System Monitoring with Metric-Correlation Models: Problems and Solutions

Miao Jiang, Mohammad A. Munawar,
Thomas Reidemeister, and Paul A. S. Ward

June 16, 2009

Introduction

- Enterprise software systems are business-critical
- These systems are often large and complex
- Faults do occur and failure is costly
- Challenge: promptly detect failures and diagnose the faulty component(s)

Introduction

One solution proposed in prior work:

- A correctly functioning enterprise-software system exhibits long-term, stable linear correlations between many of its metrics; some of these correlations break when faults occur
- Metric correlation models have been shown to be effective in detecting errors and helping localize faults

Introduction

Linear models

- Linear regression between two metrics is the most efficient modeling technique
  - $y = ax + b$
  - Parameters a and b are typically estimated using Ordinary Least Squares
- Other modeling techniques have been proposed (*e.g.*, to capture nonlinear relationships), but their cost is generally much higher
- A majority of inter-metrics correlations are linear
- In this work, we focus on linear models to capture linear relationships

Motivation

Some factors that limit the effectiveness of linear models even when the underlying relationship is linear:

- Varying coefficients
- Multi-variable correlations
- Non-constant residual variance

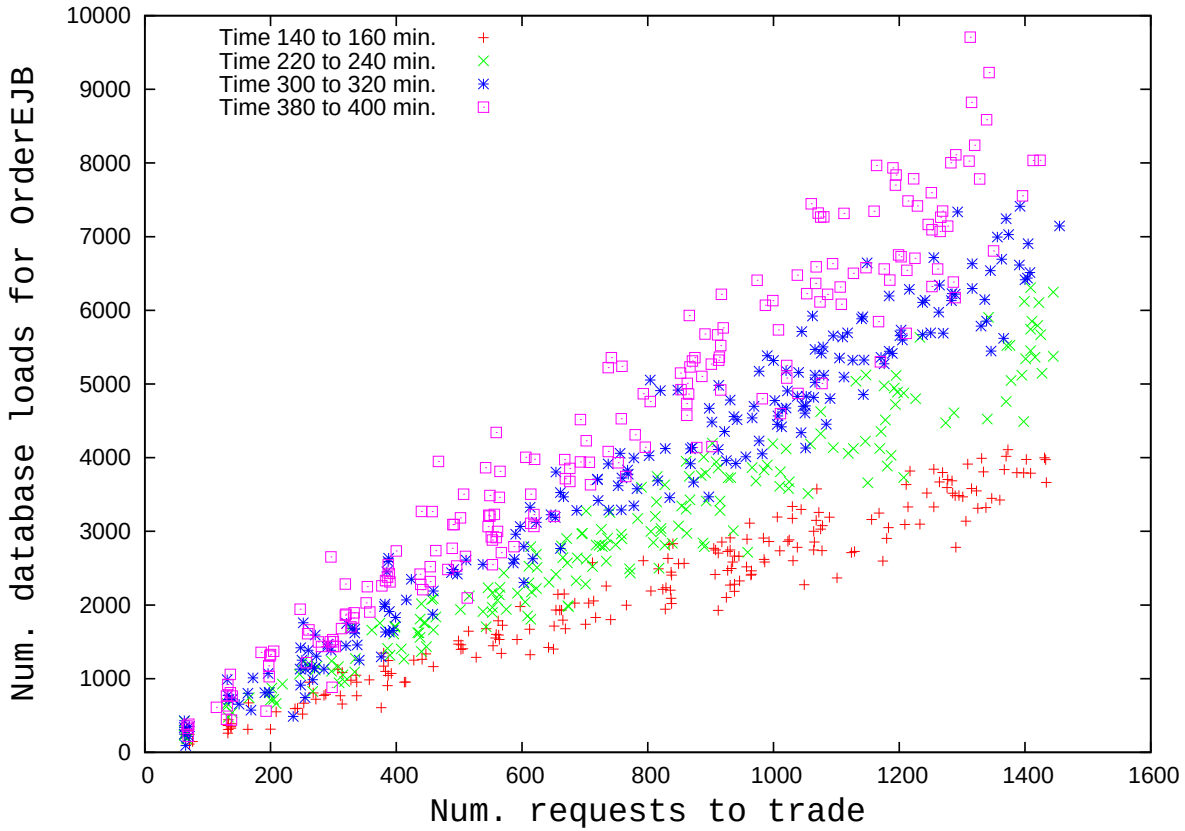
Motivation

Varying coefficients

- The model coefficients may change with circumstances, even though the underlying correlation still exists
  - Often, the coefficients change with time, or some other system-related configurations
- As a result, although the correlations still hold, models with out-of-date parameters may lead to inaccurate assessment of the modelled metrics

