Interpolation Models with Multiple Hyperparameters

David J C MacKay * Ryo Takeuchi †

January 3, 1997

Abstract

A traditional interpolation model is characterized by the choice of regularizer applied to the interpolant, and the choice of noise model. Typically, the regularizer has a single regularization constant α, and the noise model has a single parameter β. The ratio α/β alone is responsible for determining globally all these attributes of the interpolant: its 'complexity', 'flexibility', 'smoothness', 'characteristic scale length', and 'characteristic amplitude'. We suggest that interpolation models should be able to capture more than just one flavour of simplicity and complexity. We describe Bayesian models in which the interpolant has a smoothness that varies spatially. We emphasize the importance, in practical implementation, of the concept of 'conditional convexity' when designing models with many hyperparameters.

We apply the new models to the interpolation of neuronal spike data and demonstrate a substantial improvement in generalization error.

*David MacKay is with the Cavendish Laboratory, Cambridge, United Kingdom. Email: `mackay@mrao.cam.ac.uk`.

†Ryo Takeuchi is with the Electrical Engineering department, Waseda University, Tokyo, Japan. Email: `takeuchi@matsumoto.elec.waseda.ac.jp`.

1 Introduction

In this paper our philosophy of generalization is as follows: the best generalization will be achieved by a Bayesian model that is well-matched to the problem and that is accurately implemented. The aim of obtaining the best generalization is thus subsumed under the aim of searching for good models. In this paper we expand the space of interpolation models by introducing additional hyperparameters, and demonstrate that the generalization performance on a real problem is substantially improved.

A traditional linear interpolation model '$\mathcal{H}_1$' is characterized by the choice of the regularizer $\mathcal{R}$, or prior probability distribution, that is applied to the interpolant; and the choice of noise model $\mathcal{N}$. The choice of basis functions $\mathcal{A}$ used to represent the interpolant may also be important if only a small number of basis functions are used. Typically the regularizer is a quadratic functional of the interpolant and has a single associated regularization constant α, and the noise model is also quadratic and has a single parameter β. For example, the splines prior for the function $y(x)$ (Kimeldorf and Wahba 1970) is:[1]

$$\log P(y(x)|\alpha, \mathcal{H}_1) = -\frac{1}{2}\,\alpha \int dx\, [y^{(p)}(x)]^2 + \text{const}, \tag{1}$$

where $y^{(p)}$ denotes the pth derivative of y. The probability of the data measurements $D = \{t^{(m)}\}_{m=1}^{N}$ assuming independent Gaussian noise is:

$$\log P\,(\,D|\,y(x), \beta, \mathcal{H}_1) = -\frac{1}{2}\,\beta \sum_{m=1}^{N} (y(x^{(m)}) - t^{(m)})^2 + \text{const}. \tag{2}$$

(The constants in equations (1) and (2) are functions of α and β respectively.) When we use these distributions with $p = 2$ and find the most probable $y(x)$ we obtain the cubic splines interpolant. For any quadratic regularizer and quadratic

[1] Strictly this prior is improper since addition of an arbitrary polynomial of degree $p-1$ to $y(x)$ is not constrained. It can be made proper by adding terms corresponding to boundary conditions to (1). In the present implementations of the models, we enforce the boundary conditions $y(0) = 0$ and, where appropriate, $y'(0) = 0$.

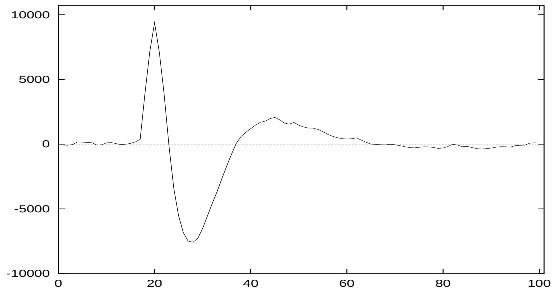


Figure 1: An inferred spike signal from a zebra finch neuron.
Courtesy of M. Lewicki and A. Doupe, California Institute of Technology.

log likelihood, the most probable interpolant depends linearly on the data values. This is the property by which we define a 'linear' interpolation model.

Such models may be optimized and compared using Bayesian methods as reviewed in MacKay (1992). In such models, for fixed β, the ratio α/β alone determines globally all the following attributes of the interpolant: its complexity, flexibility, smoothness, characteristic scale length, and characteristic amplitude. Now, whilst some of these terms may be synonyms, surely others describe distinct properties. Should not our models be able to capture more than just one flavour of simplicity and complexity? And should not the interpolant's smoothness, for example, be able to vary spatially?

1.1 Example: Neural spike modelling

An example of a function from a real system is shown in figure 1; this is the action potential of a neuron deduced from recordings of 40 distinct events (Lewicki 1994). The graph was created by fitting a simple spline model (with $p = 1$) to the data. This function has one 'spiky' region with large characteristic amplitude and short spatial scale. Elsewhere the true function is smooth. However the fitted function shown in figure 1, controlled by only one regularization constant

α, overfits the noise on the right, having a rough appearance where it should plausibly be smooth. The value of α appropriate for fitting the spiky region is too small for the rest of the curve. It would be useful here to have a model capable of capturing the concepts of local smoothness, because such a model, having a prior better matched to the real world, would require less data to yield information of the same quality. Furthermore, when different hypotheses are compared, broad priors introduce a bias toward simpler hypotheses. For example, if we ask whether one or two distinct spike functions are present in a data set, the traditional model's prior with small α will bias the conclusion in favour of the single spike function. Only with well-matched priors can the results of Bayesian hypothesis comparison be trusted.

In this paper we discuss methods for introducing multiple flavours of simplicity and complexity into a hierarchical probabilistic model in a computationally tractable way, and demonstrate new interpolation models with multiple hyperparameters that capture a spatially varying smoothness.

