

# Deciding what to observe next: adaptive variable selection for regression in multivariate data streams

<sup>1</sup>Christoforos Anagnostopoulos \*

<sup>2</sup>Niall M. Adams

<sup>1,2</sup>David J. Hand

<sup>1</sup>The Institute for Mathematical Sciences, Imperial College London, 53 Prince's Gate, South Kensington, London SW7 2PG, UK

<sup>2</sup>Department of Mathematics, Imperial College London, South Kensington Campus, London W7 2AZ, UK

## ABSTRACT

Variable selection can be valuable in the analysis of streaming data with costly measurements, as in intensive care monitoring or battery-powered sensor networks. In the presence of drift, selections must be constantly revised, calling for *adaptive* variable selection schemes. An important and novel problem arises from the fact that non-selected variables become missing variables, which induces bias upon subsequent decisions. Here, we consider adaptive variable selection in the context of linear regression, using only a fraction of the available regressors per timepoint. We suggest a scheme that fits a multivariate Gaussian over a sliding window using the *EM* algorithm and selects which variables to observe next using the *Lasso* algorithm. We experiment with simulated and real data to demonstrate that very high prediction accuracy may be retained using as little as 10% of the data.

## Categories and Subject Descriptors

G.3 [Mathematics of Computing]: Probability and Statistics—*correlation and regression analysis, multivariate statistics, robust regression*; I.2.6 [Artificial Intelligence]: Learning; H.4.2 [Information Systems Applications]: Types of Systems—*decision support*

## General Terms

Algorithms, Performance, Experimentation

## Keywords

data streams, variable selection, Lasso, EM algorithm, exploration-exploitation, sensor networks

## 1. INTRODUCTION

\* Author for correspondence: christoforos.anagnostopoulos06@imperial.ac.uk

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SAC'08 March 16-20, 2008, Fortaleza, Ceará, Brazil

Copyright 2008 ACM 978-1-59593-753-7/08/0003 ...\$5.00.

In the analysis of streaming data, existing learning problems, such as parameter estimation and model selection, pose new difficulties; the main challenge being to provide online learning algorithms that keep continuous track of incoming data, revising their models to account for drift. Here, we are interested in the classical problem of regressing a response to a set of predictor variables using only a small number  $q$  out of a total of  $p$  regressors. In the streaming data context, this calls for a scheme that, at any given timepoint, decides which regressors to observe next to minimise prediction error, on the basis of the recent history of the process — observed only partially, according to previous selections. This problem, which we call *Adaptive Variable Selection for Regression* (AVS-R), has strong links to several distinct learning problems and many natural applications; however, to the best of our knowledge it has not been addressed in the literature. We assume for simplicity that the regression is linear and that the number of regressors to be used at each tick is fixed in advance by the user.

Static variable selection for regression (see [8] for a review) is a classical topic, motivated by concerns that too many predictors may lead to loss of interpretability and overfitting, as well as by physical constraints on the total cost of measurements. The relevant literature has been dominated by two strands; one tailored to linear regression (e.g. [15]); the other more akin to Bayesian model selection (e.g. [1],[2]). The *Lasso* [16] is a notable success associated with the former approach: a robust constrained optimisation approach to sparse learning, and the method we adopt here.

Beyond the static case, it is important to distinguish between incremental variable selection schemes, wherein the choice of important predictors is viewed as a label learnt online (e.g. [12]) and schemes such as ours, wherein that choice becomes a decision about what to measure next. In this latter case, a version of the ‘exploration-exploitation’ dilemma of Reinforcement Learning [14] is raised: the immediate reward of observing predictors that are known to be correlated must be balanced against the need to explore the rest of the space; a problem exacerbated by the presence of drift.

A notable artefact is the bias that previous selection steps impose on subsequent ones, by concealing parts of the history of the process: non-selected regressors become missing values. Literature concerned with the bias of inference after variable selection treats variable selection as a one-off decision and accordingly mostly offers large-sample corrections based on asymptotics [19]. Such results are not particularly informative about the effects of cumulative small-

sample bias, which characterises AVS. We hence turn to the problem of *inference with missing data*, classically discussed in [11]. Standard approaches there include *Multiple Imputation* methods and the *Expectation-Maximization (EM)* algorithm [3], which we use here. The performance and theoretical guarantees of these methods is known to depend on the qualities of the missingness mechanism. The case where a user-specified variable selection model provides the missingness mechanism has not been addressed.

