SUPANOVA - A Sparse, Transparent Modelling Approach

Steve R. Gunn[†] and Martin Brown[‡]

[†]Image, Speech and Intelligent Systems Group
Department of Electronics and Computer Science
University of Southampton
[‡]Unilever Research Port Sunlight
Quarry Road East
Wirral, U.K.

23 April 1999

Abstract

Traditional neural networks produce opaque models that are difficult to interpret. This work describes a transparent, non-linear, modelling approach that enables constructed models to be visualised, enhancing their validation and interpretation. The technique combines the representational advantage of a sparse ANOVA decomposition, with the good generalisation ability of a Support Vector Machine.

1 Introduction

The problem of empirical data modelling is germane to many applications. In empirical data modelling a process of induction is used to build up a model of the system from examples. Ultimately the quantity and quality of the observations govern the performance of the model. By its observational nature, data obtained is finite and sampled; typically this sampling is non-uniform and due to the high dimensional nature of the problem, the data will form only a sparse distribution in the input space. Consequently, the problem is nearly always ill-posed. To address the ill-posed nature of the problem it is necessary to convert the problem to one that is well-posed. For a problem to be well-posed, a unique solution must exist that varies continuously with the data. Conversion to a well-posed problem is typically achieved with some form of capacity control, which aims to balance the fitting of the data with constraints on the model flexibility, producing a robust model that generalises successfully. Previous approaches to restoring the well posedness have included regularisation methods. In this paper, the method chosen is based around support vector methods introduced by Vapnik et al. (1997); Vapnik (1995). Girosi (1997) and Smola et al. (1998a) has shown that these methods can be placed in a regularisation framework, which guarantees the well posedness. Support Vector Machines (SVMs) employ the method of Structural Risk Minimisation to enable good generalisation for small sample sizes. However, the solution of SVMs is a summation of Kernels, and consequently the solution is opaque.

Another goal of modelling, which is often overlooked, is to produce a transparent solution. This is beneficial in that it enables the model to be validated and interpreted. Features that aid transparency are input selection and ways of decomposing the model into smaller more interpretable pieces that can be easily visualised. To address this issue we focus on the integration of an ANOVA representation to provide a transparent approach to modelling. The ANalysis Of VAriance (ANOVA) representation is motivated by the decomposition of a function into additive components, with the goal of representing the function by a subset of the terms from this expansion. A function may be decomposed into

$$f(\mathbf{x}) = f_0 + \sum_{i=1}^{n} f_i\left(x_i\right) + \sum_{i=1}^{n}\sum_{j=i+1}^{n} f_{i,j}\left(x_i, x_j\right) + \cdots + f_{1,2,\ldots,n}\left(\mathbf{x}\right), \tag{1}$$

where n is the number of inputs, f_0 represents the bias and the other terms represent the univariate, bivariate, etc., components. These basis functions are semi-local and are similar to the approaches used by Friedman (1991) in the Multivariate Adaptive Regression Splines (MARS) technique and in the Adaptive Spline Modelling of Observational Data (ASMOD) technique (Gunn et al., 1997; Kavli and Weyer, 1995). The additive representation is advantageous when the higher order terms can be ignored, so that the resulting model is represented by a small subset of the ANOVA terms, which may be easily visualised. This produces a transparent model, in contrast to the majority of neural network models, providing the modeller with structural knowledge that can be used for both validation and model interpretation. Due to the curse of dimensionality (Bellman, 1961), an exhaustive search of the possible model structures is demanding; for a fixed number of basis functions, the model space is 2^n. Accordingly, an evolutionary strategy has previously been employed to search the model space, using an approach based upon the ASMOD algorithm, (Gunn et al., 1997; Kavli and Weyer, 1995). The aim of this work is to produce transparent models that generalise well, using a global approach to the modelling problem. This paper introduces a new SUpport vector Parsimonious ANOVA (SUPANOVA) technique that fulfils this goal.

