Heteroscedastic Regularised Kernel Regression for Prediction of Episodes of Poor Air Quality

Robert J. Foxall*, Gavin C. Cawley*, Nicola L. C. Talbot,
Stephen R. Dorling† and Danilo P. Mandic*

*School of Information Systems
University of East Anglia
Norwich, U.K. NR4 7TJ
{rjf,gcc,mandic}@sys.uea.ac.uk

†School of Environmental Sciences
University of East Anglia
Norwich, U.K. NR4 7TJ
s.dorling@uea.ac.uk

Abstract.
A regularised kernel regression model is introduced for data characterised by a heteroscedastic (input dependent variance) Gaussian noise process. The proposed model provides more robust estimates of the conditional mean than standard models and also confidence intervals (error bars) on predictions. The benefits of the proposed model are demonstrated for the task of non-linear prediction of episodes of poor air quality in urban environments.

It is well known that a sum-of-squares error (SSE) metric corresponds to maximum likelihood estimation for regression tasks where the targets are assumed to have been corrupted by additive Gaussian noise with constant variance (i.e. a *homoscedastic* noise process) (e.g. [1]). The Least-Squares Support Vector Machine [2], kernel ridge-regression [3, 4] and Regularisation Machines [5] form a family of closely related techniques that perform non-linear regression using a linear model, constructed in a fixed feature space induced by a Mercer kernel, minimising a regularised sum-of-squares criterion. In this paper, we extend this family to include a formulation that is optimal for Gaussian noise with input-dependent variance (i.e. a *heteroscedastic* noise process). Linear models are constructed in a kernel induced feature space, estimating both the conditional mean and variance of the target distribution, using a regularised maximum likelihood criterion [1, 6]). This results in both more robust estimates of the conditional mean [2] and also a confidence interval on predictions (i.e. error bars). In this study, we apply the proposed method for prediction of episodes of poor air quality, in terms of both an estimate of the concentration of a given pollutant and an estimate of the probability that the concentration exceeds a given statutory threshold level.

1 Method

Suppose we are given a dataset $\mathcal{D} = \{\boldsymbol{x}_i, y_i\}_{i=1}^{\ell},\ \boldsymbol{x}_i \in \mathcal{X} \subset \mathbb{R}^d,\ y_i \in \mathbb{R}$, where the targets, y_i, are assumed to be corrupted by an independent and identically distributed (i.i.d.) sample drawn from a Gaussian noise process with a mean of zero and input dependent variance, $y_i = \mu(\boldsymbol{x}_i) + \epsilon_i,\ \epsilon_i \sim \mathcal{N}(0, \sigma(\boldsymbol{x}_i))$. The probability of observing target y_i, given input vector $\boldsymbol{x}_i$ is given by

$$p(y_i|\boldsymbol{x}_i) = \frac{1}{\sigma(\boldsymbol{x}_i)\sqrt{2\pi}} \exp\left\{-\frac{[\mu(\boldsymbol{x}_i) - y_i]^2}{2\sigma^2(\boldsymbol{x}_i)}\right\}. \tag{1}$$

The negative log-likelihood of $\mathcal{D}$ can then be written (omitting constant terms) as

$$-\log \mathcal{L}_{\mathcal{D}} = \sum_{i=1}^{\ell} \left\{\log \sigma(\boldsymbol{x}_i) + \frac{[\mu(\boldsymbol{x}_i) - y_i]^2}{2\sigma^2(\boldsymbol{x}_i)}\right\}. \tag{2}$$

To model the data, we must estimate the functions $\mu(\boldsymbol{x})$ and $\sigma(\boldsymbol{x})$. The conditional mean is estimated by a linear model, $\mu(\boldsymbol{x}) = \boldsymbol{w}^{\mu} \cdot \boldsymbol{\phi}^{\mu}(\boldsymbol{x}) + b^{\mu}$, constructed in a fixed feature space, $\mathcal{F}^{\mu}$ ($\boldsymbol{\phi}^{\mu} : \mathcal{X} \to \mathcal{F}^{\mu}$). $\mathcal{F}$ is induced by a positive definite "Mercer" kernel, $\mathcal{K}^{\mu} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, defining the inner product $\mathcal{K}^{\mu}(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\phi}^{\mu}(\boldsymbol{x}) \cdot \boldsymbol{\phi}^{\mu}(\boldsymbol{x}')$. The superscript μ is used to denote entities used to model the conditional mean $\mu(\boldsymbol{x})$. The standard deviation is a strictly positive quantity and so the *logarithm* of the standard deviation is estimated by a second linear model, $\log \sigma(\boldsymbol{x}_i) = \boldsymbol{w}^{\sigma} \cdot \boldsymbol{\phi}^{\sigma}(\boldsymbol{x}) + b^{\sigma}$, similarly constructed in a feature space $\mathcal{F}^{\sigma}$ defined by Mercer kernel $\mathcal{K}^{\sigma}$. A superscript σ is used to identify entities used to model the standard deviation, $\sigma(\boldsymbol{x})$. The parameters of the model ($\boldsymbol{w}^{\mu}, b^{\mu}, \boldsymbol{w}^{\sigma}$ and b^{σ}) are determined by minimising the objective function

