Input Variable Selection for Neural Networks: Application to Predicting the U.S. Business Cycle

Joachim Utans* **John Moody*** **Steve Rehfuss*** **Hava Siegelmann†**

*Oregon Graduate Institute
Department of Computer Science and Engineering
P.O. Box 91000, Portland, OR 97291-1000

†Technion, Faculty of Industrial Engineering
Department of Information Systems Engineering
Haifa 32000, Israel

Abstract

Selecting a "best subset" of input variables is a critical issue in forecasting. This is especially true when the number of available input series is large, and exhaustive search through all combinations of variables is computationally infeasible. Inclusion of irrelevant variables not only does not help prediction, but can reduce forecast accuracy through added noise or systematic bias. We demonstrate a technique called Sensitivity-Based Pruning (SBP) that removes irrelevant input variables from a nonlinear forecasting or regression model. The technique makes use of a saliency measure computed for each input variable and uses estimates of prediction risk for determining the number of input variables to prune. We present preliminary results of the SBP technique applied to neural network predictors of a key business cycle measure, the U.S. Index of Industrial Production.

1 Introduction: The Business Cycle and the Index of Industrial Production

Of great interest to forecasters of the economy is predicting the "business cycle", or the overall level of economic activity. The business cycle affects society as a whole by its fluctuations in economic quantities such as the unemployment rate (the misery index), corporate profits (which affect stock market prices), the demand for manufactured goods and new housing units, bankruptcy rates, investment in research and development, investment in capital equipment, savings rates, and so on. The business cycle also affects important socio-political factors such as the the general mood of the people and the outcomes of elections.

A scientific model of business cycle dynamics is not yet available due to the complexities of the economic system, the impossibility of doing controlled experiments on the economy, and the non-quantifiable factors such as mass psychology and sociology that influence economic activity. Given the absence of reliable or convincing scientific models of the business cycle, economists have resorted to analyzing and forecasting economic activity by using the empirical "black box" techniques of standard linear time series analysis. We have developed robust predictive models of the business cycle based on neural networks that outperform the standard linear AR models used by most economists.

Economic statistics for the U.S. such as the national income and product accounts and the indices of leading, coincident, and lagging indicators have been collected and computed by the Bureau of Economic Analysis of the Department of Commerce since 1946. The standard measures of economic activity used by economists to track the business cycle are the Gross Domestic Product (GDP)[1] and the Index of Industrial Production (IP).

GDP is a broader measure of economic activity than is IP. However, GDP is computed by the Department of Commerce on only a quarterly basis, while Industrial Production is computed and published monthly. We have focussed on the Index of Industrial Production rather than GDP for three reasons. First, being published monthly, there is more data available for Industrial Production than for GDP. Second, the IP series is more timely than GDP and is therefore watched more closely by business, financial, and economic professionals for making business, trading, or policy decisions. Third, due to its greater oscillation and higher noise level, the IP series is more interesting and challenging from a time series forecasting standpoint than is GDP.

Following prior work by Moody, Levin and Rehfuss (1993) and Levin, Leen and Moody (1994), we develop neural network forecasting models for IP based on monthly observations of IP and other macroeconomic and financial time series.

[1] In 1990, GDP replaced Gross National Product (GNP) as a standard measure of domestic economic activity. GNP includes so-called "factor payments" to and "factor income" from foreign sources that are not included in GDP. These factors relate to interest, dividends, and reinvested earnings by foreign subsidiaries of US companies. As such, they are not really part of the domestic economy. GDP also includes the consumption of fixed capital, an important effect that is not captured by GNP.

0-7803-2145-6/95 $4.00

2 Model Selection

2.1 Nonparametric Modeling with Limited Data

Many data modeling problems in finance, economics, and other fields are characterized by two difficulties: (1) the absence of a complete *a priori* model of the data generation process (such as the models frequently available in physics, say) and (2) by a limited quantity of data. When constructing statistical models for such applications, the issues of model selection and estimation of generalization ability or *prediction risk* are crucial and must be addressed in order to construct a better model.