2 ANOVA spline functions in Reproducing Kernel Hilbert Spaces

The following theory is based upon Reproducing Kernel Hilbert Spaces (*RKHS*) (Wahba, 1990; Girosi, 1997). An attractive kernel for deployment within an SVM is an infinite spline. The first order infinite spline kernel, which passes through the origin, is defined on the interval $[0, 1)$ by,

$$K_{spline}(u, v) = \int_0^1 (u - \tau)_+ (v - \tau)_+ \, d\tau. \tag{2}$$

and has the form of a piece-wise cubic polynomial,

$$K_{spline}(u, v) = uv + \frac{uv}{2} \min(u, v) - \frac{1}{6} (\min(u, v))^3 . \tag{3}$$

Multidimensional kernels can be obtained by forming tensor products of univariate kernels. A multivariate ANOVA spline kernel is given by the tensor product of univariate spline kernels plus a bias term,

$$K_{ANOVA}(u, v) = \prod_{d=1}^{n} \{1 + K_{spline} ({}_du, {}_dv)\} , \tag{4}$$

where left-hand subscript notation denotes the input value for an input d. To illustrate how this kernel comprises an ANOVA decomposition, consider its expansion for a three dimensional input vector,

$$\begin{aligned} K(u, v) &= \prod_{i=1}^{3} \{1 + g({}_iu, {}_iv)\} \\ &= 1 + g({}_1u, {}_1v) + g({}_2u, {}_2v) + g({}_3u, {}_3v) \\ &\quad + g({}_1u, {}_1v)g({}_2u, {}_2v) + g({}_1u, {}_1v)g({}_3u, {}_3v) \\ &\quad + g({}_2u, {}_2v)g({}_3u, {}_3v) + g({}_1u, {}_1v)g({}_2u, {}_2v)g({}_3u, {}_3v) \end{aligned} \tag{5}$$

It is evident that the tensor product produces the ANOVA terms of Equation 1, producing a flexible model. The formulation of Equation 2 constrains the univariate spline to pass through the origin, ensuring a unique expansion in the ANOVA terms is obtained. It can be shown that Equation 3 produces a positive definite kernel function. Consequently, it can be proven that each of the additive terms in Equation 5 is also positive definite since the product of two positive definite functions is also positive definite. This enables partial forms of Equation 5 to be used as valid kernels for SVMs.

3 SUPANOVA

The SUPANOVA technique is designed to select a parsimonious model representation by selecting a small set of terms from the complete ANOVA representation of Equation 1. It achieves this by decomposing the non-linear modelling problem into three stages, Figure 1. The motivation for doing this is to address the curse of dimensionality (Bellman, 1961). The first stage is used to select a complete ANOVA basis, from which the second stage selects a sparse subset that maintains good accuracy. Finally, a new model is constructed using this sparse representation. This technique contrasts with other parsimonious techniques, such as MARS and ASMOD, in that it aims to find a full model and sub-select the significant terms. MARS and ASMOD aim to construct a model from an empty, or simple, model by searching for significant terms and adding them in to the model, the so-called forward selection backward elimination strategy. The drawback with these approaches is that they are local and can suffer from entrapment in local minima within the construction process. Additionally, they may not be strictly well-posed. Other strategies for sparse selection have been proposed in the literature, notably Wahba (1990) using the basis function energy.

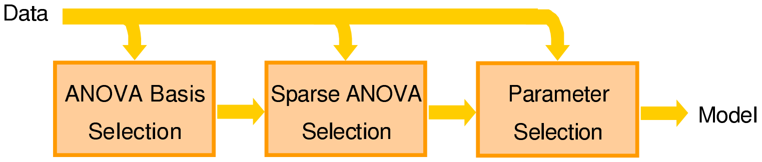


FIGURE 1: SUPANOVA Technique

3.1 Stage I : Basis Selection

The initial goal is to obtain a complete ANOVA basis that is a suitable representation for the model. This is a difficult modelling task, which is addressed by using the kernel-based method of SVMs to address the curse of dimensionality. Splines are an attractive choice for modelling (Wahba, 1990) due to their ability to approximate arbitrary functions. Accordingly, an SVM with an ANOVA spline kernel, Equation 4, is used to approximate the data. This produces a complete ANOVA basis from which the 2^n sub-models can be extracted. The actual model will be composed of a product of piecewise cubic splines with a finite number of knots located at a subset of the data points. The flexibility of the model goes against traditional thinking in statistical modelling, but the over-fitting problem is resisted by the inclusion of the capacity control of the SVM. The capacity control parameter C can be determined using 8-fold cross validation, combined with an automatic procedure, which searches for a local minimum of the validation error.

