Heteroscedastic Kernel Ridge Regression

Gavin C. Cawley [a,*,1] Nicola L. C. Talbot Robert J. Foxall, [a]

Stephen R. Dorling [b] Danilo P. Mandic [c]

[a] *School of Information Systems, University of East Anglia, Norwich, U.K. NR4 7TJ*

[b] *School of Environmental Sciences, University of East Anglia, Norwich, U.K. NR4 7TJ*

[c] *Department of Electrical and Electronic Engineering, Imperial College of Science, Technology and Medicine, London, U.K. SW7 2BT*

Abstract

In this paper we extend a form of kernel ridge regression for data characterised by a heteroscedastic (i.e. input dependent variance) Gaussian noise process, introduced in Foxall *et al.* [1]. It is shown that the proposed heteroscedastic kernel ridge regression model can give a more accurate estimate of the conditional mean of the target distribution than conventional kernel ridge regression and also provides an indication of the spread of the target distribution (i.e. predictive error bars). The leave-one-out cross-validation estimate of the conditional mean is used in fitting the model of the conditional variance in order to overcome the inherent bias in maximum likelihood estimates of the variance. The benefits of the proposed model are demonstrated on synthetic and real-world benchmark data sets and for the task of predicting episodes of poor air quality in an urban environment.

Key words: kernel methods, non-linear regression, heteroscedasticity

It is well known that 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]). The least-squares support vector machine [3], kernel ridge-regression [4, 5] and regularisation network [6] form a family of closely related techniques that implement non-linear regression using a linear model constructed in a fixed feature space induced by a Mercer kernel, minimising a regularised sum-of-squares error criterion. In this paper, we extend this family to include a formulation that is optimal for a Gaussian noise process with input-dependent (heteroscedastic) variance. Linear models are constructed in a kernel induced feature space, estimating both the conditional mean *and* conditional variance of the target distribution, using a regularised maximum likelihood criterion [7–9]. This results in both robust estimates of the conditional mean and also a more realistic credible interval on predictions (i.e. predictive error bars). Furthermore, 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 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. We then apply the proposed method to synthetic and real-world benchmark datasets, and to the practical problem of predicting episodes of poor air quality, in terms of

* Corresponding author, email: gcc@sys.uea.ac.uk

[1] 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).

both an estimate of the concentration of a given pollutant and an estimate of the probability that the predicted concentration exceeds a given statutory threshold level.

The remainder of this paper is organised as follows: The conventional homoscedastic kernel ridge regression algorithm is briefly described in section 1, introducing the notation used throughout. Section 2 introduces a heteroscedastic form of the kernel ridge regression algorithm, with an efficient training algorithm given in section 3. The elimination of the bias inherent in estimation of the conditional variance is discussed in section 4. Section 5 evaluates the conventional and heteroscedastic kernel ridge regression algorithms on a synthetic dataset (demonstrating that the estimates of conditional variance are approximately unbiased), the well-known motorcycle benchmark dataset and for predicting episodes of poor air quality in urban environments. Finally the proposed approach is summarised in section 6.

1 Kernel Ridge Regression

Ridge regression [4] is a well known technique from classical multiple linear regression that implements a regularised form of least-squares regression. Given training data,

$$\mathcal{D} = \{\boldsymbol{x}_i, y_i\}_{i=1}^{\ell}, \quad \boldsymbol{x}_i \in \mathcal{X} \subset \mathbb{R}^d, \quad y_i \in \mathcal{Y} \subset \mathbb{R},$$

the ridge regression algorithm determines the parameter vector, $\boldsymbol{w} \in \mathbb{R}^d$, and bias, $b \in \mathbb{R}$, of a linear model, $f(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{x} + b$, via minimisation of the

following objective function:

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

Clearly the objective function used in ridge regression (1) implements a form of Tikhonov regularisation [10] of a sum-of-squares error metric, where γ is a regularisation parameter controlling the bias-variance trade-off [11]. This corresponds to penalised maximum likelihood estimation of $\boldsymbol{w}$ and b, assuming the targets have been corrupted by an independent and identically distributed (i.i.d.) sample from a Gaussian noise process, with zero mean and fixed variance σ^2, i.e.

$$y_i = \boldsymbol{w}\cdot\boldsymbol{x}_i + b + \epsilon_i, \quad \epsilon \sim \mathcal{N}(0, \sigma^2).$$

A non-linear form of ridge regression, known as kernel ridge regression [5], can be obtained via the so-called "kernel trick", whereby a linear ridge regression model is constructed in a high dimensional feature space, $\mathcal{F}$ ($\boldsymbol{\phi} : \mathcal{X} \to \mathcal{F}$), induced by a non-linear kernel function defining the inner product $\mathcal{K}(\boldsymbol{x}, \boldsymbol{x'}) = \boldsymbol{\phi}(\boldsymbol{x})\cdot\boldsymbol{\phi}(\boldsymbol{x'})$. The kernel function, $\mathcal{K} : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ may be any positive definite "Mercer" kernel (for an overview of kernel learning methods, including kernel ridge regression, see Cristianini and Shawe-Taylor [12]). The objective function minimised in constructing a kernel ridge regression model is given by

$$L(\boldsymbol{w}, b) = \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{\gamma}{\ell}\sum_{i=1}^{\ell}(y_i - \boldsymbol{w}\cdot\boldsymbol{\phi}(\boldsymbol{x_i}) - b)^2.$$

The representer theorem [13] states that the solution of an optimisation problem of this nature can be written in the form of a linear combination of the training patterns, i.e. $\boldsymbol{w} = \sum_{i=1}^{\ell}\alpha_i\boldsymbol{\phi}(\boldsymbol{x}_i)$. The output of the least-squares support vector machine is then given by the kernel expansion

$$f(\boldsymbol{x}) = \sum_{i=1}^{\ell}\alpha_i\mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}) + b.$$

It can easily be shown [5, 14] that the optimal coefficients of this expansion are given by the solution of a set of $\ell + 1$ linear equations in $\ell + 1$ unknowns:

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

where $\boldsymbol{\Omega} = \boldsymbol{K} + \ell\gamma^{-1}\boldsymbol{I}$, $\boldsymbol{K} = [k_{ij} = \mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{i,j=1}^{\ell}$, $\boldsymbol{I}$ is the $\ell \times \ell$ identity matrix, $\boldsymbol{y} = (y_1, y_2, \ldots, y_\ell)^T$, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_\ell)^T$ and $\mathbf{1} = (1, 1, \ldots, 1)^T$.