When a complete *a priori* model for the data generation process does not exist, one must adopt a *nonparametric modeling* approach. In nonparametric modeling, elements of a class of functions known to have good approximation properties, such as smoothing splines (for one or two dimensional problems) or neural networks (for higher dimensional problems), are used to fit the data. An element of this class (eg. a particular neural network) is then chosen which "best fits" the data.

The notion of "best fits" can be precisely defined via an objective criterion; such as *maximum* a posteriori *probability (MAP)*, minimum *Bayesian information criterion (BIC)*, *minimum description length (MDL)*, or *minimum prediction risk (P)*. In this paper, we use the prediction risk as our selection criterion for two reasons. First, it is straightforward to compute, and second, it provides more information than MAP, BIC, or MDL, since it tells us how much confidence to put in predictions produced by our best model.

2.2 Neural Network Architecture Selection

For the discussion of architecture selection in this paper, we focus on the most widely used neural network architecture, the two-layer *perceptron* (or *backpropagation*) network. The response function for such a network with I_λ input variables, H_λ internal (hidden) neurons, and a single output is:

$$\hat{\mu}_\lambda(x) = f(v_0 + \sum_{\alpha=1}^{H_\lambda} v_\alpha \, g(w_{\alpha 0} + \sum_{\beta=1}^{I_\lambda} w_{\alpha\beta} \, x_\beta)) \quad (1)$$

Here, f and g are typically sigmoidal nonlinearities, the $w_{\alpha\beta}$ and $w_{\alpha 0}$ are input weights and thresholds, the v_α and v_0 are the output weights and threshold, and the index λ is an abstract label for the specific two layer perceptron network architecture. While we consider for simplicity this restricted class of perceptron networks in this paper, our approach can be easily generalized to networks with multiple outputs and multiple layers.

For two layer perceptrons, the *architecture selection problem* is to find a good, near-optimal architecture λ for modeling a given data set. The architecture λ is characterized by the number of hidden units H_λ, the subset of input variables I_λ, and the subset of weights v_α and $w_{\alpha\beta}$ that are non-zero. If all of the v_α and $w_{\alpha\beta}$ are non-zero, the network is referred to as *fully connected.*

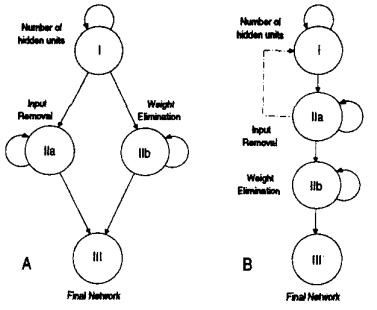

Figure 1: Heuristic Search Strategies: After selecting the number of hidden units H_λ, the input removal and weight elimination can be carried out in parallel (A) or sequentially (B). In (B), the selection of the number of hidden units and removal of inputs may be iterated (dashed line).

Since an exhaustive search over the space of possible architectures is impossible, the procedure for selecting this architecture requires a heuristic search. See Figure 1 for examples of heuristic search strategies and Moody (1994) and Moody and Utans (1994) for additional discussion.

In this paper, we focus on selecting the "best subset" of input variables for predicting the U.S. Index of Industrial Production. In order to avoid an exhaustive search over the exponentially-large space of architectures obtained by considering all possible combinations of inputs, we employ a directed search strategy using the *sensitivity-based input pruning (SBP)* algorithm (see section 3).

2.3 Architecture Selection via the Prediction Risk

The notion of generalization ability can be defined precisely as the *prediction risk* P_λ, the expected performance of an estimator in predicting new observations.