AVS-R has natural applications in fields where there is a time-lag between the arrival of the predictors and the response (e.g. gambling or sequential trials in medicine). A notable application lies with the growing field of *Adaptive Query Processing (AQP)* in battery-powered sensor networks [10], where the values of important sensors must be filtered, or predicted if at fault, using measurements from small, chosen sets of correlated sensors. Although the connections of AQP to both model-based variable selection and reinforcement learning have been recognised and the use of Gaussian distributions suggested ([4], [9], [5]), there has been no suggestion of a scheme that simultaneously tracks and selects. In this paper, we suggest precisely such a scheme, based on a windowed application of the *Lasso* and *EM* algorithms. We investigate and discuss the performance of our scheme against a simulated dataset, as well as a real dataset containing temperature readings from 60 weather stations across the UK, with promising results.

## 2. FRAMEWORK

We consider datastreams of the form:

$$(y_1, X_1), \dots, (y_{t-1}, X_{t-1}), (y_t, X_t), \dots$$

where each  $X_t$  is a  $p$ -dimensional vector of regressors and  $y_t$  a univariate response. In AVS-R, our aim is, at each tick of the clock, to select which  $q < p$  regressors we should observe so as to predict  $y_t$  with minimal error. We denote our selection at time  $t$  by  $\gamma_t$ ; the partial observation that results from this selection by  $X_t^{(\gamma_t)}$ ; so that we may form our prediction  $\hat{y}_t$  given  $X_t^{(\gamma_t)}$ . Note that in deciding what to observe at time  $t$ , only the incomplete or *opaque*<sup>1</sup> history of partial observations denoted by  $H_{[1,t-1]}^\gamma$ , is available, as opposed to the complete or *transparent* history  $H_{[1,t-1]}$ :

$$H_{[i,j]}^\gamma = ((y_i, X_i^{\gamma_i}), (y_{i+1}, X_{i+1}^{\gamma_{i+1}}), \dots, (y_{j-1}, X_{j-1}^{\gamma_{j-1}}), (y_j, X_j^{\gamma_j}))$$

$$H_{[i,j]} = ((y_i, X_i), (y_{i+1}, X_{i+1}), \dots, (y_{j-1}, X_{j-1}), (y_j, X_j))$$

In this notation, an *AVS scheme* consists of specifying a *selection* function  $S$  and a *regression* function  $f$ :

$$\text{Selection: } \gamma_t = S(t, H_{[1,t-1]}^\gamma);$$

$$\text{Regression: } \hat{y}_t = f(t, H_{[1,t-1]}^\gamma, X_t^{(\gamma_t)})$$

Note that, naturally, the computational requirements of either function must be asymptotically constant with  $t$ . Moreover, in the current work, we are assuming  $f$  to be linear.

### 2.1 The EM-Lasso scheme for AVS-R.

The range of possible AVS schemes can be narrowed down for the purposes of a first investigation by insisting that  $S$

<sup>1</sup>We follow [17] in using the terms ‘opaque’ and ‘transparent’ to distinguish between the two problems.

and  $f$  depend on the history only through a common summary statistic  $\theta_t$ , which we set to be the mean,  $\mu^{(t)}$ , and covariance,  $\Sigma^{(t)}$ , of the joint distribution of the response and covariates  $(y_t, X_t)$ . Our particular proposal then consists of using, at time  $t$ , the *EM algorithm* to produce estimates  $(\hat{\mu}^{(t)}, \hat{\Sigma}^{(t)})$  on the basis of the  $l$  most recent (partial) observations; then using these as inputs to the *Lasso algorithm* to determine the current selection  $\gamma_t$  and the prediction  $\hat{y}_t$  on the basis of  $X_t^{(\gamma_t)}$ . As the estimation problem is difficult and subject also to numerical instability issues, we additionally regularize the covariance estimates by setting  $\hat{\Sigma}^{(t)} \leftarrow \lambda \hat{\Sigma}^{(t)} + (1 - \lambda)I$  for  $\lambda = 0.001$ .