2 Heteroscedastic Kernel Ridge Regression

Suppose we are given a dataset $\mathcal{D}$ where the targets, y_i, are assumed to be corrupted by an independent and identically distributed[2] (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 conditional probability density of target y_i, given input vector $\boldsymbol{x}_i$ is given by

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

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

$$-\log \mathcal{L}_{\mathcal{D}} = \sum_{i=1}^{\ell} \left\{\log \sigma(\boldsymbol{x}_i) + \frac{[\mu(\boldsymbol{x}_i) - y_i]^2}{2\sigma^2(\boldsymbol{x}_i)}\right\}, \tag{3}$$

where $\mathcal{L}_{\mathcal{D}}$ represents the likelihood of $\mathcal{D}$. 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, $\mathcal{F}^\mu$ ($\boldsymbol{\phi}^\mu$:

[2] By identically distributed we mean that the *conditional* distribution is identical for all samples, although the variance of the noise process is different for samples collected from different regions of $\mathcal{X}$

$\mathcal{X} \rightarrow \mathcal{F}^{\mu}$). Space $\mathcal{F}^{\mu}$ is induced by a positive definite "Mercer" kernel, $\mathcal{K}^{\mu} : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$, 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

$$L(\boldsymbol{w}^{\mu}, b^{\mu}, \boldsymbol{w}^{\sigma}, b^{\sigma}) = \frac{1}{2}\gamma^{\mu}\|\boldsymbol{w}^{\mu}\|^2 + \frac{1}{2}\gamma^{\sigma}\|\boldsymbol{w}^{\sigma}\|^2 + \sum_{i=1}^{\ell} \left\{ \log \sigma(\boldsymbol{x}_i) + \frac{[\mu(\boldsymbol{x}_i) - y_i]^2}{2\sigma^2(\boldsymbol{x}_i)} \right\}. \quad (4)$$

Clearly this corresponds to quadratic regularisation [10] of a maximum likelihood cost function, where γ^{μ} and γ^{σ} are regularisation parameters, providing independent control of the bias-variance trade-off [11] for the models of the conditional mean and standard deviation. The representer theorem [13] suggests that the optimal values of $\boldsymbol{w}^{\mu}$ and $\boldsymbol{w}^{\sigma}$ can be written as expansions over training patterns (see Appendix A for details),

$$\boldsymbol{w}^{\mu} = \sum_{i=1}^{\ell} \alpha_i^{\mu} \boldsymbol{\phi}^{\mu}(\boldsymbol{x}_i)$$

and

$$\boldsymbol{w}^{\sigma} = \sum_{i=1}^{\ell} \alpha_i^{\sigma} \boldsymbol{\phi}^{\sigma}(\boldsymbol{x}_i),$$

such that

$$\mu(\boldsymbol{x}) = \sum_{i=1}^{\ell} \alpha_i^{\mu} \mathcal{K}^{\mu}(\boldsymbol{x}, \boldsymbol{x}_i) + b^{\mu}$$

and

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

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.

3 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 [15], alternating updates of the mean and standard deviation models.

3.1 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

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

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 [3], 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 $\boldsymbol{\Omega} = (\boldsymbol{K}^\mu + \boldsymbol{D})$, $\boldsymbol{K}^\mu = \left\{k_{ij}^\mu = \mathcal{K}^\mu(\boldsymbol{x}_i, \boldsymbol{x}_j)\right\}_{i,j=1}^{\ell}$, $\mathbf{1} = (1, 1, \ldots, 1)^T$, $\boldsymbol{y} = (y_i, y_2, \ldots, y_\ell)^T$, $\boldsymbol{\alpha}^\mu = (\alpha_i^\mu, \alpha_2^\mu, \ldots, \alpha_\ell^\mu)^T$ and $\boldsymbol{D}$ is a diagonal matrix with

elements $\gamma^{\mu}/(\zeta_1, \zeta_2, \ldots, \zeta_\ell)$.

3.2 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

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

where $\xi_i = \frac{1}{2}[\mu(\boldsymbol{x}_i) - y_i]^2$ and $z_i = \boldsymbol{w}^{\sigma} \cdot \boldsymbol{\phi}^{\sigma}(\boldsymbol{x}_i) + b^{\sigma} = \sum_{j=1}^{\ell} \alpha_j^{\sigma} \mathcal{K}^{\sigma}(\boldsymbol{x}_i, \boldsymbol{x}_j) + b^{\sigma}$. 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.

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

Note that while this approach generally leads to a reduction in both L^{σ} and L, it is possible to take too large a step, producing an increase in L or L^{σ}, or both. We therefore advocate a simple step-halving procedure, where steps are taken in the direction given by $-\boldsymbol{H}^{-1}\boldsymbol{\nabla}$, but the length of the step taken being halved at each iteration until a reduction in L is achieved, beginning with a full Newton step.

3.3 Convergence and Stability

Minimisation of the objective functions, L^{μ} and L^{σ}, can be shown to constitute convex optimisation problems, i.e. their respective Hessian matrices are

positive semi-definite, and therefore posses single, global minima, L however is non-convex (see Appendix B). Fortunately although convexity is a guarantee of the presence of a single, global minimum, a non-convex optimisation problem may still be free of local minima, as illustrated by figure 1. Repeated minimisation of L from different random initial starting points reliably converges to essentially identical solutions; this strongly suggests that although L is non-convex, it nevertheless has a single, global minimum. Clearly a gradient descent optimisation strategy is guaranteed to converge to a local minima as long as the learning rate is sufficiently small that an update of the model parameters never results in an increase in the objective function; provided that the objective function is also continuous, there will always be a finite learning rate such that this condition is satisfied. In the first step, in minimising L^μ we effectively minimise the overall cost function L with respect to $(\boldsymbol{\alpha}^\mu, b^\mu)$, whilst holding $(\boldsymbol{\alpha}^\sigma, b^\sigma)$ constant. This step is achieved analytically and so is guaranteed not to result in an increase in L^μ or in L itself. For the second step, note that the partial derivatives of L^σ and L with respect to the parameters of the model of the conditional standard deviation are identical, i.e.

$$\frac{\partial L^\sigma}{\partial \boldsymbol{\alpha}^\sigma} = \frac{\partial L}{\partial \boldsymbol{\alpha}^\sigma} \quad \text{and} \quad \frac{\partial L^\sigma}{\partial b^\sigma} = \frac{\partial L}{\partial b^\sigma}.$$

Following an update of the model of the conditional mean, since the μ-step is analytic, we know that

$$\frac{\partial L}{\partial \boldsymbol{\alpha}^\mu} = \mathbf{0} \quad \text{and} \quad \frac{\partial L}{\partial b^\mu} = 0,$$

As a result, there always exists a sufficiently small learning rate such that a gradient descent step minimising L^σ corresponds to a gradient descent step minimising L. A simple Newton-Raphson algorithm is used to adapt the learning rate in order to obtain rapid convergence. It is possible for the learning rate

chosen in this way to be sufficiently large that a reduction in L^σ is not accompanied by a reduction in L. We therefore adopt a simple step-halving strategy where the step selected by the Newton-Raphson process is successively halved in magnitude until a step is found that minimises L as well as L^σ, however in practise step halving is rarely required. The stability of the training algorithm and convergence to a local minimum of L are therefore assured.