Consider a set of observations $D = \{(\vec{x}_j, t_j); j = 1 \ldots N\}$ that are assumed to be generated as $t_j = \mu(x_j) + \epsilon_j$ where $\mu(x)$ is an unknown function, the inputs x_j are drawn independently with an unknown stationary probability density function $p(x)$, the ϵ_j are independent random variables with zero mean ($\bar{\epsilon} = 0$) and variance σ_ϵ^2, and the t_j are the observed target values. The learning or regression problem is to find an estimate $\hat{\mu}_\lambda(x; D)$ of $\mu(x)$ given the data set D from a class of predictors or models $\mu_\lambda(x)$ indexed by λ. In general, $\lambda \in \Lambda = (S, A, W)$, where $S \subset X$ denotes a chosen subset of the set of available input variables X, A is a selected architecture within a class of model architectures $\mathcal{A}$, and W are the adjustable parameters (weights) of architecture A.

The *prediction risk* $P(\lambda)$ (defined above) can be approximated by the expected performance on a finite test set. $P(\lambda)$ can be defined for a variety of loss functions.

For the special case of squared error, it is:

$$P(\lambda) = \int dx\, p(x)[\mu(x) - \hat{\mu}(x)]^2 + \sigma_\epsilon^2 \quad (2)$$

$$\approx E\{\frac{1}{N}\sum_{j=1}^{N}(t_j^* - \hat{\mu}_\lambda(x_j^*))^2\} \quad (3)$$

where (x_j^*, t_j^*) are new observations that were not used in constructing $\hat{\mu}_\lambda(x)$. In what follows, we shall use $P(\lambda)$ as a measure of the generalization ability of a model. Our strategy is to choose an architecture λ in the model space Λ which minimizes an estimate of the prediction risk $P(\lambda)$.

2.4 Estimation of Prediction Risk

The restriction of limited data makes the model selection and prediction risk estimation problems more difficult. This is the typical situation in economic forecasting, where the time series are short.

A limited training set results in a more severe bias/variance (or underfitting vs overfitting) tradeoff, so the model selection problem is both more challenging and more crucial. In particular, it is easier to overfit a small training set, so care must be taken not to select a model that is too large. Also, limited data sets make prediction risk estimation more difficult if there is not enough data available to hold out a sufficiently large independent test sample. In such situations, one must use alternative approaches which enable the estimation of prediction risk from the training data, such as data resampling and algebraic estimation techniques. Data resampling methods include nonlinear refinements of *v–fold cross–validation (NCV)* and *bootstrap estimation*, while algebraic estimates (in the regression context) include Akaike's *final prediction error (FPE)* (Akaike, 1970), for linear models, and the recently proposed *generalized prediction error (GPE)* for nonlinear models (Moody (1992; 1994)). For comprehensive discussions of prediction risk estimation see Eubank (1988), Hastie and Tibshirani (1990), Wahba (1990), and Moody (1994).

Since it is not possible to exactly calculate the prediction risk P_λ given only a finite sample of data, we have to estimate it. Cross-validation (CV) is a sample re–use method for estimating prediction risk; it makes maximally efficient use of the available data. We have developed a nonlinear refinement refinement of CV called NCV. For a detailed discussion, see Moody and Utans (1994) and Moody (1994).

2.5 NCV: Cross-Validation for Nonlinear Models

Let the data D be divided into v randomly selected disjoint subsets D_j of roughly equal size: $\cup_{j=1}^{v} D_j = D$ and $\forall i \neq j,\ D_i \cap D_j = \emptyset$. Let N_j denote the number of observations in subset D_j. Let $\hat{\mu}_{\lambda(D_j)}(x)$ be an estimator trained on all data except for $(x,t) \in D_j$. Then, the cross-

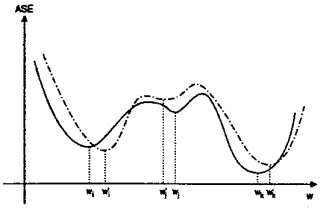

Figure 2: A nonlinear model can have many local minima in the error function. Each local minimum w_i, w_j and w_k (solid curve) corresponds to a different set of parameters and thus to a different model. Training on a different finite sample of data or retraining on a subsample, as in nonlinear cross-validation, gives rise to a slightly different error curve (dashed) and perturbed minima w_i', w_j' and w_k'. Variations due to data sampling in error curves and their minima are termed *model variance*.

