Bayesian linear regression

Thomas P. Minka

February 20, 2001

Abstract

This note derives the posterior, evidence, and predictive density for linear multivariate regression under zero-mean Gaussian noise. Many Bayesian texts, such as Box & Tiao (1973), cover linear regression. This note contributes to the discussion by paying careful attention to invariance issues, demonstrating model selection based on the evidence, and illustrating the shape of the predictive density. Piecewise regression, basis function regression, and automatic relevance determination are also discussed.

1 Introduction

The data model is that an input vector $\mathbf{x}$ of length m multiplies a coefficient matrix $\mathbf{A}$ to produce an output vector $\mathbf{y}$ of length d, with Gaussian noise added:

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e} \tag{1}$$

$$\mathbf{e} \sim \mathcal{N}(0, \mathbf{V}) \tag{2}$$

$$p(\mathbf{y}|\mathbf{x}, \mathbf{A}, \mathbf{V}) \sim \mathcal{N}(\mathbf{A}\mathbf{x}, \mathbf{V}) \tag{3}$$

This is a conditional model for $\mathbf{y}$ only: the distribution of $\mathbf{x}$ is not needed and in fact irrelevant to all inferences in this paper. As we shall see, conditional models create subtleties in Bayesian inference. In the special case $\mathbf{x} = 1$ and $m = 1$, the conditioning disappears and we simply have a Gaussian distribution for $\mathbf{y}$, with arbitrary mean and variance. This case is useful as a check on the results.

The scenario is that we are given a data set of exchangeable pairs $D = \{(\mathbf{y}_1, \mathbf{x}_1), ..., (\mathbf{y}_N, \mathbf{x}_N)\}$. Collect $\mathbf{Y} = [\mathbf{y}_1 \cdots \mathbf{y}_N]$ and $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_N]$. The distribution of $\mathbf{Y}$ given $\mathbf{X}$ under the model is

$$
\begin{aligned}
p(\mathbf{Y}|\mathbf{X}, \mathbf{A}, \mathbf{V}) &= \prod_i p(\mathbf{y}_i|\mathbf{x}_i, \mathbf{A}, \mathbf{V}) && (4)\\
&= \frac{1}{|2\pi\mathbf{V}|^{N/2}} \exp(-\frac{1}{2}\sum_i (\mathbf{y}_i - \mathbf{A}\mathbf{x}_i)^{\mathrm{T}} \mathbf{V}^{-1} (\mathbf{y}_i - \mathbf{A}\mathbf{x}_i)) && (5)\\
&= \frac{1}{|2\pi\mathbf{V}|^{N/2}} \exp(-\frac{1}{2}\mathrm{tr}(\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{A}\mathbf{X})(\mathbf{Y} - \mathbf{A}\mathbf{X})^{\mathrm{T}})) && (6)\\
&= \frac{1}{|2\pi\mathbf{V}|^{N/2}} \exp(-\frac{1}{2}\mathrm{tr}(\mathbf{V}^{-1}\left[\mathbf{A}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}} - 2\mathbf{Y}\mathbf{X}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}} + \mathbf{Y}\mathbf{Y}^{\mathrm{T}}\right])) && (7)
\end{aligned}
$$

2 The matrix-normal density

A conjugate prior for $\mathbf{A}$ is the matrix-normal density, which implies the posterior for $\mathbf{A}$ as well as for $\mathbf{Y}$ will be matrix-normal. Since this density arises so often let's take some time out to study it. A random d by m matrix $\mathbf{A}$ is matrix-normal distributed with parameters $\mathbf{M}$, $\mathbf{V}$, and $\mathbf{K}$ if the density of $\mathbf{A}$ is

$$p(\mathbf{A}) \sim \mathcal{N}(\mathbf{M}, \mathbf{V}, \mathbf{K}) \tag{8}$$

$$= \frac{|\mathbf{K}|^{d/2}}{|2\pi\mathbf{V}|^{m/2}} \exp(-\frac{1}{2}\mathrm{tr}((\mathbf{A}-\mathbf{M})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{A}-\mathbf{M})\mathbf{K})) \tag{9}$$

where $\mathbf{M}$ is d by m, $\mathbf{V}$ is d by d, and $\mathbf{K}$ is m by m. This distribution has *two* covariance matrices: $\mathbf{V}$ for the rows and $\mathbf{K}$ for the columns. If $\mathbf{V}$ is diagonal, then the rows of $\mathbf{A}$ are independent normal vectors. If $\mathbf{K}$ is diagonal, then the columns of $\mathbf{A}$ are independent normal vectors. If $\mathrm{vec}(\mathbf{A})$ denotes the stacked columns of $\mathbf{A}$, we have

$$p(\mathrm{vec}(\mathbf{A})) \sim \mathcal{N}(\mathrm{vec}(\mathbf{M}), \mathbf{V} \otimes \mathbf{K}^{-1}) \tag{10}$$

From this formula we can derive properties of the matrix-normal distribution from the normal distribution. Note that the covariance of $\mathrm{vec}(\mathbf{A})$ is highly structured. This means that, in general, the sum of two normal matrices is *not* normal, unless they have the same $\mathbf{V}$ or $\mathbf{K}$ parameter.

The matrix $\mathbf{Y} = \mathbf{XA}$ has derived density

$$p(\mathbf{Y}) \sim \mathcal{N}(\mathbf{XM}, \mathbf{XVX}^{\mathrm{T}}, \mathbf{K}) \tag{11}$$

The matrix $\mathbf{A}^{\mathrm{T}}$ has derived density

$$p(\mathbf{A}^{\mathrm{T}}) \sim \mathcal{N}(\mathbf{M}^{\mathrm{T}}, \mathbf{K}^{-1}, \mathbf{V}^{-1}) \tag{12}$$

Hence $\mathbf{Y} = \mathbf{AX}$ has derived density

$$p(\mathbf{Y}) \sim \mathcal{N}(\mathbf{MX}, \mathbf{V}, (\mathbf{X}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{X})^{-1}) \tag{13}$$

which we will use later.