4 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 (7) [16]. 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. Normally the computational expense would be prohibitive, however in this case the conditional mean is given by a model that is linear in its parameters, minimising a regularised weighted sum-of-squares cost function, and so leave-one-out cross-validation can be performed very efficiently in closed form.

4.1 Fast Leave-One-Out Cross-Validation

In this section, we introduce a fast algorithm giving a close approximation to the leave-one-out cross-validation error of regularisation networks, based on similar algorithms for multiple linear regression models that have been known to the field of statistics for some time (see e.g. Cook and Weisberg [17]). We begin by providing an alternative formulation of the solution to the optimisation problem defined in the previous section before deriving closed-form expressions for the leave-one-out cross-validation behaviour of this family of kernel learning methods.

4.1.1 Alternative Formulation of the Optimisation Problem

We begin by reformulating the solution to the optimisation problem specified by the objective function L^μ (5). Setting the partial derivatives of (5) with respect to $\boldsymbol{\alpha}^\mu$ and b^μ to zero and dividing through by two gives:

$$\sum_{i=1}^{\ell} \alpha_i^\mu \left(\frac{\gamma^\mu}{2} k_{ir}^\mu + \sum_{i=1}^{\ell} k_{ij}^\mu \zeta_j k_{rj}^\mu \right) + b^\mu \sum_{i=1}^{\ell} \zeta_i k_{ri}^\mu = \sum_{i=1}^{\ell} \zeta_i y_i k_{ri}^\mu,$$

$r = 1, 2, \ldots, \ell$, and

$$\sum_{i=1}^{\ell} \alpha_i^\mu \sum_{j=1}^{\ell} \zeta_j k_{ij}^\mu + b^\mu \sum_{i=1}^{\ell} \zeta_i = \sum_{i=1}^{\ell} \zeta_i y_i$$

It is straightforward to show that these equations can be written more concisely in the form

$$(\boldsymbol{R} + \boldsymbol{Z}^T \mathrm{diag}(\boldsymbol{\zeta}) \boldsymbol{Z}) \boldsymbol{p} = \boldsymbol{Z}^T \mathrm{diag}(\boldsymbol{\zeta}) \boldsymbol{y} \tag{9}$$

where $\boldsymbol{p} = (\boldsymbol{\alpha}^{\boldsymbol{\mu}T},\ b^{\mu})^T$, $\boldsymbol{Z} = [\boldsymbol{K}^{\mu}\ \mathbf{1}]$ and

$$\boldsymbol{R} = \begin{bmatrix} \frac{\gamma^{\mu}}{2}\boldsymbol{K}^{\mu} & \mathbf{0} \\ \mathbf{0}^T & 0 \end{bmatrix}.$$

where $\mathbf{0} = (0, 0, \ldots, 0)^T$. Note the similarity of (9) to the so-called *normal equations* arising in the solution of linear multiple regression problems, in this case including an additional matrix, $\boldsymbol{R}$, representing the regularisation term in (5) and the vector of weighting factors, $\boldsymbol{\zeta}$. This formulation permits an efficient implementation of the leave-one-out cross-validation procedure for the model of the conditional mean.

4.1.2 Leave-One-Out Cross-Validation

The similarity of the system of linear equations (9) 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 [17]. The optimal vector of model parameters $\boldsymbol{p} = (\boldsymbol{\alpha}^{\boldsymbol{\mu}T},\ b^{\mu})^T$ is given by

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

where $\boldsymbol{Z} = [\boldsymbol{K}^{\mu}\ \mathbf{1}]$. 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

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

To reduce computational complexity, the Bartlett matrix inversion formula [18] then gives

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

such that the vector of model parameters during the i^{th} iteration of the leave-one-out cross-validation procedure becomes

$$\boldsymbol{p}_{(i)} = \left(\boldsymbol{C}^{-1} + \frac{\boldsymbol{C}^{-1}\boldsymbol{u}_i\boldsymbol{z}_i^T\boldsymbol{C}^{-1}}{1 - \boldsymbol{z}_i^T\boldsymbol{C}^{-1}\boldsymbol{u}_i}\right)(\boldsymbol{d} - y_i\boldsymbol{u}_i).$$

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 $\hat{\boldsymbol{y}} = \boldsymbol{X}(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y}$ [17]); note that the i^{th} element of the leading diagonal can be written $h_{ii} = \boldsymbol{z}_i^T\boldsymbol{C}^{-1}\boldsymbol{u}_i$, so expanding the brackets we have

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

which can be rearranged to give

$$\boldsymbol{p}_{(i)} = \boldsymbol{p} + \left(\frac{\boldsymbol{z}_i^T\boldsymbol{p} - y_i}{1 - h_{ii}}\right)\boldsymbol{C}^{-1}\boldsymbol{u}_i.$$

The residual error for the i^{th} training pattern for the full model is $r_i = y_i - \boldsymbol{z}_i^T\boldsymbol{p}$ and so

$$\boldsymbol{p}_{(i)} = \boldsymbol{p} - \frac{r_i}{1 - h_{ii}}\boldsymbol{C}^{-1}\boldsymbol{u}_i.$$

Noting that $\boldsymbol{\mu} = \boldsymbol{Z}\boldsymbol{p}$, the output of the model during the i^{th} iteration of the leave-one-out cross-validation procedure can be written as

$$\boldsymbol{\mu}_{(i)} = \boldsymbol{Z}\boldsymbol{p}_{(i)} = \boldsymbol{\mu} - \frac{r_i}{1 - h_{ii}}\boldsymbol{h}_i$$

where $\boldsymbol{h}_i$ is the i^{th} column of $\boldsymbol{H}$, and therefore

$$\left\{\boldsymbol{\mu}_{(i)}\right\}_i = \mu_i - \frac{h_{ii}}{1 - h_{ii}}r_i. \tag{10}$$

The leave-one-out estimate of the mean of the target distribution given by (10) can then be substituted when fitting the model of the conditional standard deviation, such that

$$\xi_i = \frac{1}{2}\left[\{\mu_{(i)}(\boldsymbol{x}_i)\}_i - y_i\right]^2.$$

4.2 Convergence and Stability

It should be noted that while minimisation of L^μ and L^σ still represent convex optimisation problems, it is no longer straightforward to establish whether the optimisation problem corresponding to the overall cost function L, is convex or non-convex or the convergence and stability of the training procedure, as established for the first model in section 3.3. This is a result of the decoupling the models of the conditional mean and conditional variance imposed by the leave-one-out cross-validation procedure. However, the convergence and stability of the training procedure have not proved problematic in empirical work to date.