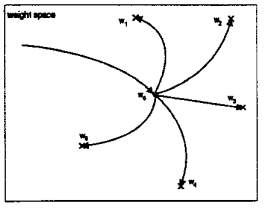

Figure 3: Illustration of the computation of 5–fold *nonlinear cross-validation (NCV)*. First, the network is trained on all data to obtain weights w_0 which are used as starting point for the cross-validation. Each data subset $D_i, i = 1 \ldots 5$ is removed from the training data D in turn. The network is trained, starting at w_0, using the remaining data. This "perturbs" the weights to obtain w_i. The test error of the "perturbed model" w_i is computed on the hold-out sample D_i. The average of these errors is the 5-fold CV estimate of the prediction risk for the model with weights w_0.

validation average squared error for subset j is defined as

$$CV_{D_j}(\lambda) = \frac{1}{N_j}\sum_{(x_k,t_k)\in D_j}\left(t_k - \hat{\mu}_{\lambda(D_j)}(x_k)\right)^2 \quad .(4)$$

These are averaged over j to obtain the v-fold cross-validation estimate of prediction risk:

$$CV(\lambda) = \frac{1}{v}\sum_j CV_{D_j}(\lambda) \quad . \quad (5)$$

Typical choices for v are 5 and 10. Leave–one–out CV is obtained in the limit $v = N$. CV is a nonparametric estimate of the prediction risk that relies only on the available data.

The frequent occurrence of multiple minima in nonlinear models (see Figure 2), each of which represents a different predictor, requires a refinement of the cross-validation procedure. This refinement, *nonlinear cross-validation (NCV)*, is illustrated in Figure 3 for $v = 5$.

A network is trained on the entire data set D to obtain a model $\hat{\mu}_\lambda(x)$ with weights w_0. These weights are used as the starting point for the v-fold cross–validation procedure. Each subset D_j is removed from the training data in turn. The network is re-trained using the remaining data starting at w_0 (rather than using random initial weights). Under

the assumption that deleting a subset from the training data does not lead to a large difference in the locally-optimal weights, the retraining from w_0 "perturbs" the weights to obtain $w_i, i = 1 \ldots v$. The Cross-Validation error computed for the "perturbed models" $\hat{\mu}_{\lambda(D_j)}(x)$ thus estimates the prediction risk for the model with locally-optimal weights w_0 as desired, and not the performance of other predictors at other local minima.

If the network would be trained from random initial weights for each subset, it could converge to a different minimum corresponding to w_i different from the one corresponding to w_0. This would correspond to a different model. Thus, starting from w_0 assures us that the cross-validation estimates the prediction risk for a particular model in question corresponding to $w \approx w_0$.

3 Pruning Inputs via Directed Search and Sensitivity Analysis

Selecting a "best subset" of input variables is a critical part of model selection for forecasting. This is especially true when the number of available input series is large, and exhaustive search through all combinations of variables is computationally infeasible. Inclusion of irrelevant variables not only does not help prediction, but can reduce forecast accuracy through added noise or systematic bias.

In Moody and Utans (1992) and Utans and Moody (1991), we proposed a *sensitivity-based pruning* method for input variables *(SBP)* (see also Moody and Utans (1994) or Moody (1994)). With this algorithm, candidate architectures are constructed by evaluating the effect of removing an input variable from the fully connected network. These are ranked in order of increasing training error. Inputs are then removed following a "Best First" strategy, i.e. selecting the input that, when removed, increases the training error least.

The SBP algorithm computes a *sensitivity measure* S_i to evaluate the change in training error that would result if input x_i were removed from the network. The sensitivity of the network model to variable i is defined as:

$$S_i = \frac{1}{N} \sum_j S_{ij} \tag{6}$$

where S_{ij} is the sensitivity computed for exemplar x_j. Since there are usually many fewer inputs than weights, a direct evaluation of S_i is feasible:

$$S_{ij} = SE(\overline{x}_i, w_\lambda) - SE(x_{ij}, w_\lambda) \tag{7}$$

$$\overline{x}_i = \frac{1}{N} \sum_{j=1}^{N} x_{ij}$$

S_i measures the effect on the training squared error (*SE*) of replacing the i^{th} input x_i by its average $\overline{x}_i$ for all exemplars (replacement of a variable by its average value removes its influence on the network output).