3 An invariant prior

When our knowledge about $\mathbf{A}$ is vague, we can use invariance arguments to derive a prior. Typically $\mathbf{Y}$ is a set of measurements like wind speed, blood pressure, or stock price and $\mathbf{X}$ is some other set of measurements like location, drug dosage, or fed interest rate. In such cases, we would like our inferences to be invariant to the units of measurement. That is, our inferences under one system of measurement should simply be a scaled version of those under another system. For measurements like 2D location, we would also want to be invariant to the choice of basis, i.e. an affine transformation of the measurement matrix should transform our inferences in precisely the same way. For example, if given $(\mathbf{Y}, \mathbf{X})$ we infer $\mathbf{A} = \mathbf{A}_0$, then given $(\mathbf{BY}, \mathbf{CX})$ we should infer $\mathbf{A} = \mathbf{BA}_0\mathbf{C}^{-1}$.

This condition is satisfied by any distribution of the form

$$p(\mathbf{A}|\mathbf{X}, \mathbf{V}, \mathbf{W}) \sim \mathcal{N}(\mathbf{0}, \mathbf{V}, \mathbf{XWX}^{\mathrm{T}}) \tag{14}$$

where $\mathbf{XWX}^{\mathrm{T}}$ is nonsingular. You can verify that this density transforms in the correct way if you transform either $\mathbf{Y}$ (and therefore $\mathbf{V}$) or $\mathbf{X}$. Other priors, such as the ridge regression prior (MacKay, 1992)

$$p(\mathbf{A}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d, \alpha\mathbf{I}_m) \tag{15}$$

are not invariant; inferences can change, even qualitatively, when you rescale a measurement. This paper focuses on the case $\mathbf{W} = \alpha\mathbf{I}$:

$$p(\mathbf{A}|\mathbf{X}, \mathbf{V}, \alpha) \sim \mathcal{N}(\mathbf{0}, \mathbf{V}, \alpha\mathbf{XX}^{\mathrm{T}}) \tag{16}$$

The Jeffreys prior for linear regression is obtained as the limit

$$p(\mathbf{A}|\mathbf{X}, \mathbf{V}) \sim \lim_{\alpha\to 0} \mathcal{N}(\mathbf{0}, \mathbf{V}, \alpha\mathbf{XX}^{\mathrm{T}}) \tag{17}$$

$$\propto \left|\mathbf{XX}^{\mathrm{T}}\right|^{d/2} |2\pi\mathbf{V}|^{-m/2} \tag{18}$$

The problem with the Jeffreys prior is that it is improper. This is avoided by leaving α as a free parameter to be optimized or integrated out. This regression technique was also advocated by Gull (1988). In this paper, α is optimized via empirical Bayes.

Note that the mean of $\mathbf{A}$ under the prior must be zero in order to achieve invariance. This is interesting because, even though shrinkage priors are widely used in regression, it finally gives a precise reason why the shrinkage point should be zero.

In the non-regression case, when we are just estimating a Gaussian distribution for $\mathbf{y}$, the invariant prior is

$$p(\mathbf{a}|\mathbf{V}) \sim \mathcal{N}(\mathbf{0}, \frac{\mathbf{V}}{\alpha N}) \tag{19}$$

4 Known V

Now back to the problem. Let

$$\mathbf{S}_{xx} = \mathbf{X}\mathbf{X}^{\mathrm{T}} + \mathbf{K} \tag{20}$$
$$\mathbf{S}_{yx} = \mathbf{Y}\mathbf{X}^{\mathrm{T}} + \mathbf{M}\mathbf{K} \tag{21}$$
$$\mathbf{S}_{yy} = \mathbf{Y}\mathbf{Y}^{\mathrm{T}} + \mathbf{M}\mathbf{K}\mathbf{M}^{\mathrm{T}} \tag{22}$$
$$\mathbf{S}_{y|x} = \mathbf{S}_{yy} - \mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{yx}^{\mathrm{T}} \tag{23}$$

Then the likelihood (7) times conjugate prior (8) is

$$p(\mathbf{Y},\mathbf{A}|\mathbf{X},\mathbf{V}) \propto \exp(-\frac{1}{2}\mathrm{tr}(\mathbf{V}^{-1}\left[\mathbf{A}\mathbf{S}_{xx}\mathbf{A}^{\mathrm{T}} - 2\mathbf{S}_{yx}\mathbf{A}^{\mathrm{T}} + \mathbf{S}_{yy}\right])) \tag{24}$$
$$\propto \exp(-\frac{1}{2}\mathrm{tr}(\mathbf{V}^{-1}\left[(\mathbf{A} - \mathbf{S}_{yx}\mathbf{S}_{xx}^{-1})\mathbf{S}_{xx}(\mathbf{A} - \mathbf{S}_{yx}\mathbf{S}_{xx}^{-1})^{\mathrm{T}} + \mathbf{S}_{y|x}\right])) \tag{25}$$

so the posterior for $\mathbf{A}$ is matrix-normal:

$$p(\mathbf{A}|D,\mathbf{V}) \sim \mathcal{N}(\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}, \mathbf{V}, \mathbf{S}_{xx}) \tag{26}$$