Prior work making use of variable hyperparameters includes the modelling of data with non-Gaussian innovations or observation noise (see, *e.g.*, (West 1984; Carter and Kohn 1994; Shephard 1994)). The interpolation models we propose might be viewed as Bayesian versions of the 'variable bandwidth' kernel regression technique (Muller and Stadtmuller 1987). The aim of our new model is also similar to the goal of inferring the locations of discontinuities in a function, studied by Blake and Zisserman (1987). Traditional interpolation models have difficulty with discontinuities: if the value of α/β is set high, then edges are blurred out in the model; if α/β is lowered the edge is captured, but ringing appears near the edge, and noise is overfitted everywhere. Blake and Zisserman introduce additional hyperparameters defining the locations of edges. The models they use are computationally non-convex, so that finding good representatives of the posterior distribution is challenging. They use 'graduated non-convexity' techniques to find good solutions. By contrast we attempt to create new hierarchical models that are, for practical purposes, convex.

2 Tractable hierarchical modelling: Convexity

Bayesian statistical inference is often implemented either by Gaussian approximations about modes of distributions, or by Markov Chain Monte Carlo methods (Smith 1991). Both methods clearly have a better chance of success if the posterior probability distribution over the model parameters and hyperparameters is not dominated by multiple distinct optima. If we know that most of the probability mass is in just one 'hump', then we know that we need not engage in a time-consuming search for the more probable optima, and we might hope that some approximating distribution (*e.g.*, involving the mode of the distribution) might be able to capture the key properties of that hump. Furthermore, convex conditional distributions may be easier to sample from with, say, Gibbs sampling methods (Gilks and Wild 1992). It would be useful if all the conditional and marginal probability distributions of our models were **log convex**:

Definition 1 *A probability distribution is* **log convex** *if there is a representation* $\mathbf{x}$ *of the variables such that the matrix* $\mathbf{M}$ *defined by*

$$M_{ij} = -\frac{\partial^2}{\partial x_i \partial x_j} \log P(\mathbf{x}) \tag{3}$$

is everywhere positive definite.

It is hard, however, to make interesting hierarchical models such that all conditional and marginal distributions are log convex. We introduce a weaker criterion:

Definition 2 *A model is* **conditionally convex** *if its variables can be divided into groups such that, for every group, their distribution conditioned on any values for the other variables is log convex.*

An example of a conditionally convex model is the traditional interpolation model with three groups of variables: D (data), $\mathbf{w}$ (parameters), and α (one hyperparameter). The probability distribution $P(D|\mathbf{w},\alpha) = P(D|\mathbf{w})$ is log convex over D (it is Gaussian). The distribution $P(\mathbf{w}|D,\alpha)$ is log convex over $\mathbf{w}$ (it is Gaussian). And the distribution $P(\alpha|\mathbf{w},D) = P(\alpha|\mathbf{w})$ is log convex over α (it is a Gamma distribution).

That a model is conditionally convex does not guarantee that *marginal* distributions of all variables are unimodal. For example the traditional model's posterior marginals $P(\mathbf{w}|D)$ and $P(\alpha|D)$ are not necessarily unimodal; but good unimodal approximations to them can often be made (MacKay 1996). So we conjecture that conditional convexity is a desirable property for a tractable model.

We now generalize the spline model of equation (1) to a model with multiple hyperparameters that is conditionally convex, and demonstrate it on the neural spike data. We then discuss general principles for hierarchical modelling with multiple hyperparameters.

3 A new interpolation model

We replace the regularizer of equation (1) by:

$$\log P(y(x)|\alpha(x),\mathcal{H}_2) = -\frac{1}{2}\int dx\,\alpha(x)[y^{(p)}(x)]^2 + \text{const}, \tag{4}$$

where $\alpha(x)$ is written in terms of hyperparameters $\mathbf{u} = \{u_h\}$ thus:

$$\alpha(x;\mathbf{u}) = \exp\left(\sum_{h=1}^{H} u_h \psi_h(x)\right), \tag{5}$$

and the constant of equation (4) is a function of $\alpha(x;\mathbf{u})$ which becomes important when $\alpha(x;\mathbf{u})$ is inferred. The exponentiated quantity has the form of a linear interpolant using basis functions $\psi_h(x)$. In the special case $H = 1, \psi_1(x) = \text{const.}$, we obtain the traditional single alpha model. This representation is chosen because (1) it embodies our prior belief that $\alpha(x)$ should be a smooth function of x, and (2) the model is conditionally convex (a partial proof is given in section 4).

When implementing this model we optimize the hyperparameters $\mathbf{u}$ and β by maximizing the marginal likelihood or 'evidence',

$$P\left(D|\mathbf{u},\beta,\mathcal{H}_2\right) = \int d^k\mathbf{y}\; P\left(D|\mathbf{y},\beta,\mathcal{H}_2\right) P\left(\mathbf{y}|\mathbf{u},\mathcal{H}_2\right), \tag{6}$$

where k is the dimensionality of our representation $\mathbf{y}$ of $y(x)$. Some authors view this 'empirical Bayes' approach as controversial and inaccurate (Wolpert 1993), but it is widely used under various names such as 'ML-II', and is closely related to 'generalized maximum likelihood' (Gu and Wahba 1991). The ideal Bayesian method would put a proper prior on the hyperparameters and marginalize over them, but optimization of the hyperparameters is computationally more convenient and often gives predictive distributions that are indistinguishable (MacKay 1996).

