

Learning Compact Representations of Time-Varying Processes

Philip Bachman and **Doina Precup**

McGill University, School of Computer Science
 McConnell Engineering Building, 111N
 3480 University St., Montreal, QC, H3A2A7, Canada

Abstract

We seek informative representations of the processes underlying time series data. As a first step, we address problems in which these processes can be approximated by linear models that vary smoothly over time. To facilitate estimation of these linear models, we introduce a method of dimension reduction which significantly reduces error when models are estimated locally for each point in time. This improvement is gained by performing dimension reduction implicitly through the model parameters rather than directly in the observation space.

Methodology

Modeling and predicting the behavior of processes that vary over time is a field rife with potential applications. Our approach to modeling such processes is related to prior work such as projection pursuit regression (Friedman and Stuetzle 1981), sliced inverse regression (Li 1991), locally-weighted regression (Atkeson, Moore, and Schaal 1997), and varying-coefficient models (Hastie and Tibshirani 1993). Additionally, our approach can serve to extend more recent work, with the learning of varying-coefficient/varying-structure models (Kolar, Song, and Xing 2009) and the learning of time-varying graphical models (Song, Kolar, and Xing 2009) being perhaps the most immediate examples.

Formally, our method seeks a set of b basis functions $\{\beta^1, \dots, \beta^b\}$ such that, at each time t , the observed output y_t can be predicted from the observed input x_t as follows:

$$y_t = \sum_{i=1}^b a_t^i (\beta^i \cdot x_t), \quad (1)$$

where we assume y_t is univariate, x_t is n -dimensional, the basis weights a_t^i vary smoothly over time, and the \cdot represents a dot product.

To learn a suitable set of b basis functions for a sequence of m observations drawn from a particular process, we first perform a locally-weighted regression on the sequence so that at each time point $y_t \approx \hat{\beta}_t \cdot x_t$, where:

$$\hat{\beta}_t = \arg \min_{\hat{\beta}} \sum_{t'=1}^m w_\sigma(t', t) \|y_{t'} - (\hat{\beta} \cdot x_{t'})\|^2, \quad (2)$$

where $w_\sigma(t', t)$ is a (Gaussian) kernel weighting function, with mean t and width σ , evaluated at t' . If the observation generating process can be approximated by (1), and the basis weights a_t^i are changing sufficiently smoothly with respect to the kernel width σ , a reasonable set of basis functions should appear as the principal components of the set of pseudo-observations comprising the coefficient vectors $\{\hat{\beta}_1, \dots, \hat{\beta}_m\}$ estimated as described in (2).

Thus, we let our set of learned bases be the first b principal components of the set of pseudo-observations produced by (2). Given the bases $\{\beta^1, \dots, \beta^b\}$ thus selected, we estimate a model for any given time t as $y_t \approx \tilde{\beta}_t \cdot \tilde{x}_t$, where:

$$\tilde{\beta}_t = \arg \min_{\tilde{\beta}} \sum_{t'=1}^m w_\sigma(t', t) \|y_{t'} - (\tilde{\beta} \cdot \tilde{x}_{t'})\|^2, \quad (3)$$

in which $\tilde{x}_{t'}$ denotes the projection of an input $x_{t'}$ onto the set of bases $\{\beta_1, \dots, \beta_b\}$. Thus, when $b < n$, the regression in (3) takes place in a lower dimension than that in (2). As we will show in the results, this reduction in dimension can significantly reduce error in the estimated models.

Results

We briefly present results from two tests. In both tests, the generative process underlying the observations fit the form given in (1), with the processes differing in the degree to which they meet the implicit assumptions of locally-weighted least-squares regression with a Gaussian kernel.

In the first test, each input x_t was drawn independently from a 21-dimensional Gaussian with identity covariance. Each output y_t was produced according to $y_t = \sum_i a_t^i \hat{x}_t^i$, in which \hat{x}_t^i represents the projection of x_t onto the i th true basis. The basis weights a_t^i were generated independently to vary smoothly over time, with a mean of zero and unit variance. We used three true bases, with each basis drawn independently from a 21-dimensional Gaussian with identity covariance. Prior to learning and prediction, we added Gaussian noise to the outputs, with variance equal to 10% of the variance in the pre-noise outputs.

When estimating a set of bases and when producing test predictions, the observation pair (x_t, y_t) was excluded from the locally-weighted regressions in (2) and (3) for time t . Kernel width during basis learning, i.e. (2), was selected

through cross-validation to maximize performance of the learned bases during prediction, i.e. (3).

