Bayesian linear regression

Thomas P. Minka

September 29, 1999

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 and basis function regression 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:

$$\begin{aligned} \mathbf{y} &= \mathbf{A}\mathbf{x} + \mathbf{e} && (1)\\ \mathbf{e} &\sim \mathcal{N}(0, \mathbf{V}) && (2)\\ p(\mathbf{y}|\mathbf{x}, \mathbf{A}, \mathbf{V}) &\sim \mathcal{N}(\mathbf{A}\mathbf{x}, \mathbf{V}) && (3) \end{aligned}$$

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

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

A conjugate prior for $\mathbf{A}$ and $\mathbf{V}$ given $\mathbf{X}$ is the matrix-Normal-Wishart density:

$$\begin{aligned}
p(\mathbf{A},\mathbf{V}|\mathbf{X}) &= p(\mathbf{A}|\mathbf{V},\mathbf{X})p(\mathbf{V}|\mathbf{X}) && (8)\\
&= \mathcal{N}(\mathbf{A};\mathbf{M},\mathbf{V},\mathbf{K})\mathcal{W}^{-1}(\mathbf{V};\mathbf{S}_0,N_0) && (9)\\
&\sim \mathcal{N}\mathcal{W}^{-1}(\mathbf{M},\mathbf{K},\mathbf{S}_0,N_0) && (10)\\
\mathcal{N}(\mathbf{A};\mathbf{M},\mathbf{V},\mathbf{K}) &= \frac{|\mathbf{K}|^{d/2}}{|2\pi\mathbf{V}|^{m/2}}\exp(-\frac{1}{2}\mathrm{tr}((\mathbf{A}-\mathbf{M})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{A}-\mathbf{M})\mathbf{K})) && (11)\\
\mathcal{W}^{-1}(\mathbf{V};\mathbf{S}_0,n) &= \frac{1}{Z_{nd}\,|\mathbf{V}|^{(d+1)/2}}\left|\frac{\mathbf{V}^{-1}\mathbf{S}_0}{2}\right|^{n/2}\exp(-\frac{1}{2}\mathrm{tr}(\mathbf{V}^{-1}\mathbf{S}_0)) && (12)\\
\text{where } Z_{nd} &= \pi^{d(d-1)/4}\prod_{i=1}^{d}\Gamma((n+1-i)/2)
\end{aligned}$$

where $\mathbf{M}$ is d by m, $\mathbf{K}$ is m by m, and $\mathbf{S}_0$ is d by d. The density of $\mathbf{A}$ given $\mathbf{V}$ in (11) is known as the matrix-Normal distribution. The marginal for $\mathbf{A}$ without $\mathbf{V}$ is a matrix-T distribution:

$$\begin{aligned}
p(\mathbf{A}|\mathbf{X}) &\sim \mathcal{T}(\mathbf{M},\mathbf{S}_0,\mathbf{K},N_0+m) && (13)\\
\mathcal{T}(\mathbf{M},\mathbf{V},\mathbf{K},n) &= \frac{\prod_{i=1}^{d}\Gamma((n+1-i)/2)}{\prod_{i=1}^{d}\Gamma((n-m+1-i)/2)}\frac{|\mathbf{K}|^{d/2}}{|\pi\mathbf{V}|^{m/2}}\left|(\mathbf{A}-\mathbf{M})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{A}-\mathbf{M})\mathbf{K}+\mathbf{I}_m\right|^{-n/2} && (14)
\end{aligned}$$

Let

$$\begin{aligned}
\mathbf{S}_{xx} &= \mathbf{X}\mathbf{X}^{\mathrm{T}}+\mathbf{K} && (15)\\
\mathbf{S}_{yx} &= \mathbf{Y}\mathbf{X}^{\mathrm{T}}+\mathbf{M}\mathbf{K} && (16)\\
\mathbf{S}_{yy} &= \mathbf{Y}\mathbf{Y}^{\mathrm{T}}+\mathbf{M}\mathbf{K}\mathbf{M}^{\mathrm{T}} && (17)\\
\mathbf{S}_{y|x} &= \mathbf{S}_{yy}-\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{yx}^{\mathrm{T}} && (18)
\end{aligned}$$