Note that the size  $l$  of a sliding window is an important quantity in drifting data streams: too large a window may lead to model mis-specification, whereas too small a window invites overfitting [18]. Here, we control this parameter offline, with a view to fitting it to the data in future work.

### 2.2 The Lasso algorithm

The Lasso algorithm [16] is proposed under the assumptions of linear regression with normal errors:

$$y = \beta X + \epsilon, \quad \epsilon_t \sim N(0, \sigma),$$

which hold under our multivariate Gaussian hypothesis. It is unique among other variable selection techniques in that it can be formulated as a convex optimisation problem, hence circumventing the exponential search implicit in classical, stepwise variable selection techniques.

Given a complete training dataset  $(y_i, X_i)_{i=1}^n$  of standardised, zero-mean i.i.d. instances of the problem, the algorithm computes the Lasso-regression coefficients,  $\beta^{(L1)}$ , by minimising the weighted sum of the residual sum of squares and the absolute values of the regression coefficients (i.e., the  $L1$ -norm of the regression vector):

$$\beta^{(L1)} \leftarrow \underset{\beta}{\operatorname{argmin}} \left\{ \left( \sum_{i=1}^n (y_i - \beta X_i)^2 \right) + \lambda \sum_{j=1}^p |\beta_j| \right\} \quad (\star)$$

It is a well-known fact that  $L1$ -norm penalties naturally favor sparse solutions [7], the size of the hyperparameter  $\lambda$  being closely related to the *dimensionality* of the learnt model (i.e. the number of regressors to be included). In fact, in recent work [6] an implementation of the Lasso has emerged which, starting from the full set of regressors (for  $\lambda = 0$ ), produces the full solution path of the Lasso as  $\lambda$  increases, eventually reaching the empty set in an *almost monotonic* fashion — i.e., variables are almost never added after deletion. These deletions ‘punctuate’ the continuous solution path, summarising it in  $p$  regression vectors: one for each dimensionality. There remains of course to choose the dimensionality: here, we assume it user-specified, and intend to address the difficult problem of learning it in future work.

Note that Lasso performs *shrinkage* as well as selection:  $\beta^{(L1)}$  will act only on a subset  $\gamma$  of selected regressors by means of its non-zero entries, but will still differ from the vector obtained by performing ordinary least squares on  $(y_i, X_i^\gamma)_{i=1}^n$ . Note also that  $(\star)$  uses the data only in computing the standardised sample covariance. It was hence possible, with minor amendments, to use estimated correlation matrices as input to an existing implementation [13].

### 2.3 The EM algorithm

The EM algorithm [3] is a procedure for maximum likelihood inference in the presence of missing data. Starting from a user-specified estimate of the ML parameter values,  $\theta(0)$ , it employs an iterative update step, each time choosing  $\theta(i+1)$  to maximise the expected log-likelihood of the observed data, where the expectation is taken over the missing data with respect to the current estimate  $\theta(i)$ . We apply this procedure on the incomplete history  $H_{[t-l, t-1]}^\gamma$ , under the assumption that these are i.i.d. samples from the multivariate Gaussian distribution of  $(y_t, X_t)$ , to obtain an estimate of the mean and covariance  $(\hat{\mu}^{(t)}, \hat{\Sigma}^{(t)})$  of the distribution.

At this point we must note that EM is only guaranteed to converge to a locally optimal solution whenever the missing data are *missing at random*; a condition that holds, roughly speaking, whenever the missingness mechanism cannot use the values the missing variables would have had, had they been observed [11]. This condition is here violated: missingness depends on selections made on the basis of past windows of estimation, which only partially overlap with the current window of estimation, and hence may contain additional information about regressors that are currently missing. Hence, the missingness mechanism at any given time ‘sees more’ of the process than just the observed data in the current window of estimation. We do not discuss this any further here, noting it as the key problem for the future.