5 Results

In this section, we evaluate the conventional homoscedastic kernel ridge regression (KRR), heteroscedastic kernel ridge regression (HKRR) and leave-one-out heteroscedastic kernel ridge regression (LOOHKRR) algorithms on three example datasets: Firstly, the well-known motorcycle benchmark provides a univariate non-linear regression task characterised by a heteroscedastic noise process that can be easily visualised. Secondly we apply the heteroscedastic methods to a synthetic regression task, where the true conditional variance is

known, demonstrating that the estimates of conditional variance given by the leave-one-out heteroscedastic algorithm are approximately unbiased. Lastly, all three algorithms are applied to the task of predicting episodes of poor air quality in urban Belfast, an application where an indication of the spread of the target distribution greatly increases the flexibility of the model.

5.1 The Motorcycle Benchmark

The Motorcycle benchmark consists of a sequence of accelerometer readings through time following a simulated motorcycle crash performed during experiments to determine the efficacy of crash helmets [19]. Figure 2 shows the output of homoscedastic, heteroscedastic and leave-one-out heteroscedastic kernel ridge regression models for the Motorcycle dataset. In each case a Gaussian radial basis kernel was used,

$$\mathcal{K}(\boldsymbol{x}, \boldsymbol{x}') = \exp\left\{-\lambda^{-2}\|\boldsymbol{x} - \boldsymbol{x}'\|_2^2\right\}. \tag{11}$$

where the optimal kernel and regularisation parameters, $\lambda^{\mu}, \lambda^{\sigma}, \gamma^{\mu}$ and γ^{σ}, were selected to minimise a 10-fold cross-validation estimate of the negative log-likelihood, using a simple Nelder-Mead simplex optimisation procedure [20] (see table B.1). Note that the error bars for both heteroscedastic kernel ridge regression models (figure 2(b) and (c)) are appropriately small where the variance of the data is least. As might be expected, the error bars for the leave-one-out variant are slightly broader than for the HKRR model, as shown in figure 2(c). The use of a heteroscedastic noise model also penalises errors more harshly in low noise regions of the data, leading to qualitatively improved estimates of the conditional mean, for example eliminating the unwarranted undulation in the output of the conventional homoscedastic kernel

ridge regression model, shown in figure 2 (a), between $\approx (3 - 12)ms$.

The leave-one-out cross-validation estimates of the sum-of-squared error and negative log-likelihood statistics for each model over the Motorcycle data set are given in table B.2. Both heteroscedastic forms of kernel ridge regression models provide quantitatively better descriptions of the dataset than the conventional homoscedastic form, as indicated by the improved negative log-likelihood scores, the LOOHKRR model performing best of all. The LOOHKRR model also provides a modest reduction in the sum-of-squares error.

5.2 Synthetic Dataset

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 [9], 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 drawn from a univariate Normal distribution with mean and variance that vary smoothly with x:

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

Figure 3, parts (a) and (b), show the arithmetic mean of the predicted conditional mean and $\pm$ 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 3, parts (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 expected value is approximately equal to the true value.

5.3 Predicting Episodes of Poor Air Quality

There are many diverse social, health-care 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 giving the predicted concentration of a given pollutant. 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,

$$p(y > Y \mid \boldsymbol{x}) = \int_Y^\infty \frac{1}{\sqrt{2\pi\sigma^2(\boldsymbol{x})}} \exp\left\{\frac{[\mu(\boldsymbol{x}) - z]^2}{2\sigma^2(\boldsymbol{x})}\right\} dz \tag{12}$$

(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.

Conventional homoscedastic and heteroscedastic and leave-one-out heteroscedastic kernel ridge regression networks were trained to predict the daily mean concentration of sulphur dioxide in urban Belfast, given inputs summarising the recent history of the SO_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 hyper-parameters were determined via manual trial-and-error exploration of the search space, maximising the likelihood over the test data.

Table B.3 shows a statistical comparison of KRR, HKRR and LOOHKRR models. The HKRR model provides more accurate estimates of the conditional mean concentration, as illustrated by a lower sum-of-squared error. The cross-entropy measure for the task of predicting exceedences, using (12), indicates that the HKRR and LOOHKRR models also provide more accurate estimates of the probability of an exceedance than the KRR model. It is well-known however that maximum likelihood estimates of the variance are bi-

ased; 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 HKRR model is inferior to that of the KRR. The LOOHKRR model improves somewhat on the negative log-likelihood of the HKRR model and also improves on the cross-entropy for the prediction task, while the sum-of-squares error remains the same. The negative log-likelihood for the LOOHKRR model however remains greater than that of the conventional homoscedastic kernel ridge regression model (KRR). This occurs because the negative log-likelihood statistic is more sensitive to outliers in the test data for the HKRR and LOOHKRR models than for the KRR model. Test data with a relatively small divergence from the predicted mean given by the model at a point in the input space where the model is most confident (i.e. the estimate of the conditional variance is low) can disproportionately inflate the negative log-likelihood. The improved sum-of-squares and cross-entropy statistics however demonstrate the superiority of the LOOHKRR model.

6 Summary

A heteroscedastic kernel ridge regression model is introduced, which jointly estimates the conditional mean and variance of the target distribution. An efficient training algorithm is provided, which can easily be shown to be stable and convergent to the global optimum of the cost function. Furthermore this model is extended to eliminate the bias inherent in maximum likelihood estimates of conditional variance through the use of the leave-one-out estimate of the conditional mean when 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. The model is then applied to the Motorcyle benchmark dataset and 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 a qualitatively and quantitatively better description of the dataset than is achieved using a conventional regularised sum-of-squares cost function (i.e. kernel ridge regression).

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 *Proceedings of the European Symposium on Artificial Neural Networks (ESANN-2002)*, 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(1–4):85–105, October 2002.

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

[5] 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.

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

[7] 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.

[8] C. M. Bishop and C. Legleye. Estimating conditional probability densities for periodic variables. In *Advances in Neural Information Processing Systems*, volume 7, pages 641–648. MIT Press, 1995.

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

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

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

[12] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines (and other kernel-based learning methods).* Cambridge University Press, Cambridge, U.K., 2000.

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

[14] J. Suykens, L. Lukas, and J. Vandewalle. Sparse approximation using least-squares support vector machines. In *Proc., IEEE Int. Symposium on Circuits and Systems*, pages 11757–11760, Geneva, Switzerland, May 2000.

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

[16] G. C. Cawley, N. L. C. Talbot, R. J. Foxall, S. R. Dorling, and D. P. Mandic. Unbiased estimation of conditional variance in heteroscedastic kernel ridge regression. In *Proceedings of the European Symposium on Artificial Neural Networks (ESANN-2003)*, pages 209–214, Bruges, Belgium, April 23–25 2003.

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

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

[19] 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.

[20] J. A. Nelder and R. Mead. A simplex method for function minimization. *Computer Journal*, 7:308–313, 1965.

[21] B. Schölkopf, R. Herbrich, and A. J. Smola. A generalised representer theorem. In *Proceedings of the Fourteenth International Conference on Computational Learning Theory*, pages 416–426, Amsterdam, the Netherlands, July 16–19 2001.

[22] B. Schölkopf and A. J. Smola. Learning with kernels: Support vector machines, regularization, optimization and beyond. MIT Press, 2002.