Another attractive feature of the SVM approach to regression is that it will accommodate a variety of loss functions (Smola et al., 1998a,b). In this paper, we employ two, an $\epsilon-$insensitive and a quadratic loss function. The quadratic loss function gives a solution which is identical to a Gaussian process, and as such this technique may also be employed within a Gaussian process framework. The optimal regression function is given by the minimum of the functional,

$$\Gamma(\mathbf{w}, \boldsymbol{\xi}^*, \boldsymbol{\xi}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\left(\sum_{i=1}^{l}\xi_i + \sum_{i=1}^{l}\xi_i^*\right), \tag{6}$$

where C is the smoothing parameter determined by cross-validation, and $\boldsymbol{\xi}$, $\boldsymbol{\xi}^*$ are slack variables representing upper and lower errors on the outputs of the model. The resultant optimisation problem for a quadratic loss function is given by,

$$\bar{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\beta_i\beta_j K(x_i, x_j) - \sum_{i=1}^{l}\beta_i y_i + \frac{1}{2C}\sum_{i=1}^{l}\beta_i^2, \tag{7}$$

and for an $\epsilon-$insensitive loss function is given by,

$$\bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{\alpha}}^* = \arg\min_{\boldsymbol{\alpha},\boldsymbol{\alpha}^*} \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)K(x_i, x_j) - \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)y_i + \sum_{i=1}^{l}(\alpha_i + \alpha_i^*)\epsilon, \tag{8}$$

with constraints,

$$0 \leq \alpha_i, \alpha_i^* \leq C, \quad i = 1, \ldots, l \tag{9}$$

The ANOVA spline kernel has an implicit bias, and hence the equality constraint is not required for either of the two loss functions (Girosi, 1997). Letting $\bar{\beta}_i = (\bar{\alpha}_i - \bar{\alpha}_i^*)$, the model resulting from the solution of either (7) or (8) is given by,

$$m(z) = \sum_{\text{SVs}} \bar{\beta}_i K(x_i, z). \tag{10}$$

In the quadratic case, there will be no sparseness in the support vector expansion and hence all data points will be support vectors. This technique produces a complete ANOVA expansion from which the 2^n components of Equation 1 can be extracted, by expanding the kernel in the manner of Equation 5. However, some terms in the ANOVA expansion will typically be negligible and the goal is to extract the significant terms, to reflect the underlying model structure. The second stage uses a technique to select the most significant terms from the ANOVA expansion to provide a parsimonious representation.

3.2 Stage II : Sparse Basis Selection

The complete ANOVA representation is difficult to interpret. By introducing a technique that extracts a subset of terms, it is possible to provide a model, which is interpretable whilst retaining good approximating capability. The individual ANOVA terms can be extracted from the SVM solution to the first stage by the appropriate sub-kernel. For example, the univariate terms are given by,

$$m_j(z) = \sum_{SVs} \beta_i K_{spline}({}_j x_i, {}_j z), \tag{11}$$

and the bi-variate terms are given by,

$$m_{j,k}(z) = \sum_{SVs} \beta_i K_{spline}({}_j x_i, {}_j z) K_{spline}({}_k x_i, {}_k z), \tag{12}$$

where β_i are the Lagrange multipliers obtained from the ANOVA kernel solution. In this stage we borrow a technique from the wavelet community (Chen, 1995), where it is desirable to select a sparse basis from an overdetermined one. A commonly employed method for determining a sparse representation is to trade off the error in approximation with the sparseness of the representation,

$$\min \text{Error} + \lambda \, \text{Sparsity}, \tag{13}$$

where λ controls this trade-off. To enforce continuity the technique retains the loss function that was employed in the basis selection stage. The optimisation problem is given by,

$$\min_{\mathbf{c}} L\left(\mathbf{y}, \mathbf{\Phi c}\right) + \lambda \, S(\mathbf{c}) \quad \text{subject to} \quad c_i \geq 0, \, \lambda \geq 0 \tag{14}$$