A further interesting property of the EM algorithm as applied to covariance estimation is that, at each iteration, it will only update covariance entries,  $\Sigma_{i,j}$ , for which variables  $x_i, x_j$  have been jointly observed at least once in the training dataset. In our scheme, we initialise at the previous window’s estimates causing a possible feedback loop: uncorrelated predictors cease being observed; they are then assumed to remain uncorrelated and are hence not revisited. Our scheme can therefore be expected to be conservatively ‘sticky’: once it locates a good subset, it will be reluctant to move away from it unless its performance deteriorates. This behavior is indeed confirmed in experiments.

### 3. EXPERIMENTATION

We experimented with both simulated and real data. The simulated dataset was created by drawing  $(y_t, X_t)_{t=1}^{900}$  from a multivariate Gaussian  $N(\mu^{(t)}, A_t^t A_t)$ , where the mean  $\mu^{(t)}$  and the upper-triangular Cholesky factor  $A_t$  follow a damped matrix random walk. We kept  $p$  small ( $p = 16$ ) so as to also implement Best Subsets. The real dataset consisted of temperature readings collected by Weather Underground, Inc.<sup>2</sup> from 60 weather stations throughout the UK over a period of 5 days. Each variable (weather station) is recorded in irregular times, ranging roughly from 2 to 15 minutes. Our scheme cannot currently accommodate this, so we used linear interpolation<sup>3</sup> to obtain a  $480 \times 60$  regular dataset.

Performance in this context for a given dataset is naturally measured by prediction error per tick:

$$(\hat{y}_1 - y_1)^2, \dots, (\hat{y}_t - y_t)^2, \dots$$

(Recall that  $y_t$  is the response at time  $t$  and  $\hat{y}_t$  is our estimate of it on the basis of selected observations from the vector of regressors  $X_t$ ). It is however difficult to draw conclusions

<sup>2</sup><http://www.wunderground.com>.

<sup>3</sup>We separate the time index into 15 minute intervals, take the mean of the observations over each interval and record it as a single observation located at its midpoint.

from this long, highly volatile sequence of numbers. To do so, we employ a combination of *benchmarking* and *averaging*.

Assessing relative performance against benchmarks allows us to separate out several distinct sources of error: sample uncertainty, changing stream dynamics, cumulative selection bias, local minima in the EM procedure and so on. Each benchmark scheme is defined by a procedure for estimating  $(\hat{\mu}^{(t)}, \hat{\Sigma}^{(t)})$  and a procedure for using these estimates to select and predict. Here, we distinguish among *opaque estimation*, which uses EM on the incomplete window  $H_{[t-l, t-1]}^\gamma$ ; *transparent estimation*, which forms the maximum likelihood estimator using the complete window  $H_{[t-l, t-1]}$  and *oracle estimation*, which is available only for simulated data and uses the true parameters  $(\mu^{(t)}, \Sigma^{(t)})$ . As for the selection and regression functions, we compare the *Lasso* with:

- *best* (for  $p < 17$ ) — exhaustive search for the subset against which the response has the lowest conditional variance, as computed from  $\hat{\Sigma}^{(t)}$ ;
- *random* — choose at random;
- *full* — use all covariates;
- *mean* — use no covariates, simply predicting  $y_t = \mu_y^{(t)}$ .

In the weather data, we also consider regressing on the  $q$  geographically nearest stations.

We also employ two averaging operations: a *sample average* and a *time average*. In the simulated datasets we could in principle use a sample average to estimate the mean squared error per tick. In practice, this is intractable since we would have to average over all possible past histories (not just windows) at each timepoint, since current error is conditional on past decisions. Thus, we approximate the *conditional mean squared error* instead, averaging at each tick over several instances of the current window of estimation only (also obtaining an estimate of the variance of the error). Moreover, to capture the additional variability of schemes that involve random choice, we average over different outcomes of such choices as well, as part of forming the sample average per tick. The resulting bias/variance sequences are smoother and easier to interpret when plotted.