[23] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

A The Generalised Representer Theorem

The representer theorem is a result arising from the field of approximation theory [13] useful in the construction of kernel learning methods. Here we provide a slightly simplified version of a generalised representer theorem due to Schölkopf *et al.* [21] sufficient for purposes of the work described here. For a basic introduction to the concepts of reproducing kernel Hilbert spaces, see the books by Cristianini and Shawe-Taylor [12] or Schölkopf and Smola [22].

Theorem 1 (Generalised Representer Theorem [21])

Let $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{\ell},\ \boldsymbol{x}_i \in \mathcal{X} \subset \mathbb{R}^d, y_i \in \mathbb{R}$ *represent the training data and* $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *the reproducing kernel generating a reproducing kernel Hilbert space (RHKS) [23] of functions* $\mathcal{H}$*. We wish to select a function,* f*, from* $\mathcal{H}$ *providing the solution to the primal optimisation problem,* P*:*

$$\min_{f \in \mathcal{H}} \left\{ \sum_{i=1}^{\ell} C(y_i, f(\boldsymbol{x}_i)) + \Omega(\|f\|_{\mathcal{H}}^2) \right\},$$

where $C(\cdot, \cdot)$ *is a convex loss function,* $\Omega(\cdot)$ *is a strictly increasing function and* $\|f\|_{\mathcal{H}}$ *represents the norm of* f *measured in the RKHS* $\mathcal{H}$*. The minimiser of* P *then admits a solution of the form*

$$f(\cdot) = \sum_{i=1}^{\ell} \alpha_i \mathcal{K}(\cdot, \boldsymbol{x}_i).$$

Proof:

Let $\mathcal{H}_{\mathcal{D}}$ *represent the subspace of* $\mathcal{H}$ *spanned by the functions* $\mathcal{K}(\cdot, \boldsymbol{x}_i),\ \forall\ i \in \{1, 2, \ldots, \ell\}$*, then every* $f \in \mathcal{H}$ *has a unique decomposition in terms of a component within* $H_{\mathcal{D}}$*,* $f_{||}(\cdot)$*, and a component orthogonal to* $H_{\mathcal{D}}$*,* $f_{\perp}(\cdot)$*, i.e.*

$$f(\cdot) = f_{||}(\cdot) + f_{\perp}(\cdot) = \sum_{i=1}^{\ell} \alpha_i \mathcal{K}(\cdot, \boldsymbol{x}_i) + f_{\perp}(\cdot)$$

Via the reproducing property of the RKHS $\mathcal{H}$*,*

$$f(\boldsymbol{x}_j) = \langle f(\cdot), \mathcal{K}(\cdot, \boldsymbol{x}_j) \rangle = \sum_{i=1}^{\ell} \alpha_i \langle \mathcal{K}(\cdot, \boldsymbol{x}_i), \mathcal{K}(\cdot, \boldsymbol{x}_j) \rangle + \langle f_{\perp}(\cdot), \mathcal{K}(\cdot, \boldsymbol{x}_j) \rangle .$$

As $f_{\perp}(\cdot)$ *is orthogonal to* $\mathcal{H}_{\mathcal{D}}$*, i.e.* $\langle f_{\perp}, \mathcal{K}(\cdot, \boldsymbol{x}_i) \rangle = 0,\ \forall\ i \in \{1, 2, \ldots, \ell\}$*, the second term vanishes, and so*

$$f(\boldsymbol{x}_j) = \sum_{i=1}^{\ell} \alpha_i \mathcal{K}(\boldsymbol{x}_j, \boldsymbol{x}_i).$$

The values of f *for points belonging to the training data,* $\mathcal{D}$*, thus depend*

only on the vector of model parameters $\boldsymbol{\alpha}$*. The first term of the optimisation criterion is a point-wise loss function therefore is independent of the orthogonal component, depending only on the value of* f *for points in* $\mathcal{D}$*. Let us now define equivalence classes for functions in* $\mathcal{H}$*, such that* f *and* f' *belong to the same equivalence class if* $f(\boldsymbol{x}_i) = f'(\boldsymbol{x}_i),\ \forall\ i \in \{1, 2, \ldots, \ell\}$*. The second term of the optimisation criterion can be written,*

$$\Omega(\|f\|_{\mathcal{H}}^2) = \Omega\left(\left\|\sum_{i=1}^{\ell} \alpha_i \mathcal{K}(\cdot, \boldsymbol{x}_i)\right\|_{\mathcal{H}}^2 + \|f_\perp\|_{\mathcal{H}}^2\right).$$

Clearly the global optimum of the optimisation problem, P*, will be the member of the optimal equivalence class for which* $\|f_\perp\|_{\mathcal{H}}^2 = 0$*.*

Clearly L^μ and L^σ represent optimisation problems of the form directly covered by the generalised representer theorem, as described above, given that they are each comprised of a convex point-wise loss function together with the usual regularisation term. The combined optimisation criterion, L, also falls into this category except that there are now two regularisation terms instead of one, however it is straight-forward to extended the arguments given above in order to accommodate this.

B On the Convexity Optimisation Problems

For an optimisation problem to be *convex*, the Hessian of the objective function with respect to the unknown variables must be positive semi-definite, i.e. the eigenvalues of the Hessian are non-negative. A simple test to determine whether a matrix is positive semi-definite is given by the following lemma:

Lemma 1 (test for a positive semi-definite matrix)

Let $\boldsymbol{A}$ *be a symmetric matrix, then* $\boldsymbol{A}$ *is positive semi-definite if and only if for any vector* $\boldsymbol{x} \neq \boldsymbol{0}$

$$\boldsymbol{x}\boldsymbol{A}\boldsymbol{x} \geq 0.$$

Kernel learning methods typically seek to minimise an objective function comprised of a linear combination of a loss function, $L_{\mathcal{D}}$, which measures the data misfit and a regularisation term, $L_{\mathcal{R}}$, i.e.

$$L = L_{\mathcal{D}} + \lambda L_{\mathcal{R}}.$$

$L_{\mathcal{R}}$ is normally a quadratic function of the model parameters, and is self-evidently convex. Clearly L then represents a convex optimisation problem provided that $L_{\mathcal{D}}$ is also convex. We must therefore show that the Hessian of $L_{\mathcal{D}}$ with respect to the model parameters is positive definite. In the case of kernel learning methods, however, rather than construct the Hessian with respect to the model parameters, we will show that we need only consider the Hessian with respect to the output of the model (before any non-linear transformation). Kernel learning methods generally implement a model of the form

$$y(\boldsymbol{x}) = f(z) = f\left(\sum_{i=1}^{\ell} \alpha_i \mathcal{K}(\boldsymbol{x}, \boldsymbol{x}_i)\right),$$

for simplicity, but without loss of generality, we will omit the usual bias term. The Hessian of $L_{\mathcal{D}}$ with respect to the model parameters, $\boldsymbol{\alpha} = \{\alpha_1, \alpha_2, \ldots, \alpha_\ell\}$ is then

$$\boldsymbol{H}_{\mathcal{D}} = \left[h_{ij} = \frac{\partial^2 L_{\mathcal{D}}}{\partial \alpha_i \partial \alpha_j}\right]_{i,j=1}^{\ell}.$$