where L is the loss function, $\mathbf{\Phi}$ is the ANOVA basis obtained from the first stage, S is the sparseness measure and $\mathbf{c}$ is the coefficient vector. The goal in selecting a sparse representation is to minimise the number of non-zero coefficients, c_i. Here non-negativity of the sub-kernel multipliers, c_i, is enforced since the requirement is to simply scale the basis functions (not to invert them). A practical measure of sparseness is a p-norm,

$$S(\mathbf{c}) = \|\mathbf{c}\|_p. \tag{15}$$

As p increases the solution becomes less sparse and the computational complexity of the resulting optimisation problem is relaxed. Ideally a value of $p = 0$, which measures the number of terms in the expansion is attractive. This case is referred to as atomic decomposition (Chen, 1995) and results in a hard combinatorial optimisation problem. Alternatively choosing a value of $p = 2$ produces a straightforward optimisation problem. This case is referred to as the method of frames or ridge regression (Chen, 1995), but crucially the sparseness within the basis is now lost. A good compromise occurs when $p = 1$ producing a sparse solution, with a practical implementation. This norm has been employed in basis-pursuit de-noising (Chen, 1995) to achieve a very similar goal. It is also advantageous in that it produces a Quadratic Program (QP) which can be solved using the same optimiser that is used to optimise the SVM. In the case of an $\epsilon-$insensitive loss function, the problem simplifies to a Linear Program (LP). Finally, the sparseness parameter, λ, can be selected by choosing a set of values, and selecting the one which gives the nearest loss to the error obtained on the validation sets in the first stage. The solution for a quadratic loss function is given by,

$$\min_{\mathbf{c}} \|\mathbf{y} - \mathbf{\Phi c}\|_2^2 + \lambda \|\mathbf{c}\|_1 \quad \text{subject to} \quad c_i \geq 0, \tag{16}$$

$$\min_{\mathbf{c}} \mathbf{c}^T \mathbf{\Phi}^T \mathbf{\Phi c} + \left(\lambda \mathbf{1} - 2\mathbf{y}^T \mathbf{\Phi}\right) \mathbf{c} \quad \text{subject to} \quad c_i \geq 0, \tag{17}$$

and for an $\epsilon-$insensitive loss function by,

$$\min_{\mathbf{c}} \|\mathbf{y} - \mathbf{\Phi c}\|_{1,\epsilon} + \lambda \|\mathbf{c}\|_1 \quad \text{subject to} \quad c_i \geq 0, \tag{18}$$

$$\min_{\mathbf{c}} \begin{pmatrix} \lambda \mathbf{1} \\ \mathbf{1} \\ \mathbf{1} \end{pmatrix}^T \begin{pmatrix} \mathbf{c} \\ \mathbf{m} \\ \mathbf{n} \end{pmatrix} \quad \text{subject to} \quad \begin{pmatrix} \mathbf{\Phi} & \mathbf{I} & -\mathbf{I} \\ -\mathbf{\Phi} & -\mathbf{I} & \mathbf{I} \\ -\mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & -\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & -\mathbf{I} \end{pmatrix} \begin{pmatrix} \lambda \mathbf{c} \\ \mathbf{m} \\ \mathbf{n} \end{pmatrix} \leq \begin{pmatrix} \epsilon + \mathbf{y} \\ \epsilon - \mathbf{y} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}. \tag{19}$$

3.3 Stage III : Parameter Selection

The first two stages together can be considered as a sparse kernel selection method. In this final stage the sparse kernel representation of stage II is deployed within an SVM to produce a sparse ANOVA model.

4 MPG Data Modelling

The performance of the SUPANOVA approach is demonstrated by application to the problem of modelling automobile miles per gallon data (Blake and Merz, 1998). The MPG data set contains the miles travelled, per gallon of fuel consumed, for various different cars. The input variables measure six characteristics of a car; the number of cylinders (discrete), displacement, horsepower, weight, acceleration and model year (discrete). The goal is to discover a relationship between the MPG and the cars' characteristics. After removing a small number of entries with missing values from the original data set, the experiments were performed using 392 examples, 352 for training and 40 for estimating the generalisation performance. The SUPANOVA algorithm was executed 50 times using quadratic and $\epsilon-$insensitive loss functions, with the whole data set "randomly" partitioned into the training and test sets.