Motivation

A case of varying coefficients



Motivation

Multi-variable correlations

- The existence of multi-variable relationships has been observed in prior work
- Such relationships have been much less studied than those involving two variables
- The problem is the high cost of identifying such relationships
  - Searching for all two-variable correlations costs $O(n^2)$ and searching for all k-variable correlations costs $O(n^k)$

Motivation

Missing variables may also cause coefficients to vary. An example involving three metrics:

Metric x: `tradeEJB.AccountProfileEJB:LiveCount`

Metric y: `tradeEJB:MethodResponseTime`

Metric z: `tradeEJB.AccountProfileEJB:PooledCount`

x =	Sample size	Models	F-score
24	617	$z = 24.25 - 4.32y$	237.15
25	90	$z = 25.44 - 4.47y$	39.89
26	312	$z = 26.64 - 5.01y$	226.77
28	981	$z = 27.93 - 3.85y$	276.00

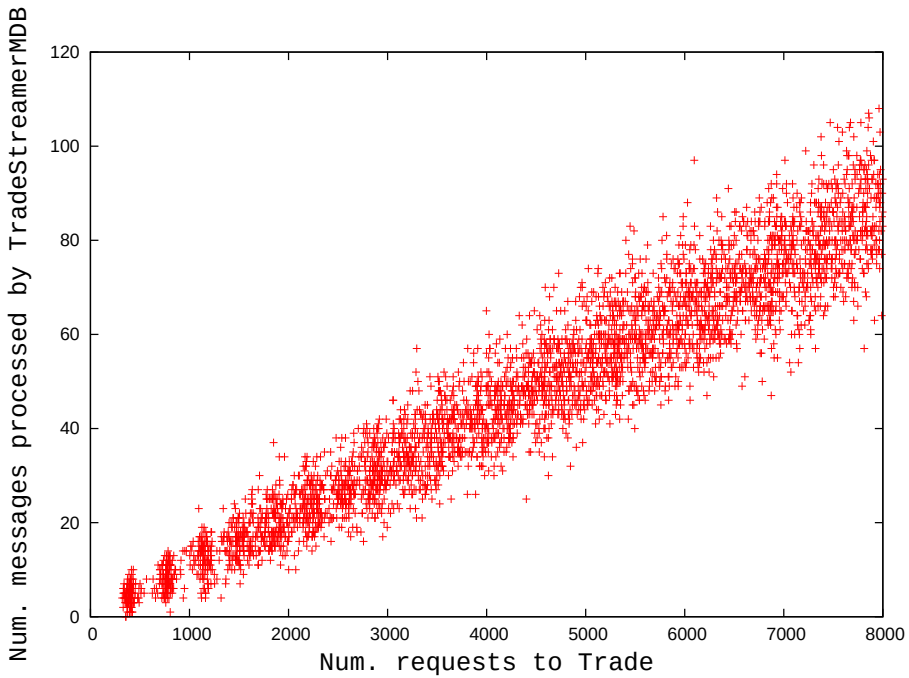
Table: Regression coefficients varying with a third variable

Motivation

Non-constant residual variance (*heteroscedasticity*)

- Heteroscedasticity refers to the fact that the variance of the residuals of a model is not constant
- Heteroscedasticity is very commonly observed in applications of regression models
- A popular example is the relationship between individuals' income and meal expenditure – there is greater variability in what an individual consumes as his/her income increases

Motivation



- Figure shows a clear, strong linear relationship
- With a linear model, residual variance is increasingly larger; at larger values of the independent variable, the model becomes less reliable

Motivation

Why is heteroscedasticity bad?