We may also average over *time* instead to come up with a single number for each scheme. This is particularly useful for the real datasets, where sample averaging is not an option. However, time averaging must still be used with some care, since there is no guarantee of ergodicity in the presence of drift. Finally, when taking both a sample and a time average for simulated data, we perform these in the said order.

### 3.1 Results

In Figure 3, we plot the percentage of the variance explained against the percentage of data used. In real datasets, the former is measured by 1 minus the ratio of the (time averaged) error of each scheme over that of mean imputation. In simulated contexts, as in Table 1, the variance of the response is fixed at 1 so that no such scaling is necessary. The percentage of data used is given by  $\frac{q}{p}$  where  $q$  is the number of regressors used and  $p$  the total number of regressors.

A comparison of the performance of opaque Lasso against that of full regression allows us to make the key observation that high prediction accuracies may be retained with fractionally less data: opaque Lasso can explain more than 97% of the variance in the response with as little as 20% of

the data, featuring a drop of only 1.5% from full regression’s 98.5% (Figure 3). This may be explained by the fact that as more regressors become available, the estimation task becomes more difficult and hence the gain in prediction accuracy decays. Indeed, for large values of  $p$  (but not for small ones), full regression could not gain significant advantages over opaque lasso by adjusting window size, as overfitting gave way to model mis-specification, in either case forcing a poor model fit. In contrast, for smaller values of  $p$  where the estimation task is manageable, full regression outperformed opaque Lasso by as wide a margin as 22% (Table 1).

Opaque Lasso performs much better than random selections: in the real data, it attains the same prediction accuracy with 5% of the data than opaque random selection attains with 50% (Figure 3). In addition, it does so with much lower variance (indicated by the dotted lines in Figure 1). Also, it outperforms geographical selections (Figure 3): a noteworthy result, since in weather temperature data geographical location is very strong domain knowledge. In fact, it is interesting to note that our scheme seems to be extracting different features of weather space in achieving such performance: no significant preference is given to neighboring over non-neighboring stations (Figure 2).

The choice of Lasso as a ‘selector module’ may be assessed by comparing it with the exponentially slow Best Subsets search in simulations. In the oracle context, where the problem of estimation is eliminated, Best Subsets is indeed by far best (Figure 1). In contrast, for both transparent and opaque estimation, Lasso selections perform similarly to best subsets: an exhaustive search of the regression space is not necessarily optimal, as it is also more sensitive to poor estimation. This is indicated by the relatively high variance in Best Subset predictions (Table 1) and emphasizes the fact that the exponentially large search space may not be the dominant obstacle in small-sample variable selection.

We now focus on dynamic behavior. The stickiness effect alluded to earlier is indeed manifested: the transparent Lasso (no selection bias) is far more exploratory than the Opaque one (selection bias), as Figure 2 establishes. Surprisingly, the two schemes perform comparably, although one uses fractionally less information than the other. This is less surprising on second thought: the ‘sticky’ behavior of the EM-Lasso *gives it the chance to base its predictions on a small but highly predictive part of the space that it becomes capable of estimating well*. This is not the case when the selections are changing very fast as in random selections, which explains the much larger difference between opaque and transparent performance there. Finally, although sticky, our scheme remains adaptive: an animated version<sup>4</sup> of Figure 2 showing the selection per tick reveals that our scheme will occasionally enter periods of exploration leading up to it getting ‘stuck’ again, but at a different part of the space.

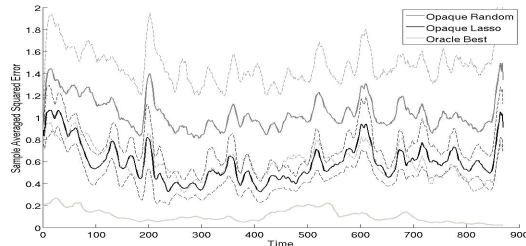
## 4. CONCLUSIONS

In this paper, we introduce the novel problem of adaptive variable selection for regression in multivariate data streams and explain how techniques and notions from static variable selection and inference with missing data may be brought to bear on it. We suggest a scheme that regresses a response variable on a small number,  $q$ , out of a total  $p$  of available re-