We use a discrete representation of $y(x)$ and $\alpha(x)$ on a finely spaced grid, $\{x_c\}$, writing $y(x) \rightarrow \mathbf{y}$, $\alpha(x;\mathbf{u}) \rightarrow \{\alpha_c \equiv \alpha(x_c;\mathbf{u})\}$ and $\psi_{hc} \equiv \psi_h(x_c)$. In this representation the Hessian of the log posterior is a sum of band-diagonal terms from the log prior and a diagonal matrix from the log likelihood, $\mathbf{A} \equiv -\nabla\nabla \log P(\mathbf{y}|D,\{\alpha\},\beta,\mathcal{H}_2) = \sum_{c=1}^{C} \alpha_c \mathbf{C}_c + \beta\mathbf{I}$. The gradient of the log evidence, which we use for the optimization, is then:

$$\frac{\partial}{\partial u_h} \log P\left(D|\mathbf{u},\beta,\mathcal{H}_2\right) = \sum_{c=1}^{C} \psi_{hc}\alpha_c \frac{\partial}{\partial\alpha_c} \log P(D|\{\alpha_c\}) \tag{7}$$

where

$$\begin{aligned} \frac{\partial}{\partial\alpha_c} \log P(D|\{\alpha_c\}) \;&=\; -\frac{1}{2}\mathbf{y}_{\rm MP}^{\rm T}\mathbf{C}_c\mathbf{y}_{\rm MP} - \frac{1}{2}{\rm Trace}[\mathbf{A}^{-1}\mathbf{C}_c] \\ &\quad +\frac{1}{2}{\rm Trace}\left[\left(\textstyle\sum_{c'} \alpha_{c'}\mathbf{C}_{c'}\right)^{-1}\mathbf{C}_c\right]. \end{aligned} \tag{8}$$

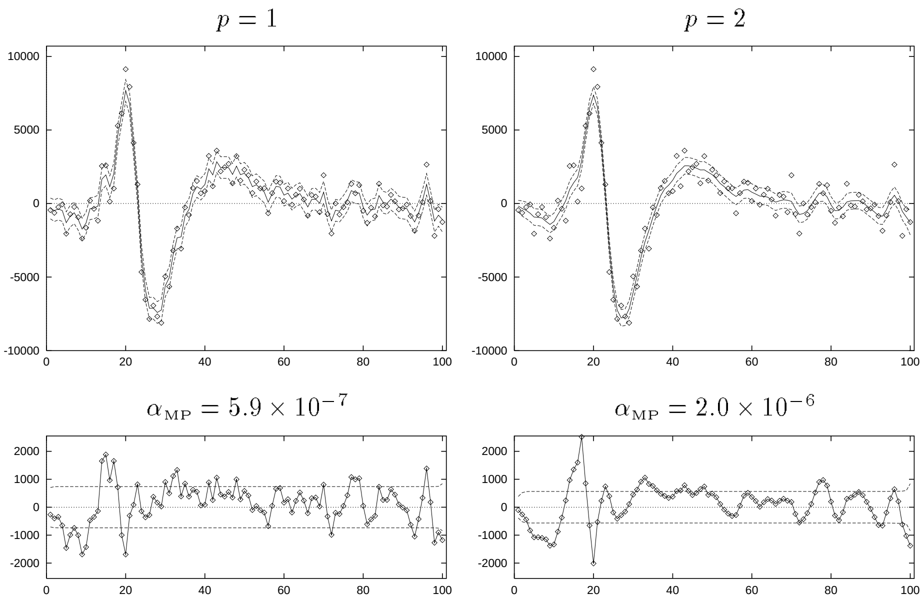


Figure 2: Traditional models: $p = 1$ and $p = 2$. The diamond-shaped points in the upper plots are the artifical data. The solid line shows the most probable interpolant found using the traditional single alpha model. The predictive error bars (dotted lines) are one-standard-deviation error bars. The lower row shows the errors between the interpolant and the original function to which the noise was added to make the artificial data. The predictive error bars are also shown. Contrast with figure 3.