$$L = \frac{1}{2}\gamma^{\mu}\|\boldsymbol{w}^{\mu}\|^2 + \frac{1}{2}\gamma^{\sigma}\|\boldsymbol{w}^{\sigma}\|^2 + \sum_{i=1}^{\ell} \left\{\log \sigma(\boldsymbol{x}_i) + \frac{[\mu(\boldsymbol{x}_i) - y_i]^2}{2\sigma^2(\boldsymbol{x}_i)}\right\}. \tag{3}$$

Clearly this corresponds to quadratic regularisation [7] of a maximum likelihood cost function, where γ^{μ} and γ^{σ} are regularisation parameters, providing independent control of the bias-variance trade-off [8] for the models of the conditional mean and standard deviation. The representer theorem [9] suggests that the optimal values of $\boldsymbol{w}^{\mu}$ and $\boldsymbol{w}^{\sigma}$ can be written as expansions over training patterns, such that

$$\mu(\boldsymbol{x}) = \sum_{i=1}^{\ell} \alpha_i^{\mu} \mathcal{K}^{\mu}(\boldsymbol{x}, \boldsymbol{x}_i) + b^{\mu} \quad \text{and} \quad \log \sigma(\boldsymbol{x}) = \sum_{i=1}^{\ell} \alpha_i^{\sigma} \mathcal{K}^{\sigma}(\boldsymbol{x}, \boldsymbol{x}_i) + b^{\sigma}. \tag{4}$$

2 An Efficient Training Algorithm

The parameters, $(\boldsymbol{\alpha}^{\mu}, b^{\mu}, \boldsymbol{\alpha}^{\sigma}, b^{\sigma})$, of the conditional mean and standard deviation models can be found via an iterative re-weighted least squares (IRLS) procedure [10], alternating updates of the mean and standard deviation models.

2.1 Updating the Model of the Conditional Mean

If $\sigma(\boldsymbol{x}_i),\ \forall i \in \{1, 2, \ldots, \ell\}$ are held constant, the optimal parameters of the model of the conditional mean, $(\boldsymbol{\alpha}^\mu, b^\mu)$, are given by the minimiser of the objective function

$$L^\mu = \frac{1}{2}\gamma^\mu \|\boldsymbol{w}^\mu\|^2 + \sum_{i=1}^{\ell} \zeta_i \{\mu(\boldsymbol{x}_i) - y_i\}^2, \tag{5}$$

where $\zeta_i^{-1} = 2\sigma^2(\boldsymbol{x}_i)$. This is equivalent to the objective function to be minimised in the weighted least-squares support vector machine [2], and so is minimised by the solution of the set of linear equations

$$\begin{bmatrix} \boldsymbol{\Omega} & \mathbf{1} \\ \mathbf{1}^T & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}^\mu \\ b^\mu \end{bmatrix} = \begin{bmatrix} \boldsymbol{y} \\ 0 \end{bmatrix}, \tag{6}$$

where $\boldsymbol{\Omega} = (\boldsymbol{K}^\mu + \boldsymbol{D})$, $\boldsymbol{K}^\mu = \{k_{ij}^\mu = \mathcal{K}^\mu(\boldsymbol{x}_i, \boldsymbol{x}_j)\}_{i,j=1}^{\ell}$, $\mathbf{1} = (1, 1, \ldots, 1)^T$, $\boldsymbol{y} = (y_i, y_2, \ldots, y_\ell)^T$, $\boldsymbol{\alpha}^\mu = (\alpha_i^\mu, \alpha_2^\mu, \ldots, \alpha_\ell^\mu)^T$ and $\boldsymbol{D}$ is a diagonal matrix with elements $\gamma^\mu/(\zeta_1, \zeta_2, \ldots, \zeta_\ell)$.