Figure 4 illustrates one of the 100 models, obtained from the SUPANOVA technique. It can be seen that it has selected 8 interaction terms (bias, 3 univariate, 3 bivariate and one trivariate) from a possible 64 terms. Table 1 demonstrates that the difference in the mean of the estimated generalisation error between a full ANOVA model and a parsimonious ANOVA model is negligible. However, it also demonstrates that the parsimonious kernel has a lower variance and hence suggest it is more robust. These results were corroborated by the quadratic loss function results. Comparing the two different loss functions shows that, for this particular data-set, there is very little performance difference. Inspection of the ANOVA terms selected by the 100 models shows a high consistency, and confirms the robustness of the technique.

Loss Function		**Estimated Generalisation Error**					
Training	Testing	Stage I		Stage III		Linear Model	
		Mean	Variance	Mean	Variance	Mean	Variance
Quadratic	Quadratic	6.97	7.39	7.08	6.19	11.4	11.0
ϵ Insensitive	ϵ Insensitive	0.48	0.04	0.49	0.03	1.80	0.11
ϵ Insensitive	Quadratic	7.07	6.52	7.13	6.04	11.72	10.94

Table 1: SUPANOVA Results for MPG Data Set. ($\epsilon = 2.5$)

Future work will investigate the nature of the regularisation term within the SVM regression framework. The regularisation is effectively zeroth order and this may be inappropriate in some circumstances. The dimension of the optimisation problem for the first and last stages is dependent upon the data size. However, the second stage is computed in the feature space and hence the dimension of the optimisation problem is exponentially related to the input dimension. This curse of dimensionality can be addressed by considering a restricted ANOVA expansion, e.g. consider only univariates and bivariates. This will reduce the complexity of the second stage to polynomial in the input dimension. Finally, the technique can just as easily be applied to the SVM classification scenario.

Acknowledgements

This work was funded by the UK Engineering and Science Research Council via grant GR/K55110.

References

R.E. Bellman. *Adaptive Control Processes*. Princeton University Press, Princeton, NJ, 1961.

