79
6.3	Final words	80
A	Householder Reflection	<b>92</b>
B	Annotated Anomalies	<b>95</b>
B.1	Abilene	95

## List of Figures

1.1	Overview of this work.	18
2.1	In this example, one variable is sufficient to maintain the reconstruction error below 4% most of the time. Illustration from Hoke et al. [2006].	21
2.2	Aberrant behavior detection with Holt Winters. Illustration from Brutlag [2000].	24
3.1	PAST algorithm.	30
3.2	PASTd algorithm.	30
3.3	Fast Recursive row-Householder subspace tracking algorithm.	32
4.1	FRAHST algorithm	41
4.2	Asymptotically faster FRAHST algorithm with recurrent QR updates.	44
4.3	Asymptotically faster FRAHST algorithm with recurrent LU updates.	45
4.4	Anomaly detection routine	47
4.5	Raw asynchronous events are collected and normalized messages are sent to the event bus.	48
4.6	We illustrate the processing flow in the system: a query is used to join raw streams into a derived complex event that feeds our algorithm, which is then apt to detect anomalies in variables under surveillance.	49
4.7	Simplified sequence diagram of the update step.	50
5.1	Network layout for the Chlorine dataset.	53
5.2	Topology of the Abilene network when the data as collected.	54
5.3	Deviations from true subspace and from orthonormality on the Artificial dataset.	59
5.4	Deviations from true subspace and from orthonormality on the Artificial dataset.	60
5.5	Wall-clock times for the update step of both versions of the algorithm. The time is linear in respect to the dimensionality of the input data.	61
5.6	Motes dataset: (b) shows the measurements and reconstruction for both algorithms FRAHST and SPIRIT on sensor 32, which is highlighted in (a). Both algorithms adapt to similar low rank values, which vary between 4-6 for most of the duration of the experiment.	62
5.7	Determinant variables and thresholds (dotted lines) for all algorithms during the anomaly detection evaluation on the Abilene Packets dataset.	66
5.8	Determinant variables and thresholds (dotted lines) for all algorithms during the anomaly detection evaluation on the Abilene Flow dataset.	67
5.9	In both figures, bottom panel shows the changes in rank for the FRAHST algorithm.	68



5.10	All four latent variables captured by FRAHST during the experiments on the Abilene Flows and Packets datasets. The main trend is clearly summarised by the first principal direction.	69
5.11	Determinant variables and thresholds (dotted lines) for all algorithms during the anomaly detection evaluation on the enhanced $N = 300$ ISP Routers dataset.	72
5.12	The latent dimension changes in order to explain the anomalous events while maintaining 96% of reconstruction accuracy.	74
5.13	Determinant variables and thresholds (dotted lines) for all algorithms during the anomaly detection evaluation on the ISP Servers dataset.	76
5.14	Top panel displays two anomalies highlighting the anomalous streams (idle and used CPU and Memory), while the bottom panel shows how FRAHST adapts the rank according to anomalous data points in order to maintain 97% of relative reconstruction error.	77
A.1	Householder reflection in two-dimensional space	93

## List of Tables

1.1	Description of notation.	16
3.1	State-of-the-art fast principal subspace trackers.	33
5.1	Description of datasets.	52
5.2	Relative squared reconstruction error for the rank adaptive algorithms. Parameters $[f_E, F_E] = [0.96, 0.98]$ imply that the relative error should be between 0.02 and 0.04.	61
5.3	Relative squared reconstruction error for repeated PCA.	63
5.4	Anomaly detections results for the Abilene Packets dataset.	65
5.5	Results for Abilene flow dataset	66
5.6	Results for the enhanced $N = 300$ ISP Routers dataset.	71
5.7	Results for the enhanced $N = 240$ ISP Servers dataset.	75
B.1	11 anomalies in the Abilene Packet dataset.	95
B.2	15 anomalies in the Abilene Flow dataset.	95