<sup>4</sup><http://imaa-dtasouli.imaa.ic.ac.uk/~christoforos/sups.html>

**Table 1: Average Errors in Simulated Dataset.**

Identity of Scheme	Av. error $\pm \sqrt{\text{variance}}$
<b>Opaque Lasso (4/16 regressors)</b>	<b>0.59 <math>\pm</math> 0.14</b>
Opaque Random (4/16 regressors)	1.03 $\pm$ 0.44
Opaque Best (4/16 regressors)	0.60 $\pm$ 0.18
Transparent Full	0.37 $\pm$ 0.18
Oracle Best	0.10 $\pm$ 0.0



**Figure 1: A plot of the average error per tick (solid lines)  $\pm$  respective standard deviations (dotted lines), for the simulated dataset, using 4 out of 16 regressors. Table 1 contains the time averages.**

gressors using the EM and Lasso algorithms and investigate its behavior in experiments with synthetic data and a real dataset of temperatures collected from around the UK over a period of 5 days. Our experiments suggest that high prediction accuracy may be retained using as little as 20% of the data, by focusing in on a small set of important regressors, regularly revising this set to adapt for drift. Several challenges remain unmet, which we intend to address in future work. First, we will be looking for ways to assess and overcome the bias in the EM estimates caused by informatively missing data. Second, as our scheme is liable to underexplore at times, we may be able to ‘tune’ its behavior using ideas from Reinforcement Learning. Moreover, a great challenge will be to drop the assumption of a window in favor of ‘forgetful’ incremental schemes. In a different direction, we would like to explore Bayesian imputation and variable selection, aiming at schemes that scale better with the dimension of the stream. Finally, we aim to study schemes that also optimise the dimensionality,  $q$ , in adaptive ways.

## 5. ACKNOWLEDGMENTS

This work was undertaken as part of the ALADDIN (Autonomous Learning Agents for Decentralised Data and Information Systems) project and is jointly funded by a BAE Systems and the EPSRC (Engineering and Physical Research Council) strategic partnership, under EPSRC grant EP/C548 051/1. The work of David Hand was partially supported by a Royal Society Wolfson Research Merit Award.

We would also like to express appreciation to Weather Underground, Inc. for use of their data and to Yoonseong Kim for aiding in the preprocessing stage.

## 6. REFERENCES

- [1] P. Brown, J., T. Fearn, and M. Vannucci. The choice of variables in multivariate regression: a non-conjugate bayesian decision theory approach. *Biometrika*, 86(3):635–648, 1999.

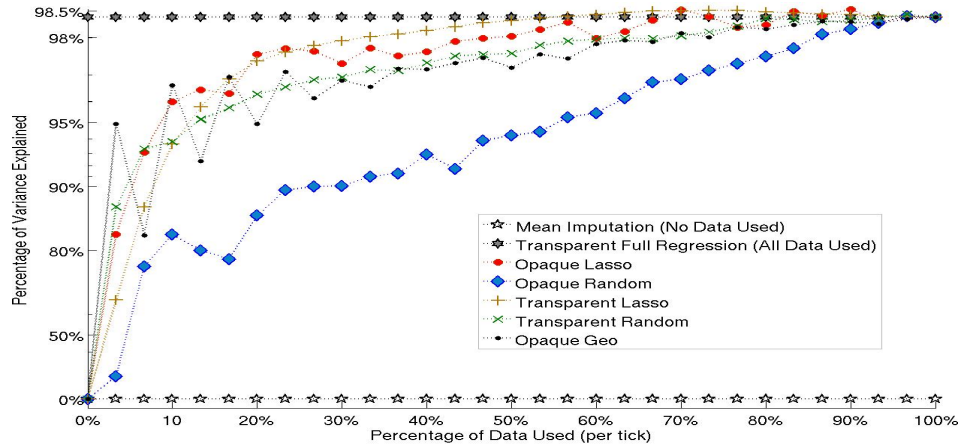


Figure 3: A plot of the percentage of the variance explained, given by  $1 - \frac{\text{error of scheme in question}}{\text{error of mean imputation}}$  where both errors are time-averaged, against the percentage of data used, given by  $\frac{q}{p}$  where  $q$  is the number of regressors used and  $p$  the number of available regressors. Note the logarithmic scale on the  $y$ -axis.