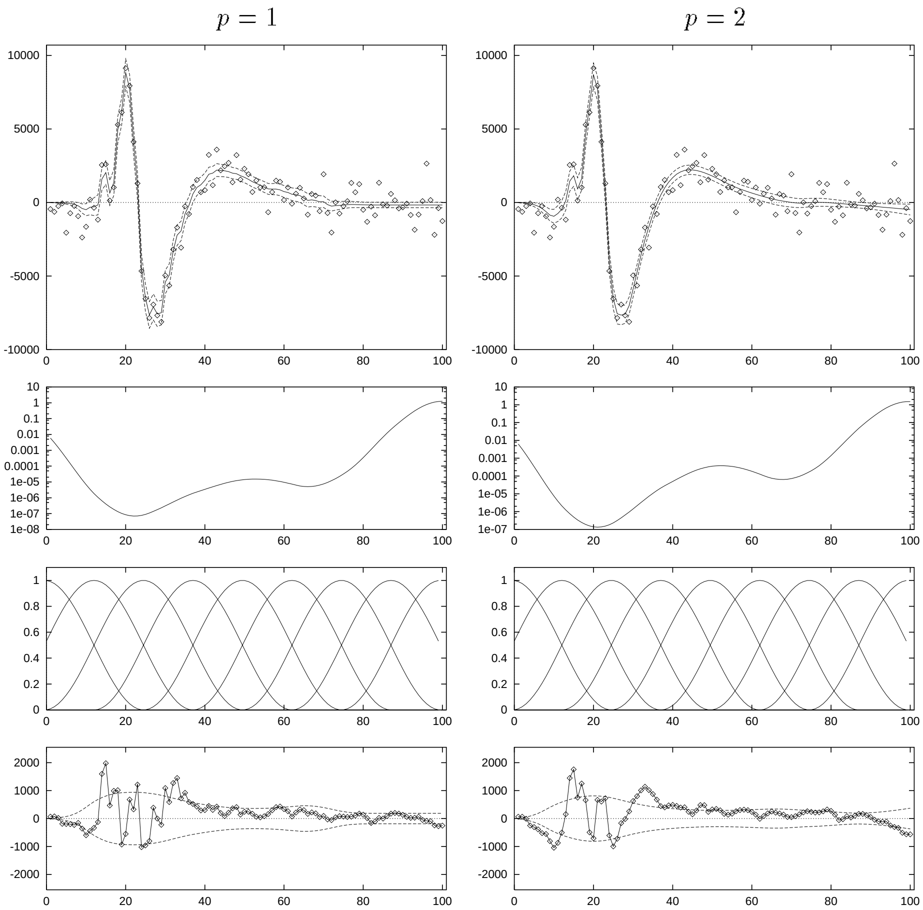


Figure 3: New models with multiple hyperparameters: $p = 1$ and $p = 2$. Top row: The diamond-shaped points are the artifical data. The solid line shows the most probable interpolant and the predictive error bars (dotted lines) are one-standard-deviation error bars. Second row: the inferred $\alpha(x)$ on a log scale (contrast with the values of 5.9×10^{-7} and 2.0×10^{-6} inferred for the traditional models). The third row shows the nine basis functions ψ used to represent $\alpha(x)$. The bottom row shows the errors between the interpolant and the original function to which the noise was added to make the artificial data. The predictive error bars are also shown. The top and bottom graphs should be compared with those of figure 2.

Table 1: Comparison of models on artificial data.

The first three columns give the evidence, the effective number of parameters, and the RMS error for each model when applied to the data shown in figures 2-3. The fourth column gives the RMS error averaged over four similar data sets.

Model	log Evidence	γ	RMS error	avg. RMS error
Trad., $p = 1$	-886.0	54.7	730	694
Trad., $p = 2$	-891.7	32.2	692	642
New, $p = 1$	-859.2	23.6	509	470
New, $p = 2$	-861.5	15.3	510	417

3.1 Demonstration

We made an artificial data set by adding Gaussian noise of standard deviation 1000 to the function depicted in figure 1. [This function plays the role, in these experiments, of a true underlying function; the presence of some actual roughness in this function is believed to be unimportant since our chosen noise level is substantially greater than the apparent size of the roughness.] Figure 2 shows the data, interpolated using the traditional single alpha models with $p = 1$ and $p = 2$. The hyperparameter α was optimized by maximizing the evidence, as in Lewicki (1994). The noise level σ_ν was set to the known noise level. In order for the spiky part of the data to be fitted, α has to be set to a small value, and the most probable interpolant is able in both models to go very close to all the data points. There is considerable overfitting everywhere, and the predictive error bars are large everywhere.

We then interpolated the data with two new models defined by equations (4) and (5), with $p = 1$ and $p = 2$. We set the basis functions ψ to the hump-shaped functions shown in figure 3. These functions define a scale length on which the smoothness is permitted to vary. This scale length was optimized roughly by maximizing the evidence. The new models had nine hyperparameters $\mathbf{u}$. These hyperparameters were set by maximizing the evidence using conjugate gradients. Because the new models are conditionally convex, we had hoped that the maximization of the evidence would lead to a unique optimum $\mathbf{u}_{\rm MP}$. However, there were multiple optima in the evidence as a function of the hyperparameters; but these did not cause insurmountable problems. We found different optima by using different initial conditions $\mathbf{u}$ for the optimization. The best evidence optima were found by initializing $\mathbf{u}$ in a way that corresponded to our prior knowledge that neuronal spike functions start and end with a smooth region; we set $\mathbf{u}$ initially to $\{u_h\} = \{0, -10, -10, -10, -10, -10, -10, 0, 0\}$. This prior knowledge was not formulated into an informative prior over $\mathbf{u}$ during the optimization, though doing so would probably be a good idea for practical purposes.

Figure 3 shows the solutions found using the new interpolation models with $p = 1$ and $p = 2$. The inferred value of α is small in the region of the spike, but elsewhere a larger value of α is inferred, and the interpolant is correspondingly smoother.

The log evidence for the four models is shown in table 1. The reported evidence values are $\log_e P(D|\alpha_{\rm MP}, \mathcal{H}_1)$, $\log_e P(D|\mathbf{u}_{\rm MP}, \mathcal{H}_2)$. If we were to make a proper model comparison we would integrate over the hyperparameters; this integration would introduce additional small subjective Occam factors penalizing the extra hyperparameters in $\mathcal{H}_2$, c.f. MacKay (1992). The root mean square errors between the interpolant and the original function to which the noise was added to make the artificial data are given in table 1, and the errors themselves are displayed at bottoms of figures 2–3.

By both the evidence value and the RMS error values, the new models are significantly superior to the traditional model. Table 1 also displays the value of the 'effective number of well-determined parameters' (Gull 1989; MacKay 1992),

γ, which, when the hyperparameters are optimized, is given by:

$$\gamma = \int dx\, \alpha(x) y^{(p)}(x)^2. \tag{9}$$

The smaller the effective number of parameters, the less overfitting of noise there is, and the smaller the error bars on the interpolant become. The total number of parameters used to represent the interpolant was in all cases 100.

3.2 Model criticism

It is interesting to assess whether the observed errors with respect to the original function are compatible with the one-standard-deviation error bars that were obtained. These are shown together at the bottom of figure 3. The errors are only significantly larger than the error bars at the leftmost five data points, where the small amount of noise in the original function is incompatible with the assumed boundary conditions $y(0) = 0$ and $y'(0) = 0$. Omitting those five data points, we find for the new $p = 1$ model that the other 95 errors have $\chi^2 = 72.5$ (c.f. expectation 95 ± 14), and for the $p = 2$ model, $\chi^2 = 122$. None of the 95 errors in either case exceed 2.5 standard deviations. We therefore see no significant evidence for the observed errors to be incompatible with the predictive error bars.

