Approximately Unbiased Estimation of Conditional Variance in Heteroscedastic Kernel Ridge Regression

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

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

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

[*]Department of Electrical and Electronic Engineering,
Imperial College of Science, Technology and Medicine,
London, U.K. SW7 2BT.
d.mandic@ic.ac.uk

Abstract.
In this paper we extend a form of kernel ridge regression for data characterised by a heteroscedastic noise process (introduced in Foxall *et al.* [1]) in order to provide approximately unbiased estimates of the conditional variance of the target distribution. This is achieved by the use of the leave-one-out cross-validation estimate of the conditional mean when fitting the model of the conditional variance. The elimination of this bias is demonstrated on synthetic dataset where the true conditional variance is known.

It is well known that the minimisation of a sum-of-squares error (SSE) metric corresponds to maximum likelihood estimation of the parameters of a regression model, where the target data are assumed to be realisations of some deterministic process that have been corrupted by additive Gaussian noise with constant variance (i.e. a *homoscedastic* noise process) (e.g. Bishop [2]). Several kernel learning methods based on the minimisation of a regularised sum-of-squares have been proposed (e.g. [3–5]). In Foxall *et al.*, we extend this family of models to include a formulation that is optimal for a Gaussian noise process with input-dependent (heteroscedastic) variance (c.f. [6]). In this paper, we overcome a major shortcoming of existing approaches, by adopting the leave-one-out cross-validation estimate of the conditional mean in fitting the model of the conditional variance, resulting in almost unbiased predictive error bars. The form of the model of the conditional mean allows a particularly efficient closed-form implementation of the leave-one-out procedure. The unbiased nature of the estimates of conditional variance is demonstrated on a synthetic dataset where the true conditional variance is known.

1 Heteroscedastic Kernel Ridge Regression

Suppose we are given data $\mathcal{D} = \{\boldsymbol{x}_i, y_i\}_{i=1}^{\ell}$, $\boldsymbol{x}_i \in \mathcal{X} \subset \mathbb{R}^d$, $y_i \in \mathcal{Y} \subset \mathbb{R}$, where the targets, y_i, are assumed to be the output of a deterministic system, 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, i.e. $y_i = \mu(\boldsymbol{x}_i) + \epsilon_i$, $\epsilon_i \sim \mathcal{N}(0,\ \sigma(\boldsymbol{x}_i))$. The conditional probability density of target y_i, given input vector $\boldsymbol{x}_i$ is given by

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

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, $\mathcal{F}^\mu$ ($\boldsymbol{\phi}^\mu : \mathcal{X} \to \mathcal{F}^\mu$). Space $\mathcal{F}^\mu$ 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 $\mathcal{K}^\sigma$. Note that the output of this model represents the natural logarithm of the standard deviation to ensure that the corresponding estimate of conditional standard deviation is strictly positive. 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

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

Clearly this corresponds to quadratic regularisation of a maximum likelihood cost function, where γ^μ and γ^σ are regularisation parameters, providing independent control of the bias-variance trade-off [2] for the models of the conditional mean and standard deviation. The optimal values of $\boldsymbol{w}^\mu$ and $\boldsymbol{w}^\sigma$ can be written as expansions over training patterns [7], such that

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

However the training algorithm for the heteroscedastic kernel ridge regression model is somewhat more complex as the variance of the noise process is no longer constant.