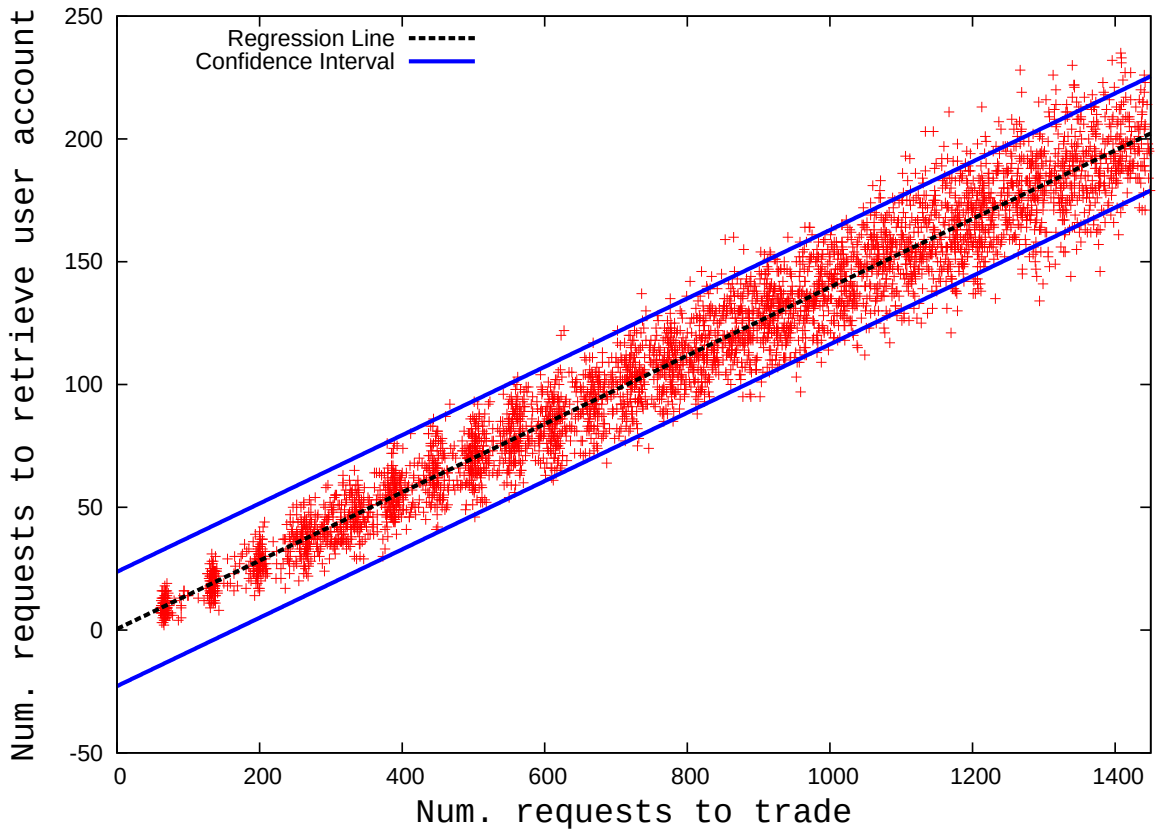
- Violates a key assumption necessary for many regression techniques (*i.e.*, errors have constant variance)
- Heteroscedasticity biases the estimated standard errors, making many diagnostic measures unreliable

Why is heteroscedasticity useful?

- Both varying coefficients and Missing variables may cause non-constant residual variance
- Heteroscedasticity gives a signal for potential varying coefficients or missing variables

Motivation

Confidence intervals for model predictions become invalid



System Modelling

Detecting non-constant error variance: **White test**

- Model the data using ordinary regression and obtain the residuals u.

$$y = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + u$$

- Regress the squared residuals against the independent variables, squared independent variables and their cross-products

$$u^2 = \gamma_0 + \sum_{i=1}^{n} \gamma_i x_i + \sum_{i=1}^{n} \delta_i x_i^2 + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \theta_{ij} x_i x_j + \epsilon$$

  Obtain R^2 of this regression.