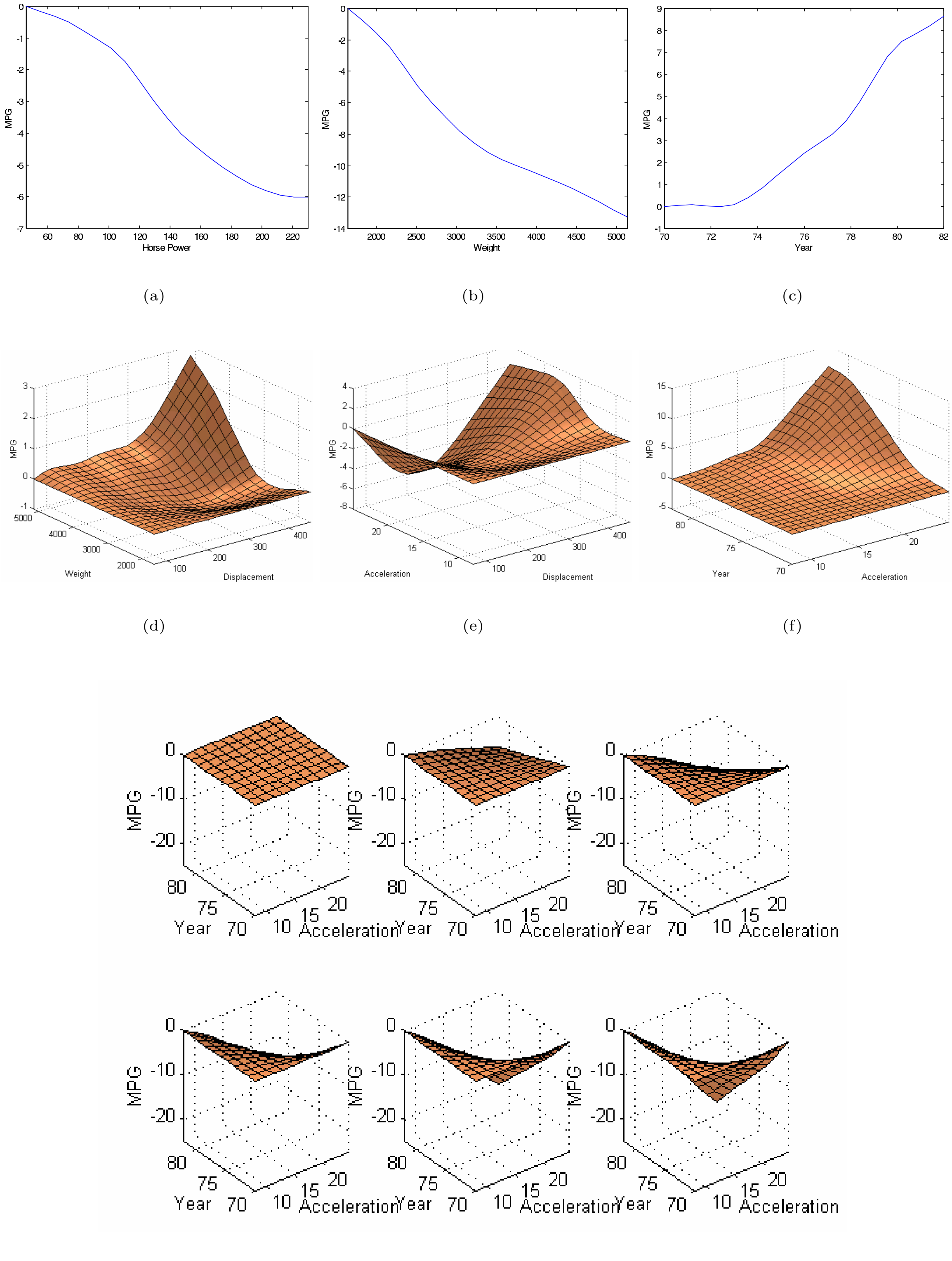
C.L. Blake and C.J. Merz. UCI repository of machine learning databases, 1998.

S. Chen. *Basis Pursuit*. PhD thesis, Department of Statistics, Stanford University, November 1995.

J.H. Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, 19:1–141, 1991.

F. Girosi. An equivalence between sparse approximation and Support Vector Machines. A.I. Memo 1606, MIT Artificial Intelligence Laboratory, 1997.

S.R. Gunn, M. Brown, and K.M. Bossley. Network performance assessment for neurofuzzy data modelling. In X. Liu, P. Cohen, and M. Berthold, editors, *Intelligent Data Analysis*, volume 1208 of *Lecture Notes in Computer Science*, pages 313–323, 1997.



(a)

(b)

(c)

(d)

(e)

(f)

(g)

FIGURE 2: ϵ-Insensitive AMPG model (1 of 50).

T. Kavli and E. Weyer. On ASMOD - an algorithm for building multivariable spline models. In G.R. Irwin K.J. Hunt and K. Warwick, editors, *Advances in Neural Networks for Control Systems*, Springer series on Advances in Industrial Control, pages 83–104. Springer Verlag, 1995.

A. J. Smola, B. Schölkopf, and K.-R. Müller. The connection between regularization operators and support vector kernels. *Neural Networks*, 11:637–649, 1998a.

A. J. Smola, B. Schölkopf, and K.-R. Müller. Convex cost functions for support vector regression. In M. Bodén L. Niklasson and T. Ziemke, editors, *Proc. of the 8th ICANN, Perspectives in Neural Computing*, pages 99–104, Berlin, Germany, 1998b. Springer Verlag.

V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, N.Y., 1995. ISBN 0-387-94559-8.

V. Vapnik, S. Golowich, and A. Smola. Support vector method for function approximation, regression estimation, and signal processing. In M. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems 9*, pages 281–287, Cambridge, MA, 1997. MIT Press.

G. Wahba. *Spline Models for Observational Data*. Series in Applied Mathematics, Vol. 59, SIAM, Philadelphia, 1990.