1.1 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 (see e.g. [8]), alternating updates of the mean and standard deviation models. If $\sigma(\boldsymbol{x}_i)$, $\forall i \in \{1, 2, \dots, \ell\}$ 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

$$L^\mu = \frac{1}{2}\gamma^\mu \|\boldsymbol{w}^\mu\|^2 + \sum_{i=1}^{\ell} \zeta_i \{\mu(\boldsymbol{x}_i) - y_i\}^2, \tag{3}$$

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 [3], and so is minimised by the solution of the set of linear equations. Likewise, if $\mu(\boldsymbol{x}_i)$, $\forall i \in \{1, 2, \dots, \ell\}$ 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

$$L^\sigma = \frac{1}{2}\gamma^\sigma \|\boldsymbol{w}^\sigma\|^2 + \sum_{i=1}^{\ell} \left[z_i + \xi_i \exp\{-2z_i\}\right], \tag{4}$$

where $\xi_i = \frac{1}{2}[\mu(\boldsymbol{x}_i) - y_i]^2$ and $z_i = \boldsymbol{w}^\sigma \cdot \boldsymbol{\phi}^\sigma(\boldsymbol{x}_i) + b^\sigma = \sum_{j=1}^{\ell} \alpha_j^\sigma \mathcal{K}^\sigma(\boldsymbol{x}_i, \boldsymbol{x}_j) + b^\sigma$. The model of the conditional standard deviation can then be updated efficiently via a simple Newton-Raphson iterative process.

2 Eliminating Bias in the Conditional Variance

It is well known that maximum likelihood estimates of variance-like quantities are biased (e.g. Bishop [2]). If the model of the conditional mean of the target distribution over-fits the training data, the apparent variance of the noise process acting on the training data is reduced. This means that the corresponding estimate of the conditional variance will be unrealistically small. To overcome this bias, the leave-one-out cross-validation estimate of the conditional mean is substituted when updating the model of the conditional variance, via minimisation of (4). It seems reasonable to suggest that the leave-one-out estimate of the conditional mean will be less susceptible to over-fitting and so the estimated conditional variance will be significantly less biased. It is straightforward to show that the minimiser of the objective function for the model of the conditional mean (3) is given by

$$\boldsymbol{p} = (\boldsymbol{R} + \boldsymbol{Z}^T \mathrm{diag}(\boldsymbol{\zeta})\boldsymbol{Z})^{-1}\boldsymbol{Z}^T \mathrm{diag}(\boldsymbol{\zeta})\boldsymbol{y} \tag{5}$$

where $\boldsymbol{p} = (\boldsymbol{\alpha}^{\boldsymbol{\mu}T},\ b^\mu)^T$, $\boldsymbol{Z} = [\boldsymbol{K}^\mu\ \boldsymbol{1}]$, $\boldsymbol{R} = \left[\frac{\gamma^\mu}{2}\boldsymbol{K}^\mu,\ \boldsymbol{0}\ ;\ \boldsymbol{0}^T,\ 0\right]$. and $\boldsymbol{0} = (0, 0, \ldots, 0)^T$. The similarity of the system of linear equations (5) giving the optimal parameters of the model of the conditional mean and the normal equations arising in multiple linear regression admits a particularly efficient implementation of the leave-one-out cross-validation procedure, well known in the field of statistics [9]. For convenience, let $\boldsymbol{U} = \mathrm{diag}(\boldsymbol{\zeta})\boldsymbol{Z}$, $\boldsymbol{C} = \boldsymbol{R} + \boldsymbol{U}^T\boldsymbol{Z}$ and $\boldsymbol{d} = \boldsymbol{U}^T\boldsymbol{t}$, such that $\boldsymbol{p} = \boldsymbol{C}^{-1}\boldsymbol{d}$. Furthermore, let $\boldsymbol{Z}_{(i)}$, $\boldsymbol{U}_{(i)}$ and $\boldsymbol{y}_{(i)}$ represent matrices $\boldsymbol{Z}$, $\boldsymbol{U}$ and vector $\boldsymbol{y}$ with the i^{th} observation deleted, then

$$\boldsymbol{C}_{(i)} = \boldsymbol{C} - \boldsymbol{u}_i\boldsymbol{z}_i^T, \quad \text{and} \quad \boldsymbol{d}_{(i)} = \boldsymbol{d} - y_i\boldsymbol{z}_i.$$