- If $nR^2 > \chi^2_{\alpha,k}$, heteroscedasticity is detected

System Modelling

Detecting non-constant error variance: **Goldfeld-Quandt test**

- Order the observations according to the values of X, a variable to which the population error variance may be related.
- Omit c middle observations and divide the rest into the two groups of $(n - c)/2$ observations
- Separately apply regression on the two groups. Then, calculate the sum of residuals squared for the two groups: SSE_1 and SSE_2
- Compute the F-statistic thus:

$$F = \frac{ESS_2}{ESS_1}$$

- If $F > F_{\alpha,d,d}$, heteroscedasticity is detected

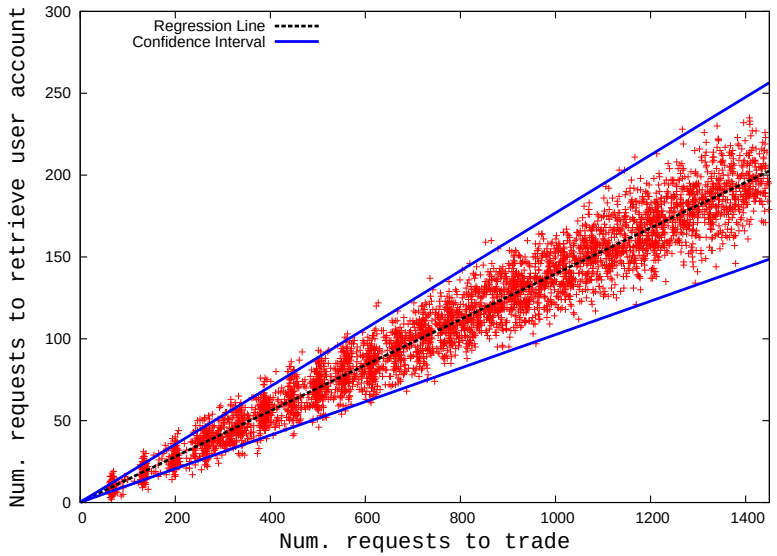
System Modelling

Generalized Least Squares

- Generalized Least Squares can be used to model linear models with heteroscedasticity
- The parameter estimation for model $\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$ is given by: $\hat{\beta} = (\mathbf{X}'\mathbf{C}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}^{-1}\mathbf{Y}$
- If the sample data passes the Goldfeld-Quandt test based on $\mathbf{X_i}$,

$$\mathbf{C} = \begin{pmatrix} x_{i1} & & & & \\ & x_{i2} & & & \\ & & x_{i3} & & \\ & & & \ddots & \\ & & & & x_{in} \end{pmatrix}$$

System Modelling



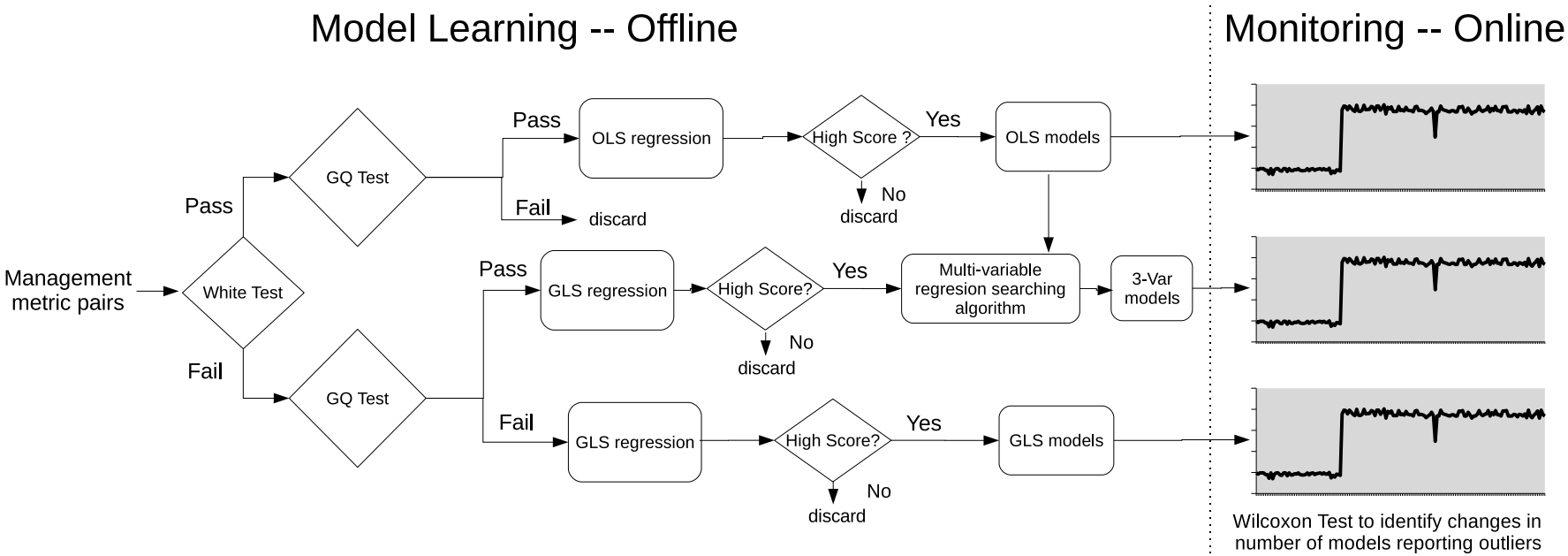
- GLS-based confidence intervals; they better fit the real data
- Fitness score given by:

$$F = \frac{\int_a^b \max(\min(U(x), U'(x)) - \max(L(x), L'(x)), 0)\, dx}{\sqrt{\int_a^b U(x) - L(x)\, dx \int_a^b U'(x) - L'(x)\, dx}}$$

System Modelling

- We develop an algorithm to identify multi-variable models fast
  - Basic idea: when heteroscadacity observed and not explained by the independent variable, start searching for the third variable in the correlation
- The varying coefficients may be addressed in two ways
  - The use of a forgetting factor to gradually remove effects of out-of-date samples
  - Filter the inaccurate models with Wilcoxon-RankSum test

System Monitoring



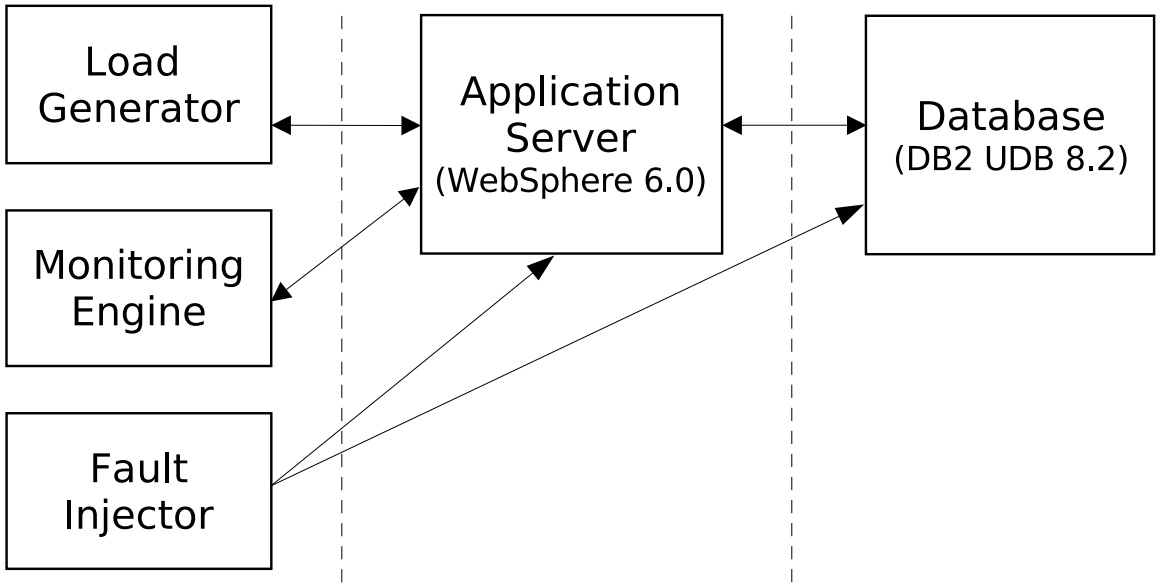
- When faults occur, we expect to see a significant change in the total number of models that report outliers