Using the chain rule, we have that

$$\frac{\partial^2 L_{\mathcal{D}}}{\partial \alpha_i \partial \alpha_j} = \sum_{m=1}^{\ell} \sum_{n=1}^{\ell} \frac{\partial^2 L_{\mathcal{D}}}{\partial z_m \partial z_n} \cdot \frac{\partial z_m}{\partial \alpha_i} \cdot \frac{\partial z_n}{\partial \alpha_j},$$

where $z_i = \sum_{j=1}^{\ell} \alpha_j \mathcal{K}(\boldsymbol{x}_j, \boldsymbol{x}_i)$. The Hessian matrix $\boldsymbol{H}_{\mathcal{D}}$ can then be written in the form

$$\boldsymbol{H}_{\mathcal{D}} = \boldsymbol{G}\hat{\boldsymbol{H}}_{\mathcal{D}}\boldsymbol{G}^T,$$

where $\hat{\boldsymbol{H}}_{\mathcal{D}}$ is the Hessian of $L_{\mathcal{D}}$ with respect to the raw output of the kernel machine,

$$\hat{\boldsymbol{H}}_{\mathcal{D}} = \left[\hat{h}_{ij} = \frac{\partial^2 L_{\mathcal{D}}}{\partial z_i \partial z_j}\right]_{i,j=1}^{\ell},$$

and $\boldsymbol{G}$ is a matrix of the partial derivatives of the output of the kernel machine with respect to the parameters, $\boldsymbol{\alpha}$,

$$\boldsymbol{G} = \left[g_{ij} = \frac{\partial z_i}{\partial \alpha_j}\right]_{i,j=1}^{\ell}.$$

The following lemma then shows that L represents a convex optimisation problem provided that $\hat{\boldsymbol{H}}_{\mathcal{D}}$ is positive definite:

Lemma 2 (positive semi-definiteness of $\boldsymbol{B} = \boldsymbol{C}\boldsymbol{A}\boldsymbol{C}^T$)

Let $\boldsymbol{A}$ be a positive semi-definite matrix, then $\boldsymbol{B} = \boldsymbol{C}\boldsymbol{A}\boldsymbol{C}^T$ is also positive semi-definite.

Proof: *If $\boldsymbol{A}$ is positive semi-definite, then for any vector $\boldsymbol{x} \neq \boldsymbol{0}$,*

$$\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} \geq 0.$$

We can then write

$$\boldsymbol{x}^T\boldsymbol{C}\boldsymbol{A}\boldsymbol{C}^T\boldsymbol{x} = (\boldsymbol{C}^T\boldsymbol{x})^T\boldsymbol{A}(\boldsymbol{C}^T\boldsymbol{x}) = \hat{\boldsymbol{x}}^T\boldsymbol{A}\hat{\boldsymbol{x}} \geq 0,$$

since $\boldsymbol{A}$ is already known to be positive semi-definite.

B.1 Convexity of L^μ

The term of the optimisation criterion for the model of the conditional mean that is dependent on the data is given by

$$L^\mu_{\mathcal{D}} = \sum_{i=1}^{\ell} \zeta_i (y_i - z_i)^2$$

where $z_i = \mu(\boldsymbol{x}_i)$ and $\zeta_i^{-1} = 2\sigma^2(\boldsymbol{x}_i)$. The Hessian of $L^\mu_{\mathcal{D}}$ with respect to z_i is then given by

$$\hat{\boldsymbol{H}}^\mu_{\mathcal{D}} = \left[\hat{h}_{ij} = \frac{\partial^2 L^\mu_{\mathcal{D}}}{\partial z_i \partial z_j}\right]_{i,j=1}^{\ell} = \mathrm{diag}(2\boldsymbol{\zeta}),$$

where $\boldsymbol{\zeta} = (\zeta_1, \zeta_2, \ldots, \zeta_\ell)$ and $\mathrm{diag}(\boldsymbol{v})$ represents a diagonal matrix with non-zero elements given by the vector $\boldsymbol{v}$. The elements of $\boldsymbol{\zeta}$ are clearly non-negative, and so the following lemma demonstrates that $\hat{\boldsymbol{H}}^\mu_{\mathcal{D}}$ is positive semi-definite and therefore L^μ represents a convex optimisation problem.

Lemma 3 (positive semi-definiteness of diagonal matrices)

A diagonal matrix $\boldsymbol{A} = \mathrm{diag}(\boldsymbol{a})$ with non-negative diagonal elements, $\boldsymbol{a} = (a_1, a_2, \ldots, a_n)$, is semi-positive definite.

Proof: *For any vector $\boldsymbol{x} \neq \boldsymbol{0}$,*

$$\boldsymbol{x}\boldsymbol{A}\boldsymbol{x}^T = \sum_{i=1}^{n} a_i x_i^2 \geq 0 \quad \text{iff } a_i \geq 0, \ \forall\ i \in \{1, 2, \ldots, n\}.$$

B.2 Convexity of L^σ

The term of the optimisation criterion for the model of the conditional standard deviation that is dependent on the data is given by

$$L^\sigma_{\mathcal{D}} = \sum_{i=1}^{\ell} \left[z_i + \xi_i \exp\{-2z_i\} \right]. \tag{B.1}$$

where $z_i = \log \sigma(\boldsymbol{x}_i)$. and $\xi_i = \frac{1}{2}[\mu(\boldsymbol{x}_i) - y_i]^2$. The Hessian of $L^\sigma_{\mathcal{D}}$ with respect to z_i is then given by

$$\hat{\boldsymbol{H}}^\sigma_{\mathcal{D}} = \left[\hat{h}_{ij} = \frac{\partial^2 L^\sigma_{\mathcal{D}}}{\partial z_i \partial z_j} \right]^{\ell}_{i,j=1} = \mathrm{diag}(4\boldsymbol{\xi}\boldsymbol{\nu}),$$

where $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_\ell)$ and $\log \boldsymbol{\nu} = -2(z_1, z_2, \ldots, z_\ell)$. As the elements of $\boldsymbol{\xi}$ and $\boldsymbol{\nu}$ are non-negative, via lemma 3, we know that $\hat{\boldsymbol{H}}^\sigma_{\mathcal{D}}$ is positive semi-definite and so L^σ also represents a convex optimisation problem.