3.3 Discussion

These new models offer two practical benefits. First, while the new models still fit the spiky region well (indeed the errors are slightly reduced there), they give a smoother interpolant elsewhere. This reduction in overfitting allows more information to be extracted from any given quantity of experimental data; neuronal spikes will be distinguishable given fewer samples. To quantify the potential savings in data we fitted the four models to fake data equivalent to $1, 2, \ldots 10$ independent observations of the function shown in figure 1, that is, $N = 100, 200, \ldots 1000$ data points with noise level $\sigma_\nu = 1000$ (we we did this

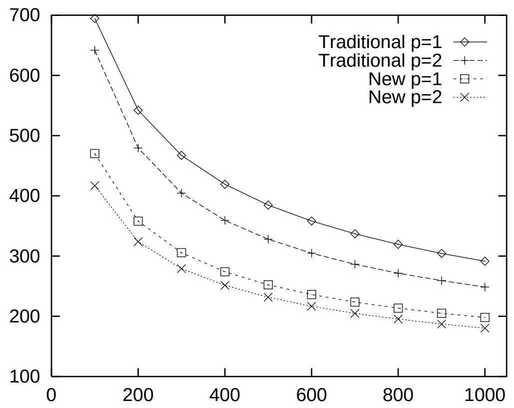


Figure 4: Average RMS error of the traditional and new models as a function of amount of data

by decreasing the actual noise level in the artificial data). The figures and tables shown thus far correspond to the case of one observation, $N = 100$. In figure 4 we show the RMS error of each model as a function of the number of data points, averaged over four runs with different artificial noise. To achieve the same performance (RMS error) as the new models, the traditional models require about three times as much data.

Second, the new models have greater values of the evidence. This does not only mean that they are more probable models (assuming that the omitted Occam factors for the hyperparameters are smaller than these evidence differences). It also means that model comparison questions can be answered in a more reliable way. For example, if we wish to ask 'are two distinct spike types present in several data sets or just one?' then we must compare two hypotheses: $\mathcal{H}_B$, which explains the data in terms of two spike functions, and $\mathcal{H}_A$, which just uses one function. In such model comparisons, the 'Occam factors' that penalize the extra parameters of $\mathcal{H}_B$ are important. If we used the traditional interpolation model, we would obtain Occam factors about e^{20} bigger than those obtained using the new interpolation model. Broad priors bias model comparisons toward simpler models. The new interpolation model, when optimized, produces a prior in which the effective number of degrees of freedom of the interpolant is reduced so that the prior is less broad.

Of course, inference is open-ended, and we expect that these models will in turn be superceded by even better ones. Close inspection of figure 3 reveals that the smoothness assumption on the regularizer may be imperfect — we know from prior experience that the true function's spikiness is confined to a very small time interval, but the new model gives a jagged interpolant in the time interval before the spike too because the function $\alpha(x)$ is assumed to vary smoothly. Future models might include a continuum of alternative values of p (non-integer values of p can be implemented in a Fourier representation). It might also make sense for the characteristic length scale of the basis functions ψ with which $\alpha(x)$ is represented to be shorter where α is small.

The advantages conferred by the new models are not accompanied by a significant increase in computational cost. The optimization of the hyperparameters requires that the Hessian matrix be inverted a small number of times.

Other approaches to the implementation of models with multiple hyperparameters could be considered. The confidence intervals in the present approach, in which the hyperparameters are optimized, are likely to be too small. One could use Markov chain Monte Carlo methods such as Gibbs sampling or hybrid Monte Carlo, both of which would involve a similar computational load (see Neal (1993) for an excellent review). We have used the Gibbs sampling software 'BUGS' (Thomas *et al.* 1992) to implement a similar interpolation model in which the Gaussian noise level is a spatially varying function $\beta(x)$ (MacKay 1995).

4 Some Generalizations

4.1 Strategies for making models with multiple hyperparameters

We now discuss more generally the construction of hierarchical models with multiple hyperparameters.

Consider a Gaussian prior on some parameters $\mathbf{w}$, equivalent to the function $y(x)$ in the earlier example. There are various ways of defining a model with multiple hyperparameters such that each hyperparameter controls a different flavour of simplicity or complexity in $\mathbf{w}$.

Sum Model

Firstly, one might define the inverse covariance matrix as a sum:

$$P(\mathbf{w}|\{\alpha\}) = \frac{1}{Z}\exp\left(-\frac{1}{2}\sum_{c=1}^{C}\alpha_c\mathbf{w}^{\mathrm{T}}\mathbf{C}_c\mathbf{w}\right), \tag{10}$$

where $\{\mathbf{C}_c\}$ are arbitrary positive semi-definite matrices and $\alpha_c \geq 0, \forall c$.

Covariance Sum Model

Secondly, one might define the covariance matrix as a sum:

$$P(\mathbf{w}|\{\theta\}) = \frac{1}{Z}\exp\left(-\frac{1}{2}\mathbf{w}^{\mathrm{T}}\left[\sum_{c=1}^{C}\theta_c\mathbf{C}_c\right]^{-1}\mathbf{w}\right), \tag{11}$$

with hyperparameters $\theta_c \geq 0, \forall c$.

Exponential Sum Model

Thirdly, we can take a sum model of the form (10) (though not necessarily using the same matrices $\{\mathbf{C}_c\}$) and rewrite the coefficients as an exponential sum:

$$\alpha_c = \exp\left(\sum_h u_h\psi_{hc}\right), \tag{12}$$

with hyperparameters $u_h \in (-\infty, \infty)$, so that

$$P(\mathbf{w}|\{u\}) = \frac{1}{Z} \exp\left(-\frac{1}{2}\sum_{c=1}^{C} \exp\left(\sum_h u_h \psi_{hc}\right) \mathbf{w}^{\mathrm{T}}\mathbf{C}_c\mathbf{w}\right), \tag{13}$$

These models have very different capabilities.