The inverse of the "downdated" matrix, $\boldsymbol{C}_{(i)}$, can then be found with a computational complexity of only $\mathcal{O}(\ell^2)$ operations via the the Bartlett matrix inversion formula [10],

$$\boldsymbol{C}_{(i)}^{-1} = \boldsymbol{C}^{-1} + \frac{\boldsymbol{C}^{-1}\boldsymbol{u}_i\boldsymbol{z}_i^T\boldsymbol{C}^{-1}}{1 - \boldsymbol{z}_i^T\boldsymbol{C}^{-1}\boldsymbol{u}_i},$$

Let $\boldsymbol{H} = \boldsymbol{Z}\boldsymbol{C}^{-1}\boldsymbol{U}^T$ represent the *hat* matrix (in multiple linear regression the hat, or projection matrix $\boldsymbol{H}$ maps the desired output $\boldsymbol{y}$ onto the output of the model $\hat{\boldsymbol{y}} = \boldsymbol{X}(\boldsymbol{X}\boldsymbol{X}^T)^1\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y}$ [9]). Following a straight-forward, but somewhat lengthy series of algebraic manipulations, we obtain

$$\left\{\boldsymbol{\mu}_{(i)}\right\}_i = \mu_i - \frac{h_{ii}}{1 - h_{ii}}r_i. \tag{6}$$

where $r_i = u_i - \boldsymbol{z}_i^T\boldsymbol{p}$ is the residual error for the i^{th} training pattern for the full model. The leave-one-out estimate of the mean of the target distribution given by (6) can then be substituted when fitting the model of the conditional standard deviation, such that $\xi_i = \frac{1}{2}\left[\{\mu_{(i)}(\boldsymbol{x}_i)\}_i - y_i\right]^2$.

3 Results

In this section we demonstrate that the leave-one-out kernel ridge regression model provides almost unbiased estimates of the conditional standard deviation using a synthetic regression problem, taken from Williams [6], in which the true conditional standard deviation is known exactly. The univariate input patterns, x, are drawn from a uniform distribution on the interval $(0,\ \pi)$, the corresponding targets, y, are draw from a univariate Normal distribution with mean and variance that vary smoothly with x:

$$x \sim \mathcal{U}(0, \pi), \text{ and } y \sim \mathcal{N}\left(\sin\left\{\frac{5x}{2}\right\}\sin\left\{\frac{3x}{2}\right\}, \frac{1}{100} + \frac{1}{4}\left[1 - \sin\left\{\frac{5x}{2}\right\}\right]^2\right).$$

Figure 1 (a) and (b) show the arithmetic mean of the predicted conditional mean and ± one standard deviation credible interval for simple and leave-one-out heteroscedastic kernel ridge regression models respectively, over 1000 randomly generated datasets of 64 patterns each. A radial basis function kernel was used, with width parameter, $\lambda = 2$, for both the model of the conditional mean and the model of the conditional standard deviation, the regularisation parameters were set as follows: $\gamma^\mu = \gamma^\sigma = 1$. In both cases the fitted mean is, on average, in good agreement with the true mean. Figure 1 (c) and (d) show the arithmetic mean of the predicted conditional standard deviation for the simple and leave-one-out heteroscedastic kernel ridge regression models. The simple heteroscedastic kernel ridge regression model, on average, consistently under-estimates the conditional standard deviation, and so the predicted credible intervals are optimistically narrow. The mean predicted conditional standard deviation for the leave-one-out heteroscedastic kernel ridge regression model is very close to the true value. This suggests that the estimation of the conditional standard deviation is (almost) unbiased as the the expected value is approximately equal to the true value.

4 Summary

An improved heteroscedastic kernel ridge regression model is introduced, which jointly estimates the conditional mean and variance of the target distribution, eliminating the bias inherent in maximum likelihood estimates of conditional variance through the use of the leave-one-out estimate of the conditional mean in fitting the model of the conditional variance. The resulting estimates of conditional variance are shown experimentally to be approximately unbiased using a synthetic dataset where the true variance is known.