B.3 Convexity of L

The term of the optimisation criterion for the combined model that is dependent on the data is given by

$$L_{\mathcal{D}} = \sum_{i=1}^{\ell} \left[z_i + \frac{1}{2}(\mu_i - y_i)^2 \exp\{-2z_i\} \right]$$

where $z_i = \log \sigma(\boldsymbol{x}_i)$ and $\mu_i = \mu(\boldsymbol{x}_i)$. In this case the model has two outputs and so the Hessian of $L_{\mathcal{D}}$ with respect to z_i and μ_i is then described by a block matrix

$$\hat{\boldsymbol{H}}_{\mathcal{D}} = \begin{bmatrix} \boldsymbol{H}_1 & \boldsymbol{H}_2 \\ \\ \boldsymbol{H}_3 & \boldsymbol{H}_4 \end{bmatrix}$$

where

$$\boldsymbol{H}_1 = \left[\frac{\partial^2 L_{\mathcal{D}}}{\partial z_i^2}\right]_{i,j=1}^{\ell} = \hat{\boldsymbol{H}}_{\mathcal{D}}^{\sigma},$$
$$\boldsymbol{H}_2 = \left[\frac{\partial^2 L_{\mathcal{D}}}{\partial z_i \partial \mu_j}\right]_{i,j=1}^{\ell} = \boldsymbol{H}_3 = \left[\frac{\partial^2 L_{\mathcal{D}}}{\partial \mu_i \partial z_j}\right]_{i,j=1}^{\ell} = \mathrm{diag}(-2(\mu_i - y_i)\exp(-2z_i)),$$
$$\boldsymbol{H}_4 = \left[\frac{\partial^2 L_{\mathcal{D}}}{\partial \mu_i^2}\right]_{i,j=1}^{\ell} = \hat{\boldsymbol{H}}_{\mathcal{D}}^{\mu}.$$

If $\hat{\boldsymbol{H}}_D$ is positive definite, $\boldsymbol{v}^T \hat{\boldsymbol{H}}_D \boldsymbol{v} > 0$ for any vector $\boldsymbol{v}$. Let $\boldsymbol{v} = [\boldsymbol{v}^{\mu T} \quad \boldsymbol{v}^{\sigma T}]^T$, then using block matrix algebra

$$\boldsymbol{v}^T \hat{\boldsymbol{H}}_D \boldsymbol{v} = \boldsymbol{v}^{\mu T} \boldsymbol{H}_1 \boldsymbol{v}^{\mu} + \boldsymbol{v}^{\sigma T} \boldsymbol{H}_3 \boldsymbol{v}^{\mu} + \boldsymbol{v}^{\mu T} \boldsymbol{H}_2 \boldsymbol{v}^{\sigma} + \boldsymbol{v}^{\sigma T} \boldsymbol{H}_4 \boldsymbol{v}^{\sigma}$$

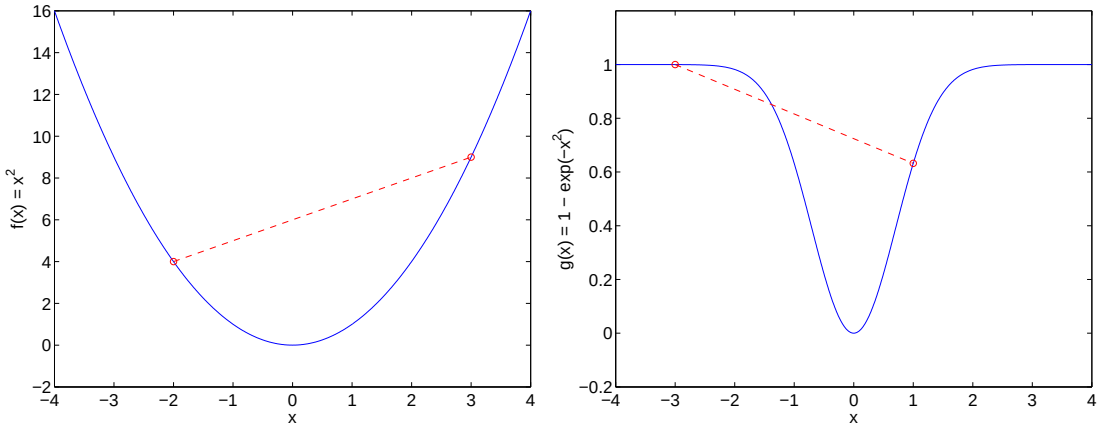
since $\boldsymbol{H}_1, \ldots, \boldsymbol{H}_4$ are diagonal this can be expanded to

$$= \sum_{i=1}^{\ell} \exp(-2z_i)[(v_i^{\mu})^2 + 2(v_i^{\sigma})^2(\mu_i - y_i)^2 - 4v_i^{\sigma} v_i^{\mu}(\mu_i - y_i)]$$

Note that if $v_i^{\mu} = v_i^{\sigma}(\mu_i - y_i)$ $i = 1, \ldots, \ell$ then

$$= \sum_{i=1}^{\ell} \exp(-2z_i)[(v_i^{\sigma}(\mu_i - y_i))^2 + 2(v_i^{\sigma})^2(\mu_i - y_i)^2 - 4v_i^{\sigma} v_i^{\sigma}(\mu_i - y_i)(\mu_i - y_i)]$$
$$= -\sum_{i=1}^{\ell} \exp(-2z_i)(v_i^{\sigma})^2(\mu_i - y_i)^2$$

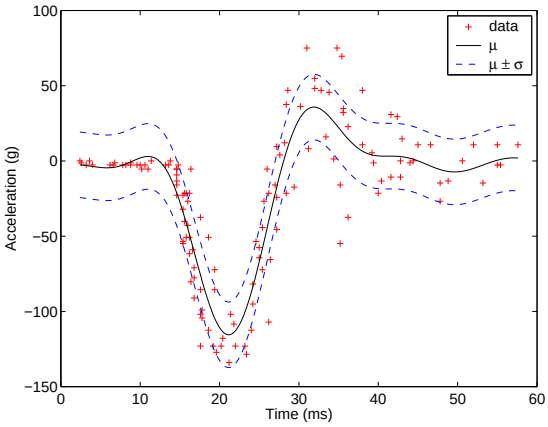
This is negative, since the elements of the summation are non-negative. Therefore, as there exists a vector $\boldsymbol{v}$ such that $\boldsymbol{v}^T \hat{\boldsymbol{H}}_D \boldsymbol{v} < 0$, $\hat{\boldsymbol{H}}_D$ is not positive definite and L corresponds to a non-convex optimisation problem.



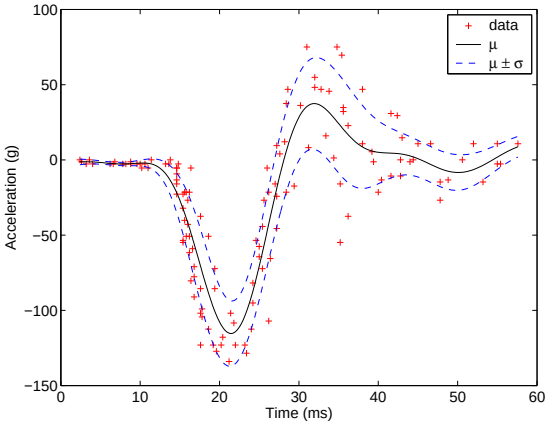
(a) (b)

Fig. 1. A straight line joining any two points on a convex function does not cut the graph of that function other than at the end-points, as shown in (a). The function shown in (b) is clearly non-convex, note however that although $g(x)$ is non-convex it is nevertheless unimodal.