The sum model implements the paradigm of starting from a flexible distribution, then adding in extra terms $\alpha_c \mathbf{C}_c$ so as to kill degrees of freedom. This model has no way of introducing selective flexibility. If one hyperparameter α_c is large, there is no way that other hyperparameters can be set to undo the stiffness introduced.

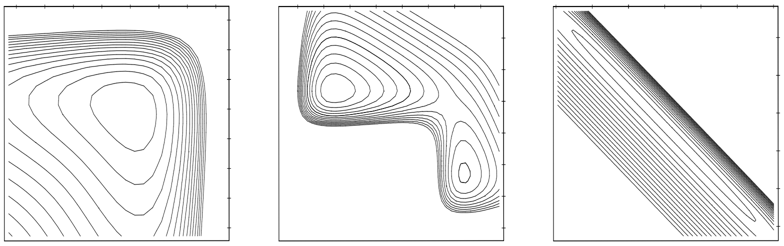
The covariance sum model uses an alternative paradigm, starting from a stiff distribution, and introducing *lacunae of flexibility* into it.

The important difference between these two paradigms is that whereas the sum model is conditionally convex, the covariance sum model is not; it is possible for there to be multiple optima over the hyperparameters even in the limit of perfect data. This will be demonstrated and explained subsequently.

The exponential sum model, of which the interpolation model of section 3 is an example, is intended to combine the best of both worlds. Consider the case where the matrix elements ψ_{ch} are non–negative. As one hyperparameter u_h is increased, it introduces selective stiffness, and as it is decreased, it introduces selective flexibility. The model, being a reparameterization of the sum model, is still conditionally convex (as long as ψ does not have pathological properties).

4.2 Convexity of the sum model

We give a partial proof of conditional convexity for the sum model. It is straightforward to confirm that the conditional distributions $P(D|\mathbf{w}, \{\alpha\})$ and $P(\mathbf{w}|D, \{\alpha\})$ are log convex. The non-trivial property is that $P(\{\alpha\}|\mathbf{w}, D) \propto P(\{\alpha\})P(\mathbf{w}|\{\alpha\})$ is convex. We assume that the prior over $\{\alpha\}$ is defined to be



a) Sum model b) Covariance sum c) Exponential sum

Figure 5: Toy problem probability contours.

Each figure shows the likelihood of two hyperparameters given $\mathbf{w} = (0.01, 2.0, 0.1)$. Hyperparameters α_a, c_a and u_a are on the horizontal axes, and α_b, c_b and u_b on the vertical axes. In all figures the top e^8 of the function is shown and the contours are equally spaced in log probability.

convex and examine the second factor. Defining $\mathbf{M} = \sum_{c=1}^{C} \alpha_c \mathbf{C}_c$, we find:

$$\frac{\partial^2}{\partial\alpha_c \partial\alpha_d} \log P(\mathbf{w}|\{\alpha\}) = -\frac{1}{2}\mathrm{Trace}\left[\mathbf{M}^{-1}\mathbf{C}_c\mathbf{M}^{-1}\mathbf{C}_d\right] \tag{14}$$

This second derivative is negative definite.

Proof

For arbitrary $\mathbf{x}$,

$$\sum_{c,d} x_c \mathrm{Trace}[\mathbf{M}\mathbf{C}_c\mathbf{M}\mathbf{C}_d] x_d \tag{15}$$

$$= \mathrm{Trace}\left[\left(\sum_d \mathbf{M}^{1/2}\mathbf{C}_d\mathbf{M}^{1/2} x_d\right)^2\right] > 0. \tag{16}$$

4.3 A toy illustration

As an illustration, we examine the conditional convexity of a model that assigns a zero-mean Gaussian distribution to a three component vector $\mathbf{w}$. This distribution is to be parameterized by two hyperparameters. For simplicity, we assume $\mathbf{w}$ is directly observed: $\mathbf{w} = (0.01, 2.0, 0.1)$. This choice of $\mathbf{w}$ favours priors that give flexibility to component 2. Components 1 and 3 do not call for such flexibility.

Sum model: We build $\mathbf{M}$ as a sum of two matrices, diag(1,1,0) and diag(0,1,1).

$$\mathbf{M} = \mathrm{diag}(\alpha_a, \alpha_a + \alpha_b, \alpha_b). \tag{17}$$

Figure 5a shows the log probability $\log P(\mathbf{w}|\{\alpha\})$ as a function of $\log \alpha_a$ and $\log \alpha_b$. The function is convex.

Covariance sum model: We now build $\mathbf{M}^{-1}$ as a sum of diag(1,1,0) and diag(0,1,1), letting:

$$\mathbf{M} = \mathrm{diag}\left(\frac{1}{c_a}, \frac{1}{c_a + c_b}, \frac{1}{c_b}\right). \tag{18}$$

Figure 5b shows the log probability $\log P(\mathbf{w}|\{c\})$ as a function of $\log c_a$ and $\log c_b$. The function is not convex. The two alternative flavours of flexibility compete with each other to give the required variance for component 2 of $\mathbf{w}$. Either we may switch on c_a to a large value, or we may switch on c_b — but we may not switch on both to an intermediate degree.

Exponential sum model: We build $\mathbf{M}$ as a sum of three matrices, diag(1,0,0), diag(0,1,0). and diag(0,0,1), with the aid of basis functions ψ_a = (1,1,0) and ψ_b = (0,1,1). Thus

$$\mathbf{M} = \mathrm{diag}\left(e^{u_a}, e^{u_a + u_b}, e^{u_b}\right). \tag{19}$$

This model has the same number of hyperparameters as the previous two models but uses them differently. Figure 5c shows the log probability $\log P(\mathbf{w}|\{u\})$ as a function of u_a and u_b. The function is convex. Two alternative flavours of flexibility are embodied, but (just) do not compete with each other destructively.