With the invariant prior (16) this becomes

$$p(\mathbf{A}|D,\mathbf{V},\alpha) \sim \mathcal{N}(\mathbf{Y}\mathbf{X}^{\mathrm{T}}(\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1}(\alpha+1)^{-1}, \mathbf{V}, \mathbf{X}\mathbf{X}^{\mathrm{T}}(\alpha+1)) \tag{27}$$

For the non-regression model, the posterior is

$$p(\mathbf{a}|\mathbf{Y},\mathbf{V},\alpha) \sim \mathcal{N}((\alpha+1)^{-1}\bar{\mathbf{y}}, \frac{1}{(\alpha+1)N}\mathbf{V}) \tag{28}$$
$$\bar{\mathbf{y}} = \frac{1}{N}\sum_i \mathbf{y}_i \tag{29}$$

The mode of the posterior differs from the likelihood maximum by the shrinkage factor $(\alpha+1)^{-1}$, which provides some protection against overfitting.

4.1 Model selection via the evidence

Returning to the joint distribution (25) and integrating out $\mathbf{A}$ gives the evidence for linearity, with $\mathbf{V}$ known:

$$p(\mathbf{Y}|\mathbf{X},\mathbf{V}) = \frac{|\mathbf{K}|^{d/2}}{|\mathbf{S}_{xx}|^{d/2}\,|2\pi\mathbf{V}|^{N/2}}\exp(-\frac{1}{2}\mathrm{tr}(\mathbf{V}^{-1}\mathbf{S}_{y|x})) \tag{30}$$
$$\sim \mathcal{N}(\mathbf{M}\mathbf{X}, \mathbf{V}, \mathbf{I}_N - \mathbf{X}^{\mathrm{T}}\mathbf{S}_{xx}^{-1}\mathbf{X}) \tag{31}$$

This can also be derived from (1), (8), and the properties of the matrix-normal distribution. The invariant prior reduces this to

$$p(\mathbf{Y}|\mathbf{X},\mathbf{V},\alpha) = \left(\frac{\alpha}{\alpha+1}\right)^{md/2} |2\pi\mathbf{V}|^{-N/2} \exp(-\frac{1}{2}\mathrm{tr}(\mathbf{V}^{-1}\mathbf{S}_{y|x})) \tag{32}$$

$$\mathbf{S}_{y|x} = \mathbf{Y}\mathbf{Y}^{\mathrm{T}} - (\alpha+1)^{-1}\mathbf{Y}\mathbf{X}^{\mathrm{T}}(\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1}\mathbf{X}\mathbf{Y}^{\mathrm{T}} \tag{33}$$

When $\alpha = 0$, $\mathbf{S}_{y|x}$ attains its smallest value of

$$\mathbf{S}_{y|x} = \mathbf{Y}(\mathbf{I}_N - \mathbf{X}^{\mathrm{T}}(\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1}\mathbf{X})\mathbf{Y}^{\mathrm{T}} \tag{34}$$

which has an intuitive geometrical interpretation. The matrix $\mathbf{X}^{\mathrm{T}}(\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1}\mathbf{X}$ is the projection matrix for the subspace spanned by the columns of $\mathbf{X}$. Therefore $\mathbf{I}_N - \mathbf{X}^{\mathrm{T}}(\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1}\mathbf{X}$ extracts the component of $\mathbf{Y}$ orthogonal to the input, which is precisely the noise $\mathbf{e}$.

However, even though $\alpha = 0$ provides the best fit to the data, the probability of the data is zero. This is because the prior is so broad that any particular dataset must get vanishingly small probability. The only way to increase the probability assigned to D is to make the prior narrower, which also means shrinking the regression coefficients toward zero.

By zeroing the gradient with respect to α, we find that the evidence is maximized when

$$\alpha = \frac{md}{\mathrm{tr}(\mathbf{V}^{-1}\mathbf{Y}\mathbf{X}^{\mathrm{T}}(\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1}\mathbf{X}\mathbf{Y}^{\mathrm{T}}) - md} \tag{35}$$

This estimator behaves in a reasonable way: it shrinks more when N is small, m is large, or the noise level $\mathbf{V}$ is large, in order to reduce overfitting.

The evidence for linearity is useful for selecting among different linear models, namely models with different inputs. The different inputs might be different nonlinear transformations of the measurements. If we consider the different inputs as separate models with separate priors, then we compute (35) and (30) for each model and see which is largest. Figure 1 has an example of using this rule to select polynomial order. The data is synthetic with $N = 50$ and known variance $\mathbf{V} = 10$. For order k, the input vector is $\mathbf{x} = [1\ x\ x^2 \cdots x^k]^{\mathrm{T}}$. Because of the invariant prior, it doesn't matter if we use monomials versus Legendre polynomials or Hermite polynomials (though for MacKay (1992), it did matter).

Another approach is to construct a composite model with all possible inputs and determine which coefficients to set to zero. This method is mathematically identical to the first except that all models use the same value of α. Unfortunately, this makes model selection more difficult because typically the best model depends on α.