(a)



(b)



(c)

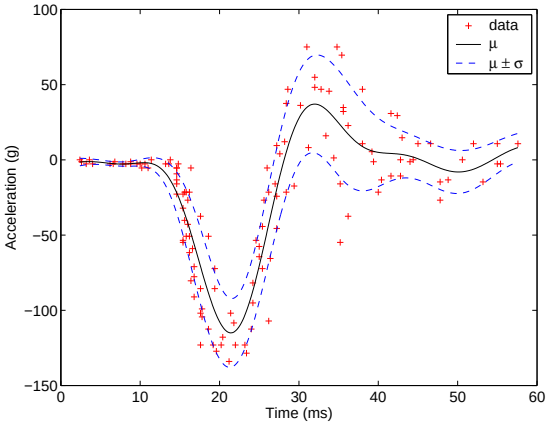


Fig. 2. Homoscedastic (a) heteroscedastic (b) and leave-one-out heteroscedastic (c) kernel ridge regression models of the Motorcycle benchmark dataset.

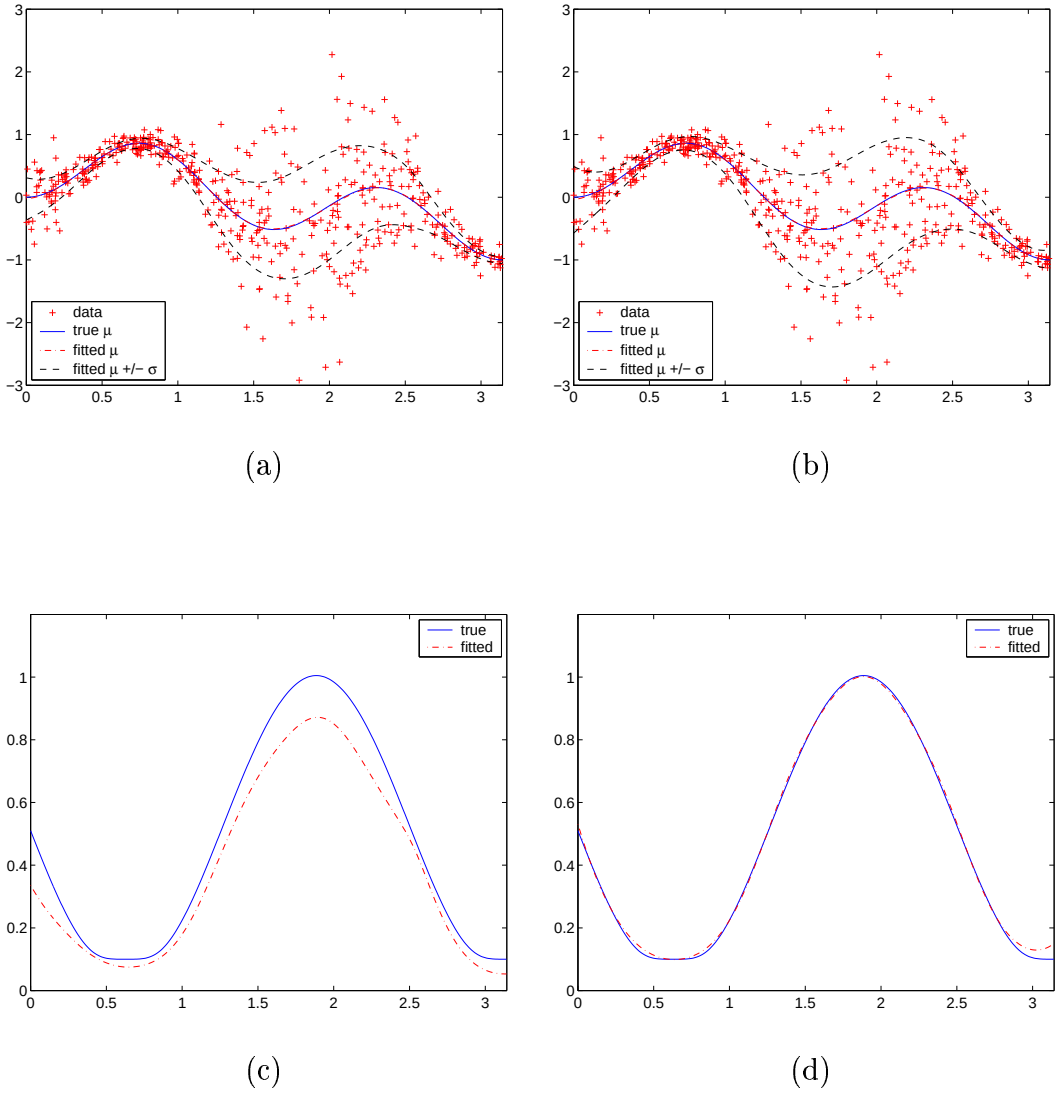


Fig. 3. Arithmetic mean of the estimate of the conditional mean and ± one standard deviation credible interval for (a) simple heteroscedastic kernel ridge regression (HKRR) and (b) leave-one-out heteroscedastic kernel ridge regression (LOOHKRR) models for a synthetic regression problem, (c) and (d) display the corresponding means of the estimated conditional standard deviation for the HKRR and LOOHKRR models respectively. All graphs show average results computed over 1000 randomly generated datasets (see text for details).

Table B.1

Kernel and regularisation parameters employed for the Motorcycle benchmark dataset.

Model	λ^{μ}	γ^{μ}	λ^{σ}	γ^{σ}
KRR	6.776	1.337	-	-
HKRR	8.705	5.68×10^{-4}	6.762	2.776
LOOHKRR	8.137	5.91×10^{-4}	7.736	1.487

Table B.2

Leave-one-out cross-validation estimates of the sum-of-squares error and negative log-likelihood for kernel ridge-regression models of the Motorcycle benchmark dataset.

Model	**SSE**	$-\log \mathcal{L}_{\mathcal{D}}$
KRR	71702.2	487.262
HKRR	71922.6	440.221
LOOHKRR	71528.0	436.585

Table B.3

Comparison of conventional, heteroscedastic and leave-one-out heteroscedastic kernel ridge regression models (KRR, HKRR and LOOHKRR respectively) for prediction of daily mean SO_2 concentration in urban Belfast.

Model	**SSE**	$-\log \mathcal{L}_{\mathcal{D}}$	**X-ENT**
KRR	15.81	517.8	6.57
HKRR	14.92	2313	3.48
LOOHKRR	14.92	1581	3.47