

## STATISTICAL ANALYSIS AND DATA MINING

Research Article

**Sparse partial least squares regression for on - line variable selection with multivariate data streams**

Brian McWilliams, Giovanni Montana ✉

First published: 12 May 2010

<https://doi.org/10.1002/sam.10074>

Cited by: 3



About



Access

als



ire

## Abstract

Data streams arise in several domains. For instance, in computational finance, several statistical applications revolve around the real - time discovery of associations between a very large number of co - evolving data feeds representing asset prices. The problem we tackle in this paper consists of learning a linear regression function from multivariate input and output streaming data in an incremental fashion while also performing dimensionality reduction and variable selection. When input and output streams are high - dimensional and correlated, it is plausible to assume the existence of hidden factors that explain a large proportion of the covariance between them. The methods we propose build on recursive partial least squares (PLS) regression. The hidden factors are dynamically inferred and tracked over time and, within each factor, the most important streams are recursively identified by means of sparse matrix decompositions. Moreover, the recursive regression model is able to adapt to sudden changes in the data generating mechanism and also identifies the number of latent factors. Extensive simulation results illustrate how the methods perform and compare with alternative penalized regression models for streaming data. We also apply the algorithm to solve a multivariate version of the *enhanced index tracking* problem in computational finance. Copyright © 2010 Wiley Periodicals, Inc. Statistical Analysis and Data Mining 3: 170 - 193, 2010

Citing Literature



**Help & Support**



**Opportunities**



**Connect with Wiley**



Copyright © 1999-2018 John Wiley & Sons, Inc. All rights reserved

**WILEY**