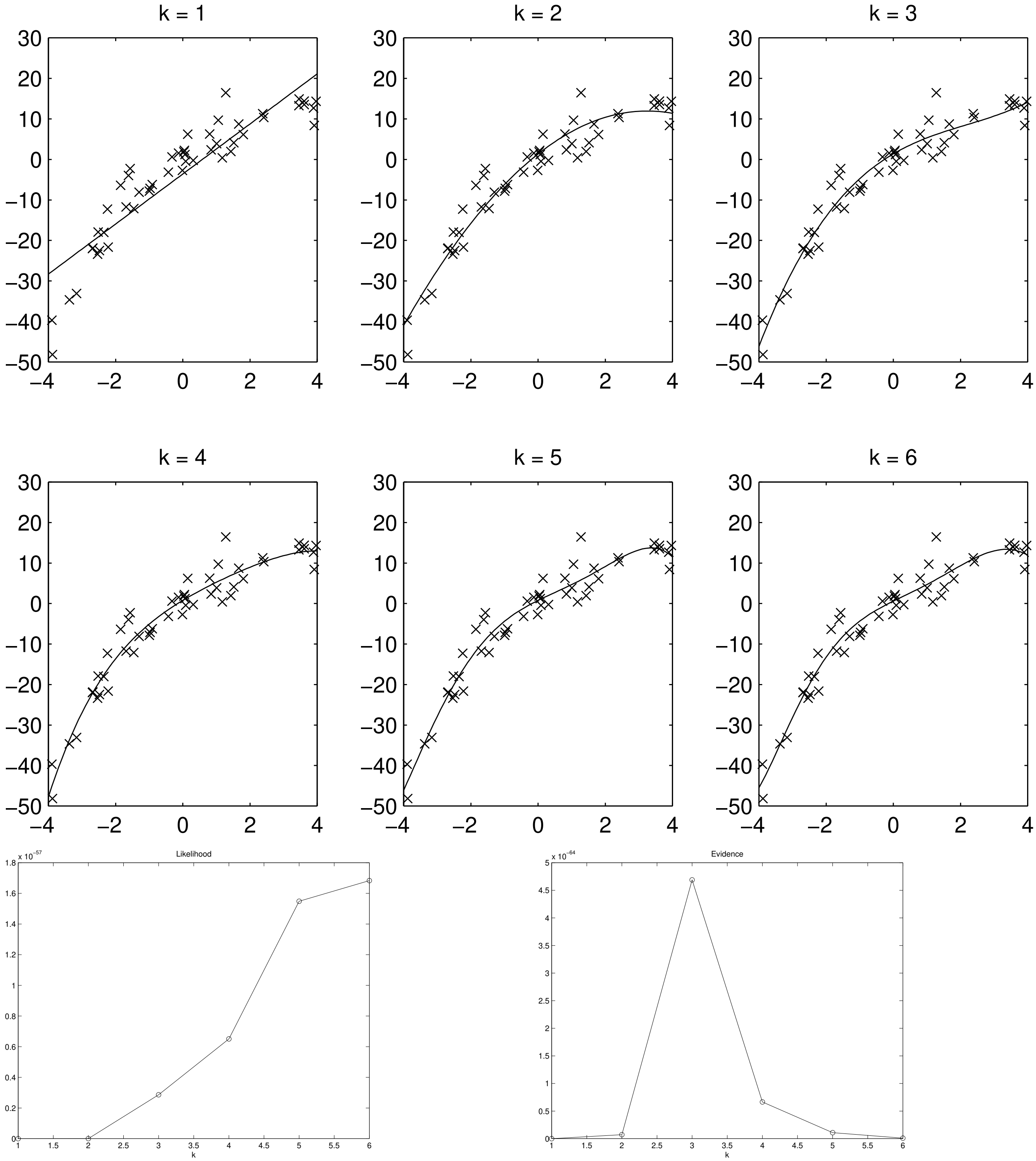

Figure 1: Example of using the evidence to select model order. A synthetic data set is approximated by polynomials of varying order. The likelihood curve always increases with increasing order while the evidence curve has a clear maximum at $k = 3$ (the true order in this case).

In the non-regression case, the evidence for Gaussianity is

$$p(D|\mathbf{V},\alpha) = \left(\frac{\alpha}{\alpha+1}\right)^{md/2} |2\pi\mathbf{V}|^{-N/2} \exp(-\frac{1}{2}\mathrm{tr}(\mathbf{V}^{-1}\mathbf{S}) \quad (36)$$

$$\mathbf{S} = (\sum_i \mathbf{y}_i\mathbf{y}_i^{\mathrm{T}}) - \frac{N}{(\alpha+1)}\bar{\mathbf{y}}\bar{\mathbf{y}}^{\mathrm{T}} \quad (37)$$

$$= \mathbf{Y}(\mathbf{I}_N - \frac{1}{(\alpha+1)N}\mathbf{1}\mathbf{1}^{\mathrm{T}})\mathbf{Y}^{\mathrm{T}} \quad (38)$$

which incorporates shrinkage of the mean. This formula can be used for a variety of purposes, such as testing whether two populations with the same variance also have the same mean. The evidence is maximized when

$$\alpha = \frac{d}{N\bar{\mathbf{y}}^{\mathrm{T}}\mathbf{V}^{-1}\bar{\mathbf{y}} - d} \quad (39)$$

4.2 Predicting new outputs

To predict the $\mathbf{y}$ output for a new $\mathbf{x}$ input, combine the matrix-normal distribution (26) for $\mathbf{A}$ with the definition (1) of $\mathbf{y}$ to get

$$p(\mathbf{y}|\mathbf{x},D,\mathbf{V}) = \mathcal{N}(\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{x}, \mathbf{V}c^{-1}) \quad (40)$$

$$c = (1+\mathbf{x}^{\mathrm{T}}\mathbf{S}_{xx}^{-1}\mathbf{x})^{-1} = 1 - \mathbf{x}^{\mathrm{T}}(\mathbf{S}_{xx}+\mathbf{x}\mathbf{x}^{\mathrm{T}})^{-1}\mathbf{x} \quad (41)$$

This result can also be derived by considering an augmented data set $D' = \{\mathbf{y},\mathbf{x}\} \cup D$ and then computing the evidence ratio $p(D'|\mathbf{V})/p(D|\mathbf{V})$.