Note that in computing S_i, no retraining is done in evaluating $SE(\overline{x}_i, w_\lambda)$. Also note that it is not sufficient to just set $x_{ij} = 0 \; \forall \; j$, because the value of the bias of each hidden unit was determined during training and would not be offset properly by setting the input arbitrarily to zero. Of course, if the inputs are normalized to have zero mean prior to training, then setting an input variable to zero is equivalent to replacing it by its mean.

4 Empirical Results

Following prior work by Moody *et al.* (1993) and Levin *et al.* (1994), we construct neural network models for predicting the rate of change of the U.S. Index of Industrial Production (IP). The prediction horizon for the IP results presented here is 12 months.

The data set consists of monthly observations of IP and other macroeconomic and financial series for the period from January 1950 to December 1989. The data set thus has a total of 480 exemplars. Input series are derived from around ten raw time series, including IP, the Index of Leading Indicators, the Standard & Poors 500 Index, and so on. Both the "unfiltered" series and various "filtered" versions are considered for inclusion in the model, for a total of 48 possible input variables. The target series and all 48 candidate input series are normalized to zero mean and unit standard deviation.

For the results reported here, networks with three sigmoidal units and a single linear output unit are used (see previous work of Moody *et al.* (1993) and Levin *et al.* (1994)).

Figures 4 and 5 show the results of the sensitivity analysis for the case where the training-set consists of 360 exemplars randomly chosen from the 40 year period, the remaining 120 monthly observations constitute the test-set.

Local optima for the number of inputs are found at 15 on the FPE curve and 13 on the NCV curve. Due to the variability in the FPE and NCV estimates (readily apparent in figure 5 for NCV), we favor choosing the first good local minimum for these curves rather than a slightly better global minimum. This local minimum for NCV corresponds to a global minimum for the test error. Choosing it leads to a reduction of 35 in the number of input series and a reduction in the number of network weights from 151 to 46. Inclusion of additional input variables, while decreasing the training error, does not improve the test-set performance.

5 Summary

We have demonstrated the effectiveness of the *sensitivity-based pruning (SBP)* algorithm for selecting a small subset of input variables from a large number of available inputs. The SBP algorithm as implemented here uses estimates of

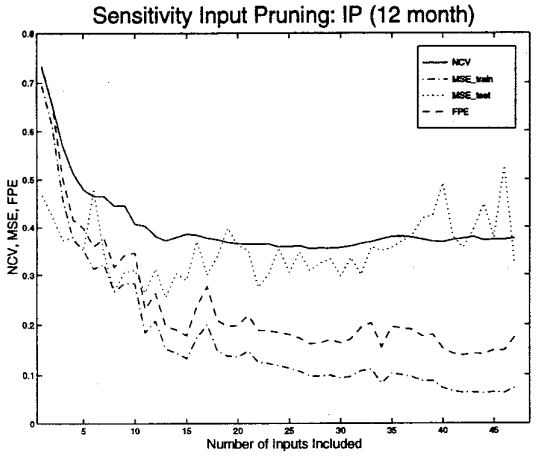

Figure 4: Sensitivity Input Pruning for IP (12 month prediction horizon). The figure shows the NCV, FPE and MSE for both the training and test-set.