System Monitoring

- We use Wilcoxon Rank-Sum test to identify such changes
- Two sliding windows are kept during monitoring, and the test tells if there is likely a shift between the two windows
- Larger window size usually leads to less false positives, more false negatives, and longer detection time

Evaluation

Experiment setup



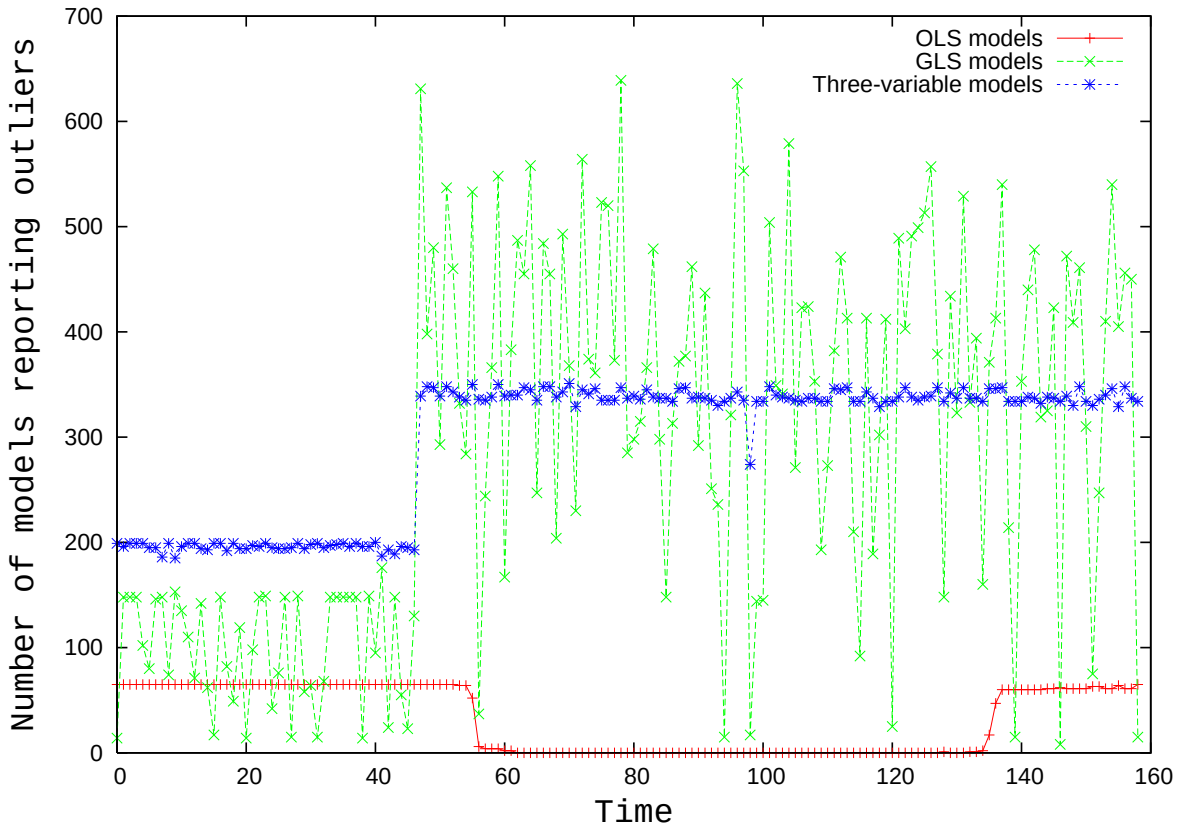
Evaluation

Faults injected

Fault	**C**omponents **I**njected	**D**escription
mis-ds-authentication	JDBC data souce	Misspecified authentication credentials
mis-ds-connection-pool	JDBC data source	Misconfigured size
mis-thread-pool	Web container thread pool	Misconfigured size
del-accountjsp del-displayquotejsp del-tradehomejsp del-marketsummjsp del-orderjsp del-portfoliojsp del-quotejsp	account.jsp displayquote.jsp tradehome.jsp marketsummary.jsp order.jsp portfolio.jsp quote.jsp	Deletion of component