Since $E[\mathbf{y}|\mathbf{x},D] = E[\mathbf{A}|D]\mathbf{x}$, the expected value of $\mathbf{y}$ given $\mathbf{x}$ is identical to substituting the posterior mean (or mode) for $\mathbf{A}$. But the variance of $\mathbf{y}$ is not simply $\mathbf{V}$; it depends on the input $\mathbf{x}$. Figure 2 plots the contours of the predictive density conditional on x. The mean is a straight line with slope $\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}$, while the standard deviation lines are curved to account for uncertainty in the model. That is, the model is allowed to wiggle within the constraints provided by the data. Only near the training data can predictions be considered reliable.

In the non-regression case, we have $c^{-1} = 1 + \frac{1}{N(\alpha+1)}$ so the predictive density is

$$p(\mathbf{y}|D) \sim \mathcal{N}((\alpha+1)^{-1}\bar{\mathbf{y}}, \frac{N(\alpha+1)+1}{N(\alpha+1)}\mathbf{V}) \quad (42)$$

which again incorporates wiggle of the unknown mean.

The predictive density for K new samples $(\mathbf{Y}',\mathbf{X}')$ is derived in the same way, via (13):

$$p(\mathbf{Y}'|\mathbf{X}',D,\mathbf{V}) \sim \mathcal{N}(\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{X}', \mathbf{V}, \mathbf{C}) \quad (43)$$

$$\mathbf{C} = (\mathbf{I}_K + (\mathbf{X}')^{\mathrm{T}}\mathbf{S}_{xx}^{-1}\mathbf{X}')^{-1} = \mathbf{I}_K - (\mathbf{X}')^{\mathrm{T}}(\mathbf{S}_{xx}+\mathbf{X}'(\mathbf{X}')^{\mathrm{T}})^{-1}\mathbf{X}' \quad (44)$$

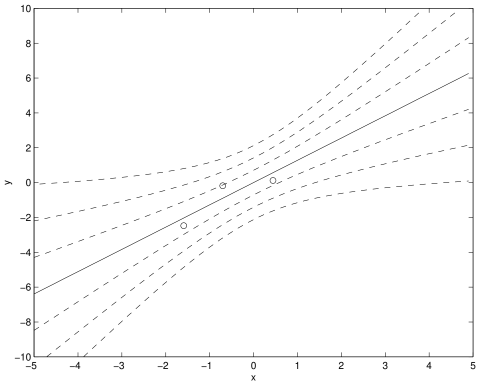

Figure 2: Contours of the predictive density for y, given three training points (circles). The model is $y = ax$ (a line through the origin).

In the non-regression case, we have

$$\mathbf{C} = \mathbf{I}_K - \mathbf{1}\mathbf{1}^{\mathrm{T}}/(N(\alpha+1)+K) \tag{45}$$

$$p(\mathbf{Y}'|\mathbf{Y},\mathbf{V}) \sim \mathcal{N}((\alpha+1)^{-1}\bar{\mathbf{y}}\mathbf{1}^{\mathrm{T}}, \mathbf{V}, \mathbf{C}) \tag{46}$$

5 Unknown V

A conjugate prior for $\mathbf{V}$ is the inverse Wishart distribution:

$$\begin{aligned} p(\mathbf{V}) &\sim \mathcal{W}^{-1}(\mathbf{S}_0, n) && (47) \\ &= \frac{1}{Z_{nd}\,|\mathbf{V}|^{(d+1)/2}} \left|\frac{\mathbf{V}^{-1}\mathbf{S}_0}{2}\right|^{n/2} \exp(-\frac{1}{2}\mathrm{tr}(\mathbf{V}^{-1}\mathbf{S}_0)) && (48) \\ \text{where } Z_{nd} &= \pi^{d(d-1)/4}\prod_{i=1}^{d}\Gamma((n+1-i)/2) \end{aligned}$$

The invariant prior used in this paper is

$$p(\mathbf{V}|N_0) \sim \mathcal{W}^{-1}(N_0\mathbf{I}_d, N_0) \tag{49}$$

where N_0 is optimized via empirical Bayes. The Jeffreys prior for $\mathbf{V}$ is obtained in the limit $N_0 \to 0$.

The marginal likelihood (30) times the conjugate prior is

$$p(\mathbf{Y},\mathbf{V}|\mathbf{X}) \propto \frac{1}{|\mathbf{V}|^{(d+1)/2}\,|2\pi\mathbf{V}|^{(N+N_0)/2}} \exp(-\frac{1}{2}\mathrm{tr}(\mathbf{V}^{-1}(\mathbf{S}_{y|x}+\mathbf{S}_0))) \tag{50}$$

so the posterior for $\mathbf{V}$ is inverse Wishart:

$$p(\mathbf{V}|D) \sim \mathcal{W}^{-1}(\mathbf{S}_{y|x} + \mathbf{S}_0, N + N_0) \tag{51}$$

Integrating $\mathbf{V}$ out of (50) gives the evidence for linearity:

$$\begin{aligned} p(\mathbf{Y}|\mathbf{X}) &= \frac{Z_{(N+N_0)d}}{Z_{N_0 d}} \frac{|\mathbf{K}|^{d/2}}{|\mathbf{S}_{xx}|^{d/2}} \frac{|\mathbf{S}_0|^{N_0/2}}{\pi^{Nd/2}\left|\mathbf{S}_{y|x} + \mathbf{S}_0\right|^{(N+N_0)/2}} & (52) \\ &\sim \mathcal{T}(\mathbf{M}\mathbf{X}, \mathbf{S}_0, \mathbf{I} - \mathbf{X}^{\mathrm{T}}\mathbf{S}_{xx}^{-1}\mathbf{X}, N + N_0) & (53) \end{aligned}$$