The sum model starts from flexibility and adds in constraints of stiffness that kill degrees of freedom in $\mathbf{w}$. The covariance sum representation starts from stiffness and adds in selective flexibility to create required degrees of freedom. The covariance sum model is not convex because different forms of flexibility *compete* to account for the data. There is a struggle for existence, because any potential piece of flexibility is penalized by Occam factors in the $\det \mathbf{M}$ term, encouraging it to stay switched off. In contrast, alternative ways of introducing stiffness (as in the sum model and the exponential sum model) do not compete. If two sorts of stiffness are compatible with the data, they can both be switched on without incurring any penalty. This is why the sum model is convex. The exponential sum model, we conjecture, pushes flexibility to the limits of convexity. We believe these ideas may be relevant to the design of computationally tractable Gaussian process models for non-linear regression (Williams and Rasmussen 1996).

4.4 How to represent a covariance matrix

In this paper we have used interpolation of neural spike data as a test bed for the new models. We now discuss another task to which the general principles we have discussed may apply.

Imagine that we wish to model correlations between k variables $\mathbf{y} = (y_1 \ldots y_k)^{\mathrm{T}}$ that are assumed to be Gaussian with a covariance matrix $\mathbf{V}$ that varies with other variables $\mathbf{x}$. How should this varying covariance matrix $\mathbf{V}(\mathbf{x})$ be parameterized? We assume that a representation $\mathbf{V}(\mathbf{U}(\mathbf{x}))$ is to be used. We would like the parameterization $\mathbf{V}(\mathbf{U})$ to satisfy the following desiderata.

1. Any setting of the parameters $\mathbf{U}$ should produce a valid positive definite matrix $\mathbf{V}$.

2. Any positive definite matrix $\mathbf{V}$ should be realizable by a unique value of the parameters $\mathbf{U}$.

3. The parameterization and its inverse should be continuous and differentiable.

4. The representation should treat all indices of the covariance matrix symmetrically; for example, the first row of $\mathbf{V}$ should not be treated differently from the second row.

5. $\mathbf{U}$ should have $k(k+1)/2$ degrees of freedom, that being the number of independent elements in the symmetric matrix $\mathbf{V}$.

6. Finally we would like the representation to be conditionally convex; that is, given one or more vectors $\mathbf{y}$, the conditional probability of $\mathbf{U}$ should be log convex.

These desiderata rule out most obvious representations of $\mathbf{V}$. The raw matrix $\mathbf{V}$ is not permitted because it violates desideratum 1. A triangular decomposition violates 4. An eigenvector / eigenvalue representation violates 2,3,5. The 'variance component model' representation used for example in Gu and Wahba (1991) is a covariance sum representation and violates desiderata 5 and 6.

The ideas of this paper motivate the following representation, which is conditionally convex. Let $\mathbf{y}$ be k dimensional, and let $\mathcal{R}_{k-1}$ be the unit spherical surface, with $\mathbf{v}$ being a unit vector in that space. As parameters we introduce a symmetric matrix $\mathbf{U}$ that is not constrained to be positive definite. Then we represent $\mathbf{V}$ as the inverse of a sum of outer products thus:

$$\mathbf{V}(\mathbf{U}) = \left[\int_{\mathcal{R}_{k-1}} d^{k-1}\mathbf{v} \; \exp(\mathbf{v}^{\mathrm{T}}\mathbf{U}\mathbf{v}) \; \mathbf{v}\mathbf{v}^{\mathrm{T}} \right]^{-1} \tag{20}$$

This representation satisfies all the desiderata. Since this may not be self-evident, we include a sketch of a proof of half of property 2, namely, that the mapping from $\mathbf{U}$ to $\mathbf{V}$ is one to one. We first transform into the eigenvector basis of $\mathbf{U}$ (by an orthogonal transformation that leaves $\mathcal{R}_{k-1}$ invariant) and prove that the eigenvectors $\{\mathbf{e}\}$ of $\mathbf{U}$ are also eigenvectors of $\mathbf{V}$. Let $\{w_i\}$ be the components of $\mathbf{v}$ in the eigenvector basis so that $\mathbf{v} = \sum w_i \mathbf{e}_{(i)}$, where the eigenvectors and eigenvalues of $\mathbf{U}$ satisfy $\mathbf{U}\mathbf{e}_{(i)} = \lambda_i^U \mathbf{e}_{(i)}$. Then from equation

(20) we can write

$$\mathbf{V}^{-1} = \sum_{i,j} \int_{\mathcal{R}_{k-1}} d^{k-1}\mathbf{w} \, \exp\left(\sum_j \lambda_j^U w_j^2\right) \, w_i w_j \mathbf{e}_{(i)} \mathbf{e}_{(j)}^{\mathrm{T}} \tag{21}$$

The integrand, for $i \neq j$, is antisymmetric in w_i and w_j, so the integral is zero in these cases. Thus

$$\mathbf{V}^{-1} = \sum_i \mathbf{e}_{(i)} \mathbf{e}_{(i)}^{\mathrm{T}} \int_{\mathcal{R}_{k-1}} d^{k-1}\mathbf{w} \, \exp\left(\sum_j \lambda_j^U w_j^2\right) \, w_i^2, \tag{22}$$

that is, $\mathbf{V}$ has the same eigenvectors as $\mathbf{U}$, and its eigenvalues are given by:

$$(\lambda_i^V)^{-1} = \int_{\mathcal{R}_{k-1}} d^{k-1}\mathbf{w} \, w_i^2 \exp\left(\sum_j \lambda_j^U w_j^2\right), \tag{23}$$

Then the mapping from $\mathbf{U}$ to $\mathbf{V}$ is one to one if the above mapping from the eigenvalues of $\mathbf{U}$, $\{\lambda^U\}$, to the eigenvalues of $\mathbf{V}$, $\{\lambda^V\}$, is one to one. We differentiate equation (23) to obtain the Jacobian; if the Jacobian is full-rank then the mapping is one to one.