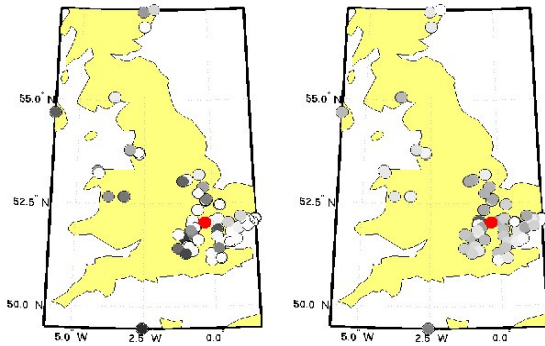


Figure 2: Two plots depicting the preference shown to each weather station by Opaque (left) and Transparent (right) Lasso schemes. The red dot indicates the response, whereas greyscale shadings represent the percentage of the time each regressor was selected and used, black for ‘always’; white for ‘never’.

- [2] P. Dellaportas, J. Forster, and I. Ntzoufras. On Bayesian Model and Variable Selection using MCMC. *Statistics and Computing*, 12(1):27–36, 2002.
- [3] A. Dempster, N. Laird, and D. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.
- [4] A. Deshpande, C. Guestrin, and S. Madden. Using Probabilistic Models for Data Management in Acquisitional Environments. *Proc. of the Biennial Conf. on Innovative Data Sys. Res.(CIDR)*, pages 317–328, 2005.
- [5] A. Deshpande, C. Guestrin, S. Madden, J. Hellerstein, and W. Hong. Model-Driven Data Acquisition in Sensor Networks. *Proc. of the 30th VLDB Conf.*, 2004.
- [6] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least Angle Regression. *Annals of Statistics*, 32(2):407–499, 2004.
- [7] W. Fu. Penalized Regressions: The Bridge versus the Lasso. *Journal of Computational and Graphical Statistics*, 7(3):397–416, 1998.
- [8] E. George. The Variable Selection Problem. *Journal of the American Statistical Association*, 95(452), 2000.
- [9] S. Han, E. Chan, R. Cheng, and K. Lam. A Statistics-Based Sensor Selection Scheme for Continuous Probabilistic Queries in Sensor Networks. *Real-Time Systems*, 35(1):33–58, 2007.
- [10] J. Hellerstein, M. Franklin, S. Chandrasekaran, A. Deshpande, K. Hildrum, S. Madden, V. Raman, and M. Shah. Adaptive Query Processing: Technology in Evolution. *IEEE Data Engineering Bulletin*, 23(2):7–18, 2000.
- [11] R. Little and D. Rubin. *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics, 2002.
- [12] M. Sato. Online model selection based on the variational bayes. *Neural Computation*, 12:1649–1681, 2001.
- [13] K. Sjöstrand. Matlab implementation of LASSO, LARS, the elastic net and SPCA, jun 2005. Ver 2.0.
- [14] R. Sutton and A. Barto. *Introduction to Reinforcement Learning*. MIT Press Cambridge, MA, USA, 1998.
- [15] M. Thompson. Selection of variables in multiple regression: Part i. a review and evaluation. *International Statistical Review/Revue Internationale de Statistique*, 46(1):1–19, 1978.
- [16] R. Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.
- [17] J. Vermorel and M. Mohri. Multi-Armed Bandit Algorithms and Empirical Evaluation. *Proc of the 16th European Conf. on Machine Learning*, pages 437–448.
- [18] H. Wang, J. Yin, J. Pei, P. Yu, and J. Yu. Suppressing Model Overfitting in Mining Concept-Drifting Data Streams. *Proc. of the 12th ACM SIGKDD International Conf. on Knowledge Discovery and Data Mining*, pages 736–741, 2006.
- [19] P. Zhang. Inference after Variable Selection in Linear Regression Models. *Biometrika*, 79(4):741–746, 1992.