This is a matrix-T distribution, analogous to the matrix-normal. A $d \times m$ matrix $\mathbf{A}$ is matrix-T distributed if

$$\begin{aligned} p(\mathbf{A}) &\sim \mathcal{T}(\mathbf{M}, \mathbf{V}, \mathbf{K}, n) & (54) \\ &= \frac{\prod_{i=1}^{d} \Gamma((n+1-i)/2)}{\prod_{i=1}^{d} \Gamma((n-m+1-i)/2)} \frac{|\mathbf{K}|^{d/2}}{|\pi\mathbf{V}|^{m/2}} \left|(\mathbf{A}-\mathbf{M})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{A}-\mathbf{M})\mathbf{K} + \mathbf{I}_m\right|^{-n/2} & (55) \end{aligned}$$

Substituting the invariant prior gives

$$p(\mathbf{Y}|\mathbf{X}) = \frac{\prod_{i=1}^{d} \Gamma((N+N_0+1-i)/2)}{\prod_{i=1}^{d} \Gamma((N_0+1-i)/2)} \left(\frac{\alpha}{\alpha+1}\right)^{md/2} (\pi N_0)^{-Nd/2} \left|\frac{\mathbf{S}_{y|x}}{N_0} + \mathbf{I}_d\right|^{-(N+N_0)/2} \tag{56}$$

The optimum (α, N_0) can be computed by iterating the fixed-point equations

$$\begin{aligned} \hat{\mathbf{V}} &= (\mathbf{S}_{y|x} + N_0\mathbf{I}_d)/(N + N_0) & (57) \\ \alpha &= \frac{md}{\mathrm{tr}(\hat{\mathbf{V}}^{-1}\mathbf{Y}\mathbf{X}^{\mathrm{T}}(\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1}\mathbf{X}\mathbf{Y}^{\mathrm{T}}) - md} & (58) \\ N_0^{new} &= d - 1 + (N_0 + 1 - d)\frac{\sum_{i=1}^{d} \Psi((N+N_0+1-i)/2) - \Psi((N_0+1-i)/2)}{\log\left|\frac{\mathbf{S}_{y|x}}{N_0} + \mathbf{I}_d\right| + \mathrm{tr}(\hat{\mathbf{V}}^{-1}) - d} & (59) \end{aligned}$$

Similarly, we can multiply the predictive distribution (40) by the posterior for $\mathbf{V}$ and integrate out $\mathbf{V}$, to get the predictive density

$$p(\mathbf{y}|\mathbf{x}, D) = \mathcal{T}(\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{x}, (\mathbf{S}_{y|x} + \mathbf{S}_0)c^{-1}, N + N_0 + 1) \tag{60}$$

In the non-regression case, $c^{-1} = 1 + \frac{1}{N(\alpha+1)}$ so

$$p(\mathbf{y}|D) \sim \mathcal{T}((\alpha+1)^{-1}\bar{\mathbf{y}}, \frac{N(\alpha+1)+1}{N(\alpha+1)}(\mathbf{S} + \mathbf{S}_0), N + N_0 + 1) \tag{61}$$

The prediction for K new samples is

$$p(\mathbf{Y}'|\mathbf{X}', D) \sim \mathcal{T}(\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{X}', \mathbf{S}_{y|x} + \mathbf{S}_0, \mathbf{C}, N + N_0 + K) \tag{62}$$

In the non-regression case, this is

$$p(\mathbf{Y}'|D) \sim \mathcal{T}((\alpha+1)^{-1}\bar{\mathbf{y}}\mathbf{1}^{\mathrm{T}}, \mathbf{S} + \mathbf{S}_0, \mathbf{C}, N + N_0 + K) \tag{63}$$

6 Piecewise regression

Piecewise regression allows different parts of the data to follow different regression laws. Consider the model

$$p(\mathbf{y}_i|\mathbf{x}_i, \mathbf{A}_1, \mathbf{V}_1, \mathbf{A}_2, \mathbf{V}_2) \sim \begin{cases} \mathcal{N}(\mathbf{A}_1\mathbf{x}_i, \mathbf{V}_1) & \text{if } i < t \\ \mathcal{N}(\mathbf{A}_2\mathbf{x}_i, \mathbf{V}_2) & \text{if } i \geq t \end{cases} \tag{64}$$

This is known as a *changepoint* model: the first $t-1$ observations follow one model, and the rest follow another. The changepoint t is unknown and must be estimated. This model and its generalizations are useful for segmenting time-series data such as speech. See Broemeling (1985) for more discussion of this model.

In this model, the $\mathbf{x}$ values need not be increasing or have any other pattern, though in the examples they will be increasing. Also, the linear pieces do not necessarily meet.

Given a prior $p(t)$ on the changepoint location, the posterior can be readily computed via

$$p(t|D) \propto p(t)p(\mathbf{y}_1..\mathbf{y}_{t-1}|\mathbf{x}_1..\mathbf{x}_{t-1})p(\mathbf{y}_t..\mathbf{y}_N|\mathbf{x}_t..\mathbf{x}_N) \tag{65}$$

where the last two terms are given by separate applications of (52). The normalizing constant is the evidence for the existence of a changepoint:

$$p(D|\text{changepoint}) = \sum_{t=1}^{N} p(t)p(\mathbf{y}_1..\mathbf{y}_{t-1}|\mathbf{x}_1..\mathbf{x}_{t-1})p(\mathbf{y}_t..\mathbf{y}_N|\mathbf{x}_t..\mathbf{x}_N) \tag{66}$$

Fitting a line is meaningless if there are less than two data points, so a reasonable $p(t)$ is uniform from 3 to $N-1$.