$$\frac{\partial(\lambda_i^V)^{-1}}{\partial\lambda_h^U} = \int_{\mathcal{R}_{k-1}} d^{k-1}\mathbf{w} \, w_i^2 w_h^2 \exp\left(\sum_j \lambda_j^U w_j^2\right). \tag{24}$$

This Jacobian is a sum of outer products of positive vectors $\mathbf{z}$ given by $z_i = w_i^2$, so it either defines a positive semi-definite or a positive definite matrix. The matrix can only be positive semi-definite if there is some direction $\mathbf{h}$ such that $\sum_i h_i w_i^2 = 0$ for all $\mathbf{w}$ having non-zero measure under the integral over $\mathcal{R}_{k-1}$. Because the integral is over all of $\mathcal{R}_{k-1}$, there is no such vector $\mathbf{h}$. Thus the matrix is full rank, and the mapping from $\mathbf{U}$ to $\mathbf{V}$ is one to one.

The only problem with this representation is that it involves a high-dimensional integral. We propose for practical purposes the following approximation:

$$\mathbf{V}(\mathbf{U}) = \left[\frac{1}{C} \sum_{c=1}^{C} \exp(\mathbf{v}_c^{\mathrm{T}} \mathbf{U} \mathbf{v}_c) \, \mathbf{v}_c \mathbf{v}_c^{\mathrm{T}}\right]^{-1} \tag{25}$$

where $\{\mathbf{v}_c\}_{c=1}^{C}$ are fixed unit vectors lying in $\mathcal{R}_{k-1}$, selected either at random or systematically. This representation is conditionally convex and is able to represent arbitrary $\mathbf{V}$ in the limit $C \rightarrow \infty$.

5 Conclusions

This work builds on a data modelling philosophy previously illustrated by work on the 'Automatic Relevance Determination' model for neural networks (MacKay 1994; Neal 1996): use a huge, flexible model with an essentially infinite number of parameters; and control the complexity of the model with sophisticated regularizers. Models with large numbers of hyperparameters can, if carefully designed, be practically implemented. The hyperparameters reduce the effective number of degrees of freedom of the model in a manner appropriate to the the properties of the data, leading to substantial improvements in generalization error.

Acknowledgements

D.J.C.M. thanks the Isaac Newton Institute and T. Matsumoto, Waseda University, for hospitality, and Radford Neal, Mike Lewicki, David Mumford and Brian Ripley for helpful discussions. R.T. thanks T. Matsumoto for his support. We also thank the referees for helpful feedback.

References

Blake, A., and Zisserman, A. (1987) *Visual Reconstruction*. Cambridge Mass.: MIT Press.

Carter, C. K., and Kohn, R. (1994) On Gibbs sampling for state-space models. *Biometrika* **81** (3): 541–553.

Gilks, W., and Wild, P. (1992) Adaptive rejection sampling for Gibbs sampling. *Applied Statistics* **41**: 337–348.

Gu, C., and Wahba, G. (1991) Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method. *SIAM J. Sci. Stat. Comput.* **12**: 383–398.

Gull, S. F. (1989) Developments in maximum entropy data analysis. In *Maximum Entropy and Bayesian Methods, Cambridge 1988*, ed. by J. Skilling, pp. 53–71, Dordrecht. Kluwer.

Kimeldorf, G. S., and Wahba, G. (1970) A correspondence between Bayesian estimation of stochastic processes and smoothing by splines. *Annals of Mathematical Statistics* **41** (2): 495–502.

Lewicki, M. (1994) Bayesian modeling and classification of neural signals. *Neural Computation* **6** (5): 1005–1030.

MacKay, D. J. C. (1992) Bayesian interpolation. *Neural Computation* **4** (3): 415–447.

MacKay, D. J. C. (1994) Bayesian non-linear modelling for the prediction competition. In *ASHRAE Transactions, V.100, Pt.2*, pp. 1053–1062, Atlanta Georgia. ASHRAE.

MacKay, D. J. C. (1995) Probabilistic networks: New models and new methods. In *ICANN '95*, pp. 331–337, Paris. EC2 and Cie.

MacKay, D. J. C. (1996) Hyperparameters: Optimize, or integrate out? In *Maximum Entropy and Bayesian Methods, Santa Barbara 1993*, ed. by G. Heidbreder, pp. 43–60, Dordrecht. Kluwer.

Muller, H. G., and Stadtmuller, U. (1987) Variable bandwidth kernel estimators of regression-curves. *Annals of Statistics* **15** (1): 182–201.

Neal, R. M. (1993) Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG–TR–93–1, Dept. of Computer Science, University of Toronto.

Neal, R. M. (1996) *Bayesian Learning for Neural Networks*. Number 118 in Lecture Notes in Statistics. New York: Springer.

Shephard, N. (1994) Partial non-Gaussian state-space. *Biometrika* **81** (1): 115–131.

Smith, A. (1991) Bayesian computational methods. *Philosophical Transactions of the Royal Society of London A* **337**: 369–386.

Thomas, A., Spiegelhalter, D. J., and Gilks, W. R. (1992) BUGS: A program to perform Bayesian inference using Gibbs sampling. In *Bayesian Statistics 4*, ed. by J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, pp. 837–842. Oxford: Clarendon Press.

West, M. (1984) Outlier models and prior distributions in Bayesian linear-regression. *Journal of the Royal Statistical Society Series B-Methodological* **46** (3): 431–439.

Williams, C. K. I., and Rasmussen, C. E. (1996) Gaussian processes for regression. In *Advances in Neural Information Processing Systems 8*, ed. by D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo. MIT Press.

Wolpert, D. H. (1993) On the use of evidence in neural networks. In *Advances in Neural Information Processing Systems 5*, ed. by C. L. Giles, S. J. Hanson, and J. D. Cowan, pp. 539–546, San Mateo, California. Morgan Kaufmann.