2.2 Updating the Conditional Standard Deviation Model

If $\mu(\boldsymbol{x}_i),\ \forall i \in \{1, 2, \ldots, \ell\}$ are held constant, the optimal parameters of the model of the conditional standard deviation, $(\boldsymbol{\alpha}^\sigma, b^\sigma)$, are given by the minimiser of the objective function

$$L^\sigma = \frac{1}{2}\gamma^\sigma \|\boldsymbol{w}^\sigma\|^2 + \sum_{i=1}^{\ell} \left[z_i + \xi_i \exp\{-2z_i\}\right], \tag{7}$$

where $\xi_i = \frac{1}{2}[\mu(\boldsymbol{x}_i) - y_i]^2$ and $z_i = \boldsymbol{w}^\sigma \cdot \boldsymbol{\phi}(\boldsymbol{x}_i) + b^\sigma = \sum_{j=1}^{\ell} \alpha_j^\sigma \mathcal{K}^\sigma(\boldsymbol{x}_i, \boldsymbol{x}_j) + b^\sigma$. It is straightforward to obtain the gradient vector, $\boldsymbol{\nabla}$, and Hessian matrix, $\boldsymbol{H}$ with respect to the vector of model parameters $(\boldsymbol{\alpha}^\sigma, b^\sigma)$. The model of the conditional standard deviation can then be updated via a simple Newton-Raphson algorithm, i.e.

$$(\boldsymbol{\alpha}^\sigma, b^\sigma)_{t+1} = (\boldsymbol{\alpha}^\sigma, b^\sigma)_t - \boldsymbol{H}^{-1}\boldsymbol{\nabla}. \tag{8}$$

2.3 Convergence and Stability

The objective functions, L, L^μ and L^σ, can all be shown to constitute convex optimisation problems (i.e. their respective Hessian matrices are positive semi-definite), and therefore posess single, global minima. Furthermore, decreases in the values of L^μ or L^σ during alternating steps of the training algorithm produce corresponding reductions in L. Both μ and σ steps are guaranteed to reduce L^μ and L^σ respectively, unless the corresponding minima have already been found. The stability of the training algorithm and convergence to the global minimum of L are therefore assured.

3 Results

3.1 The Motorcycle Benchmark

The Motorcycle benchmark consists of a sequence of accelerometer readings through time following a simulated motorcycle crash during an experiment to determine the efficacy of crash helmets [11]. Figure 1(a) shows the output of a heteroscedastic regularised kernel regression model for the Motorcycle dataset, using a common Gaussian radial basis kernel for both conditional mean and variance models,

$$\mathcal{K}^{\mu}(\boldsymbol{x}, \boldsymbol{x}') = \mathcal{K}^{\sigma}(\boldsymbol{x}, \boldsymbol{x}') = \exp\left\{-\lambda^{-2}\|\boldsymbol{x} - \boldsymbol{x}'\|^2\right\}. \tag{9}$$

where $\lambda = 13.1$, $\gamma^{\mu} = 2 \times 10^{-5}$ and $\gamma^{\sigma} = 1$. Note that the error bars are appropriately small where the variance of the data is least. The use of a heteroscedastic noise model also penalises errors more harshly in low noise regions of the data, leading to improved estimates of the conditional mean, for example eliminating the unwarranted undulation in the output of a conventional least-squares support vector machine, shown in figure 1(b), between $\approx 3 - 12ms$.

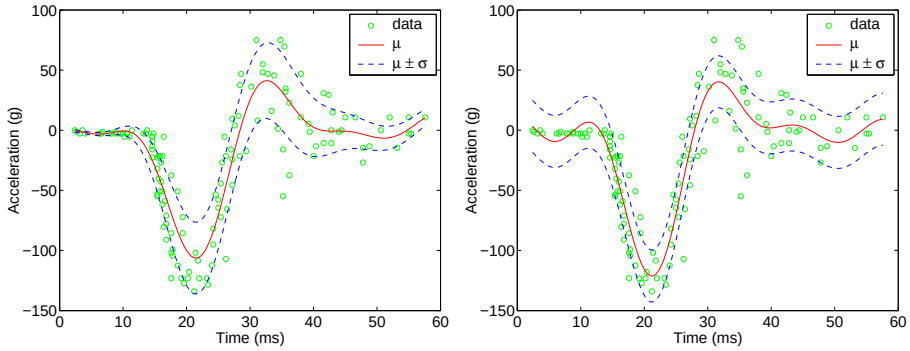

Figure 1: Heteroscedastic (a) and homoscedastic (b) regularised kernel regression model of the Motorcycle benchmark dataset.