Figure 1 shows results from our first test. These illustrate the way in which prediction error varied as a function of the number of learned bases and the kernel width used during the regression in (3). Prediction error was measured as the variance of the difference between the predicted output and the pre-noise output, divided by the variance of the pre-noise output. Prediction errors less than 0.1, as produced by our method when using the “correct” number of bases, correspond to predictions more accurate than the noisy observations available in the learning and prediction phases. Figure 1 shows that the optimal kernel width increases with the number of bases used, as one would expect from the effective increase in dimension. Our method performed best when using three learned bases, because these bases closely approximate the subspace spanned by the true bases and the effective dimension reduction permits a smaller kernel width during prediction, which allows closer tracking of changes in the true model.

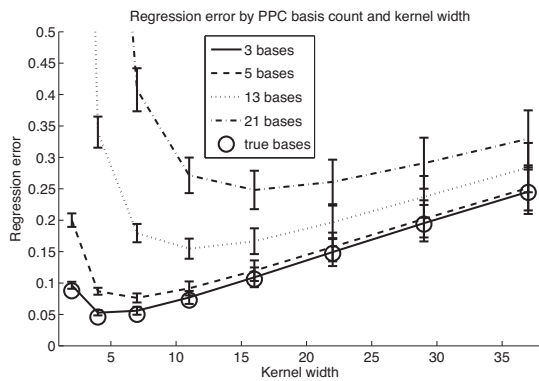


Figure 1: A plot of prediction error as a function of learned basis function count and kernel width when modeling a process defined by a smoothly-varying linear combination of three fixed basis functions. Using the minimal set of learned bases capable of spanning the true process space produces much better estimates of process behavior/state.

Our second test involves a process that less closely matches the assumptions of our method. For this test, the observed inputs were 50-dimensional and 10 true bases were used. Each true basis was first drawn from a 50-dimensional Gaussian with identity covariance, after which its entries were set to zero with probability 0.9, thus producing sparse bases. The basis weights a_t^i were set to vary more abruptly than in the first test, with only a strict subset of the bases having non-zero weights at each point in time, thus leading to sparse effective models.

For this test, during both basis learning and prediction, we used ℓ_1 -regularized regression to better match the sparsity of the underlying process and to mitigate the effects of high-dimensional inputs combined with an abruptly varying process structure that mandated smaller kernel widths. Figure 2 shows results from this test, comparing the prediction error produced by various combinations of basis set size and

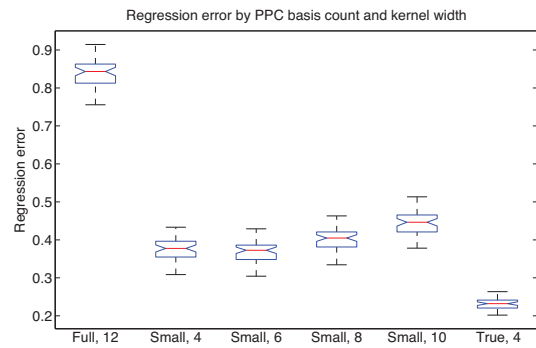


Figure 2: A boxplot of prediction error as a function of approximate basis function count and kernel width when modeling a time-varying process with ten sparse basis functions. The boxes are labeled (x,y), where x is the learned basis set size and y is the prediction kernel width. The “full” set used 50 bases, the small set used 10 bases, and the “true” set used the true bases.

kernel width. As in the first test, selecting a smaller set of basis functions, thus reducing the dimension of the regression in (3) by restricting ourselves to only the most “important” dimensions of the parameter space, permits more accurate model estimation and more precise tracking of variation in the underlying model with a smaller kernel width. The kernel width used with a full set 50 of bases was selected to optimize prediction performance, while the range of widths used with a reduced set of 10 bases was selected to include the optimum.

Our tests show that the specific method we have introduced should prove useful, while the underlying approach to dimension reduction in model parameter space is readily extensible in a way that immediately suggests worthwhile directions for future work.

References

- Atkeson, C. G.; Moore, A. W.; and Schaal, S. 1997. Locally-Weighted Learning. *Artificial Intelligence Review* 11(1-5):11–73.
- Friedman, J., and Stuetzle, W. 1981. Projection Pursuit Regression. *Journal of the American Statistical Association* 76(376):817–823.
- Hastie, T., and Tibshirani, R. 1993. Varying-Coefficient Models. *Journal of the Royal Statistical Society B* 55(4):757–796.
- Kolar, M.; Song, L.; and Xing, E. P. 2009. Sparsistent Learning of Varying-Coefficient Models with Structural Changes. In *Neural Information Processing Systems 23*.
- Li, K.-C. 1991. Sliced Inverse Regression for Dimension Reduction. *Journal of the American Statistical Association* 86(414):316–327.
- Song, L.; Kolar, M.; and Xing, E. P. 2009. Time-Varying Dynamic Bayesian Networks. In *Neural Information Processing Systems 23*.