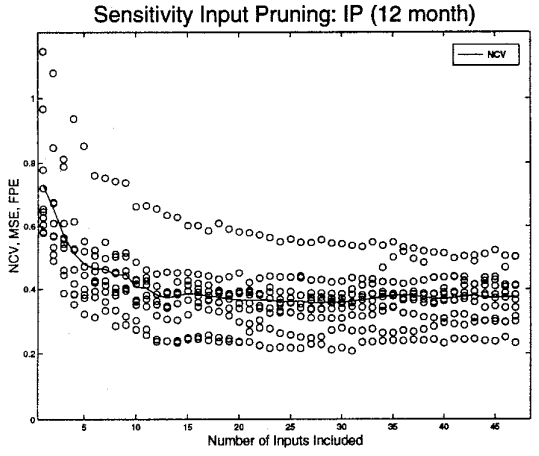

Figure 5: Sensitivity Input Pruning for IP (12 month prediction horizon). The figure illustrates the spread in test-set error for each of the 10 subsets used to calculate NCV (denoted by circles). The NCV error is the average of these test-set errors.

prediction risk $P(\lambda)$, such as our recently proposed *Nonlinear Cross-Validation (NCV)* procedure, to determine the number of inputs to prune from a network model. In the experiments presented here, 35 out of 48 available input time series can be eliminated from a neural network model that predicts the U.S. Index of Industrial Production. The resulting network models exhibit better prediction performances, as measured by either estimates of prediction risk or errors on actual test sets, than models that make use of all 48 input series.

Acknowledgements

We gratefully acknowledge support for our recent work from ARPA and ONR under grants N00014-92-J-4062 and N00014-94-1-0071.

References

Akaike, H. (1970), 'Statistical predictor identification', *Ann. Inst. Statist. Math.* **22**, 203–217.

Eubank, R. L. (1988), *Spline Smoothing and Nonparametric Regression*, Marcel Dekker, Inc.

Hastie, T. J. and Tibshirani, R. J. (1990), *Generalized Additive Models*, Vol. 43 of *Monographs on Statistics and Applied Probability*, Chapman and Hall.

Levin, A. U., Leen, T. K. and Moody, J. E. (1994), Fast pruning using principal components, *in* J. Cowan, G. Tesauro and J. Alspector, eds, 'Advances in Neural Information Processing Systems 6', Morgan Kaufmann Publishers, San Francisco, CA.

Moody, J. (1994), Prediction risk and neural network architecture selection, *in* V. Cherkassky, J. Friedman and H. Wechsler, eds, 'From Statistics to Neural Networks: Theory and Pattern Recognition Applications', Springer-Verlag.

Moody, J. E. (1991), Note on generalization, regularization and architecture selection in nonlinear learning systems, *in* B. H. Juang, S. Y. Kung and C. A. Kamm, eds, 'Neural Networks for Signal Processing', IEEE Signal Processing Society, pp. 1–10.

Moody, J. E. (1992), The effective number of parameters: an analysis of generalization and regularization in nonlinear learning systems, *in* J. E. Moody, S. J. Hanson and R. P. Lippmann, eds, 'Advances in Neural Information Processing Systems 4', Morgan Kaufmann Publishers, San Mateo, CA, pp. 847–854.

Moody, J. E. and Utans, J. (1992), Principled architecture selection for neural networks: Application to corporate bond rating prediction, *in* J. E. Moody, S. J. Hanson and R. P. Lippmann, eds, 'Advances in Neural Information Processing Systems 4', Morgan Kaufmann Publishers, San Mateo, CA, pp. 683–690.

Moody, J. and Utans, J. (1994), Architecture selection strategies for neural networks: Application to corporate bond rating prediction, *in* A. N. Refenes, ed., 'Neural Networks in the Captial Markets', John Wiley & Sons.

Moody, J., Levin, A. and Rehfuss, S. (1993), 'Predicting the U.S. index of industrial production', *Neural Network World* **3**(6), 791–794. special issue: *Proceedings of Parallel Applications in Statistics and Economics '93*, M. Novak (ed).

Utans, J. and Moody, J. (1991), Selecting neural network architectures via the prediction risk: Application to corporate bond rating prediction, *in* 'Proceedings of the First International Conference on Artificial Intelligence Applications on Wall Street', IEEE Computer Society Press, Los Alamitos, CA.

Wahba, G. (1990), *Spline Models for Observational Data*, Vol. 59 of *Regional Conference Series in Applied Mathematics*, SIAM Press, Philadelphia.