3.2 Predicting Episodes of Poor air Quality

There are many diverse social, healthcare and economic problems associated with poor air quality. While government bodies have established threshold concentrations for a range of pollutants, the use of statistical modelling techniques to predict episodes of poor air quality is problematic, firstly because episodes of poor air quality are rare and on the decline due to a reduction in emissions, but also because different end users have different costs associated with false-positive and false-negative predictions. The output of a heteroscedastic regularised kernel regression model provides a full description of the target distribution describing the predicted level of a given concentration. Given a

vector, $\boldsymbol{x}$, summarising current meteorological and emissions data, the model provides not only a forecast of the most likely concentration, $\mu(\boldsymbol{x})$, but also of the probability that the observed concentration, y, exceeds a fixed threshold level, Y. The latter is obtained via integration of the upper tail of the predictive distribution,

$$p(y > Y \mid \boldsymbol{x}) = \int_Y^\infty \frac{1}{\sqrt{2\pi\sigma^2(\boldsymbol{x})}} \exp\left\{\frac{[\mu(\boldsymbol{x}) - z]^2}{2\sigma^2(\boldsymbol{x})}\right\} dz \tag{10}$$

(initial studies indicate that a heteroscedastic Gaussian distribution provides a reasonable approximation to the observed noise process). A single model can then be used for analysis of air quality time-series data, without the need for retraining to accommodate changes in threshold concentrations or misclassification costs.

Table 1: Comparison of LS-SVM and heteroscedastic regularised kernel regression (HRKR) models for prediction of daily mean SO_2 concentration in urban Belfast.

Statistic	**HRKR**	**LS-SVM**
Root-Mean-Squared Error	14.92	15.81
Negative Log Likelihood	2313	517.8
Cross-Entropy	3.48	6.572

LS-SVM and heteroscedastic regularised kernel regression (HRKR) networks were trained to predict the daily mean concentration of sulphur dioxide in urban Belfast, given inputs summarising the recent history of the $S0_2$ time-series and current meteorological conditions. Data from the years 1993-1996 were used in training and the models evaluated on data from the year 1998. In each case, the hyperparameters were determined via manual trial-and-error exploration of the search space. Table 1 shows a statistical comparison of LS-SVM and HRKR models. The HRKR model provides more accurate estimates of the conditional mean concentration, as illustrated by a lower root-mean-square error. The cross-entropy measure indicates that the HRKR also provides more accurate estimates of the probability of an exceedance than the LS-SVM model. It is well-known however that maximum likelihood estimates of the variance are biased; if over-fitting occurs in estimation of the conditional mean, the apparent noise density is unrealistically small. As a result the negative log-likelihood of the HRKR model is inferior to that of the LS-SVM.

4 Summary

A heteroscedastic regularised kernel regression model is introduced, which jointly estimates the conditional mean and variance of the target distribution. The model is then applied to the task of predicting episodes of poor air quality

in an urban environment. The use of a heteroscedastic noise model is demonstrated to provide improved estimates of the conditional mean of the target distribution and useful, although biased, error bars on predictions.

5 Acknowledgements

The authors would like to thank the anonymous referees for their helpful comments on the previous draft of this paper. This work was supported by the European Commission (grant number IST-99-11764), as part of its Framework V IST programme and by the Royal Society (research grant RSRG-22270).

References

[1] P. M. Williams. Using neural networks to model conditional multivariate densities. *Neural Computation*, 8:843–854, 1996.

[2] J. A. K. Suykens, J. De Brabanter, L. Lukas, and J. Vandewalle. Weighted least squares support vector machines : robustness and sparse approximation. *Neurocomputing (in press)*, 2001.

[3] C. Saunders, A. Gammerman, and V. Vovk. Ridge regression learning algorithm in dual variables. In *Proc., 15th Int. Conf. on Machine Learning*, pages 515–521, Madison, WI, July 24–27 1998.

[4] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[5] T. Poggio and Girosi F. Networks for approximation and learning. *Proc. of the IEEE*, 78(9), September 1990.

[6] D. A. Nix and A. S. Weigend. Estimating the mean and variance of the target probability distribution. In *Proc., Int. Conf. on Neural Networks*, volume 1, pages 55–60, 1994.

[7] A. N. Tikhonov and V. Y. Arsenin. *Solutions of ill-posed problems.* John Wiley, New York, 1977.

[8] S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58, 1992.

[9] G. S. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *J. Math. Anal. Applic.*, 33:82–95, 1971.

[10] I. T. Nabney. Efficient training of RBF networks for classification. Technical Report NCRG/99/002, Aston University, Birmingham, UK, 1999.

[11] B. W. Silverman. Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *J. Royal Statistical Society, B*, 47(1):1–52, 1985.