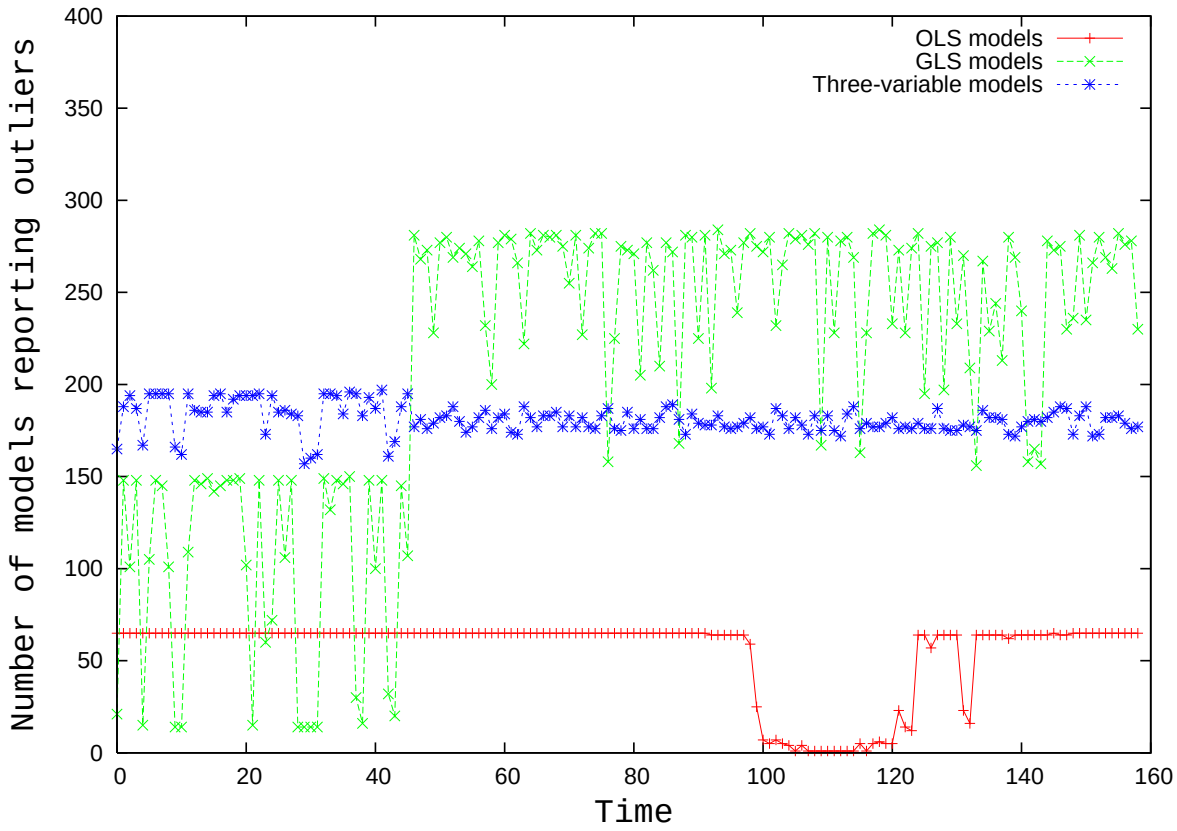
Evaluation

Sample fault detection I



Evaluation

Sample fault detection II



Evaluation

	OLS	GLS	3-var.	Combined
mis-ds-authentication	57	54	**52**	52
mis-ds-connection-pool	61	58	**56**	56
del-accountjsp	64	**52**	-	52
del-displayquotejsp	100	**55**	-	55
del-tradehomejsp	60	**58**	-	58
del-marketsummjsp	**57**	-	152	57
del-orderjsp	**55**	-	105	55
del-portfoliojsp	**57**	61	113	57
del-quotejsp	59	**52**	58	52
mis-thread-pool	71	**55**	58	55

Table: Time (sampling interval) at which anomalies are detected

- Based on window size of 12, which is large enough to eliminate false positives in this evaluation

Summary

- We identify factors that prevent linear models based on OLS from being effective in system monitoring
- We propose methods to deal with these factors, all as efficient as OLS
- We evaluate our overall approach and show that it works well
- In the future, we plan to extend this study to the problem of diagnosis

Questions?

Thank you!