Figure 3 shows two examples: one where there is a changepoint and one where there is not. In the first example, the odds of a changepoint ((66) divided by (52)) are overwhelming, while in the second example the odds are 300:1 *against* a changepoint. The optimal (α, N_0) was used in each evaluation of (52).

Seber & Wild (1989) describe a variety of ways to enforce continuity of the piecewise linear function. For example, we could use a coupled prior on $\mathbf{A}_1$ and $\mathbf{A}_2$ that requires the lines (or planes) to meet at a given point (or edge). But the simplest and most general way to get a continuous piecewise regression is the method of basis functions, described in section 7.

A model more flexible than the changepoint model is the *switching regression* model, where the regression law can switch back and forth throughout the data. For a recent paper see Chen & Liu (1996).

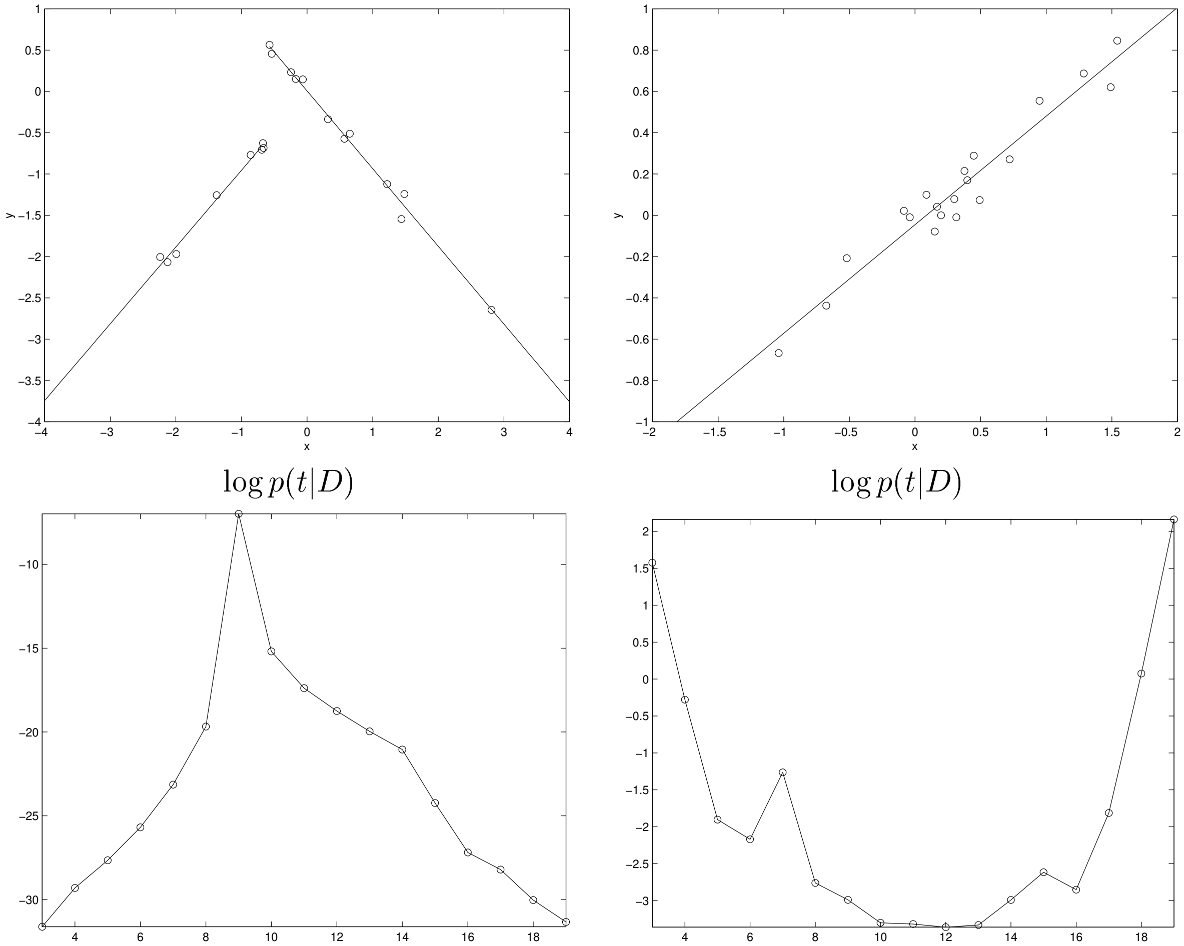

Figure 3: Example of changepoint analysis. On the left, the changepoint at $t = 9$ is correctly found. On the right, there is no changepoint.

7 Basis function regression

Basis function regression is the special case where the inputs x_i are functions of a common quantity $\mathbf{z}$:

$$x_i = f_i(\mathbf{z}) \tag{67}$$

All formulas remain the same, but now the predictive density $p(\mathbf{y}|\mathbf{x}, D)$ can be viewed as a function of $\mathbf{z}$. This technique was already used in figure 1, where $f_i(z) = z^{i-1}$. Other choices include $f_i(z) = |z - t_i|$, which yields a piecewise linear regression with changepoints t_i, $f_i(z) = \exp(-\frac{1}{2h}(z - t_i)^2)$, which superimposes smooth bumps, and $f_i(z) = \tanh(h_i(x - t_i))$, which superimposes smooth ramps. Figure 4 shows examples of these three bases.