5 Acknowledgements

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).

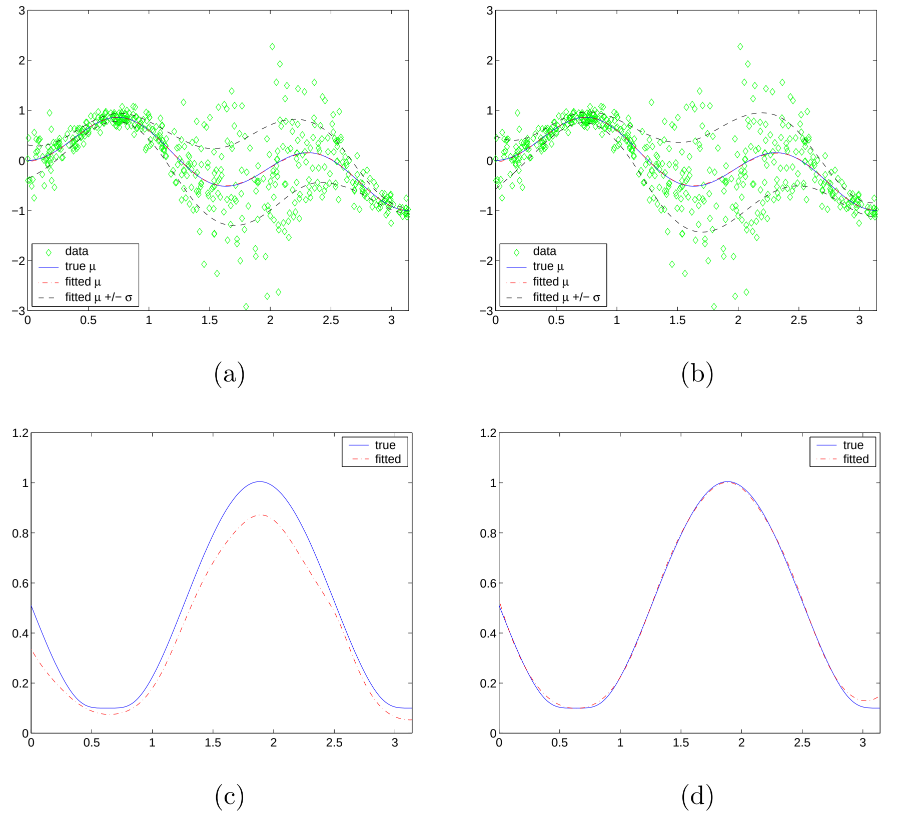
References

[1] R. J. Foxall, G. C. Cawley, N. L. C. Talbot, S. R. Dorling, and D. P. Mandic. Heteroscedastic regularised kernel regression for prediction of episodes of poor air quality. In *Proc. ESANN*, pages 19–24, Bruges, Belgium, April 2002.

[2] C. M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, 1995.

[3] J. A. K. Suykens, J. De Brabanter, L. Lukas, and J. Vandewalle. Weighted least squares support vector machines : robustness and sparse approximation. *Neurocomputing*, 48, October 2002.

[4] 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.



(a)

(b)

(c)

(d)

Figure 1: Arithmetic mean of the estimate of the conditional mean and ± one standard deviation credible interval for (a) HKRR and (b) LOOHKRR models for a synthetic regression problem, (c) and (d) display the corresponding means of the estimated conditional standard deviation. All graphs show average results computed over 1000 randomly generated datasets.

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

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

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

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

[9] R. D. Cook and S. Weisberg. *Residuals and Influence in Regression*. Monographs on Statistics and Applied Probability. Chapman and Hall, New York, 1982.

[10] M. S. Bartlett. An inverse matrix adjustment arising in disciminant analysis. *Annals of Mathematical Statistics*, 22(1):107–111, 1951.