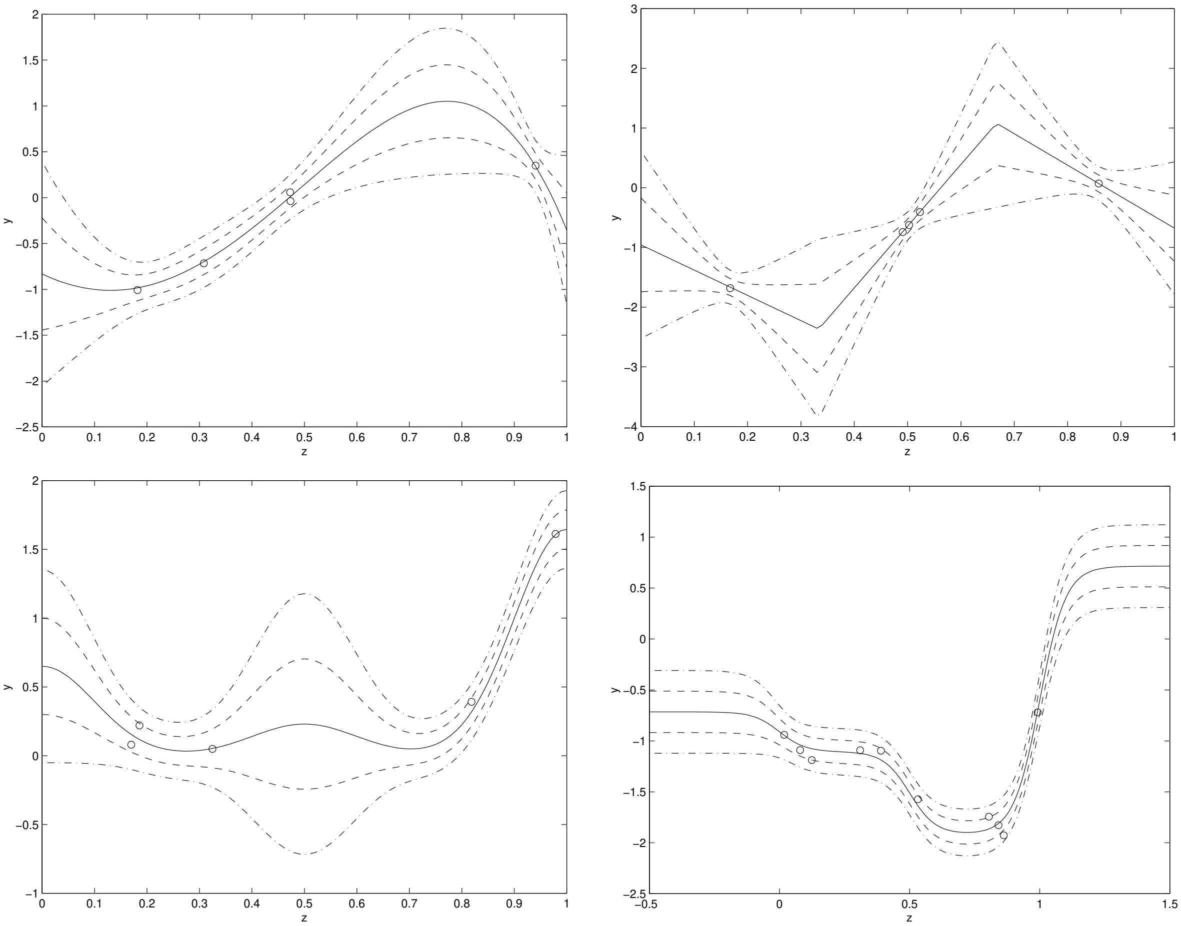

Figure 4: Example of basis function regression. The contours of the predictive density are shown for a polynomial basis, a piecewise linear basis, a Gaussian basis, and a tanh basis.

The evidence formula (52) can be used to tune parameters within the basis functions, such as the t_i's. This is an alternative to least-squares procedures for "generalized basis functions" (Poggio & Girosi, 1990). Bretthorst (1988) used the evidence technique for spectrum analysis: the basis functions were sinusoids with flexible frequency and phase parameters. Multi-layer perceptrons (MLPs) can also be viewed as basis function regressors with flexible basis functions. The basis functions are the hidden unit responses; typically tanh functions.

References

[1] George E. P. Box and George C. Tiao. *Bayesian Inference in Statistical Analysis.* Addison-Wesley, 1973.

[2] G. Larry Bretthorst. *Bayesian Spectrum Analysis and Parameter Estimation.* Springer-Verlag, 1988. `http://bayes.wustl.edu/glb/book.pdf`.

[3] Lyle Broemeling. *Bayesian analysis of linear models.* Marcel Dekker, 1985.

[4] R. Chen and J. S. Liu. Predictive updating methods with application to Bayesian classification. *Journal of the Royal Statistical Society B*, 58:397–415, 1996. `http://playfair.stanford.edu/reports/jliu/pre-up.ps.Z`.

[5] S. F. Gull. Bayesian inductive inference and maximum entropy. In G. J. Erickson and C. R. Smith, editors, *Maximum Entropy and Bayesian Methods*, pages 53–74. Kluwer Academic Publishers, 1988. `http://bayes.wustl.edu/sfg/gull.html`.

[6] D. J. C. MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 1992. `http://wol.ra.phy.cam.ac.uk/mackay/inter.nc.ps.gz`.

[7] T. Poggio and F. Girosi. Networks for approximation and learning. *Proc. of IEEE*, 78:1481–1497, 1990.

[8] G. A. F. Seber and C. J. Wild. *Nonlinear Regression.* John Wiley & Sons, 1989.