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

$$\begin{aligned}
p(\mathbf{Y},\mathbf{A},\mathbf{V}|\mathbf{X}) &\propto |\mathbf{V}|^{(N+N_0+d+1)/2}\exp(-\frac{1}{2}\mathrm{tr}(\mathbf{V}^{-1}\left[\mathbf{A}\mathbf{S}_{xx}\mathbf{A}^{\mathrm{T}}-2\mathbf{S}_{yx}\mathbf{A}^{\mathrm{T}}+\mathbf{S}_{yy}+\mathbf{S}_0\right])) && (19)\\
&= |\mathbf{V}|^{(N+N_0+d+1)/2}\exp(-\frac{1}{2}\mathrm{tr}(\mathbf{V}^{-1}\left[(\mathbf{A}-\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1})\mathbf{S}_{xx}(\mathbf{A}-\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1})^{\mathrm{T}}+\mathbf{S}_{y|x}+\mathbf{S}_0\right])) && (20)
\end{aligned}$$

so the posterior is

$$p(\mathbf{A},\mathbf{V}|D)\sim\mathcal{N}\mathcal{W}^{-1}(\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1},\mathbf{S}_{xx},\mathbf{S}_{y|x}+\mathbf{S}_0,N+N_0) \quad (21)$$

from which inferences can be made. The prior used in this paper is approximately noninformative, with free parameters α and N_0:

$$p(\mathbf{A},\mathbf{V}|\mathbf{X})\sim\mathcal{N}\mathcal{W}^{-1}(\mathbf{0},\alpha\mathbf{X}\mathbf{X}^{\mathrm{T}},N_0\mathbf{I}_d,N_0) \quad (22)$$

As $(\alpha, N_0) \to 0$, this prior approaches the noninformative Jeffreys prior:

$$p(\mathbf{A}, \mathbf{V}|\mathbf{X}) \propto \left|\mathbf{X}\mathbf{X}^{\mathrm{T}}\right|^{d/2} |2\pi\mathbf{V}|^{-m/2} |\mathbf{V}|^{-(d+1)/2} \tag{23}$$

The Jeffreys prior is conditional on $\mathbf{X}$ in order to be invariant to input rescaling and is conditional on $\mathbf{V}$ in order to be invariant to output rescaling. Other priors, such as that used by MacKay (1992), are not invariant; inferences change when you rescale the input space. Unfortunately, the Jeffreys prior is improper so sometimes the limit can be taken and other times (α, N_0) must be left as parameters to be optimized or integrated out. This is the regression technique advocated by Gull (1988). Note that if $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ is singular then we must choose a different $\mathbf{K}$ matrix based on problem-specific knowledge. In the non-regression case, this prior reduces to

$$p(\mathbf{A}|\mathbf{V})p(\mathbf{V}) \sim \mathcal{MW}^{-1}(\mathbf{0}, \alpha N, N_0\mathbf{I}_d, N_0) \tag{24}$$

2 Known V

From (20) we see that the posterior for $\mathbf{A}$ given $\mathbf{V}$ is matrix-Normal:

$$\begin{aligned} p(\mathbf{A}|D, \mathbf{V}) &\sim \mathcal{M}(\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}, \mathbf{V}, \mathbf{S}_{xx}) && (25) \\ &\sim \mathcal{M}(\mathbf{Y}\mathbf{X}^{\mathrm{T}}(\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1}(\alpha+1)^{-1}, \mathbf{V}, \mathbf{X}\mathbf{X}^{\mathrm{T}}(\alpha+1)) && (26) \end{aligned}$$

For the non-regression model the posterior simplifies to $\mathcal{M}(\frac{\sum_i \mathbf{y}_i}{(\alpha+1)N}, \frac{\mathbf{V}}{(\alpha+1)N})$. The mode of the posterior differs from maximum-likelihood by the factor $(\alpha+1)^{-1}$, which provides some protection against overfitting.

2.1 Model selection via the evidence

Multiplying (7) times (11) and integrating out $\mathbf{A}$ gives the evidence for linearity, with $\mathbf{V}$ known:

$$\begin{aligned} p(\mathbf{Y}|\mathbf{X}, \mathbf{V}) &= \frac{|\mathbf{K}|^{d/2}}{|\mathbf{S}_{xx}|^{d/2} |2\pi\mathbf{V}|^{N/2}} \exp(-\frac{1}{2}\mathrm{tr}(\mathbf{V}^{-1}\mathbf{S}_{y|x})) && (27) \\ &\sim \mathcal{M}(\mathbf{M}\mathbf{X}, \mathbf{V}, \mathbf{I} - \mathbf{X}^{\mathrm{T}}\mathbf{S}_{xx}^{-1}\mathbf{X}) && (28) \\ &= \left(\frac{\alpha}{\alpha+1}\right)^{md/2} |2\pi\mathbf{V}|^{-N/2} \exp(-\frac{1}{2}\mathrm{tr}(\mathbf{V}^{-1}\mathbf{S}_{y|x})) && (29) \end{aligned}$$

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

$$\mathbf{S}_{y|x} = \mathbf{Y}(\mathbf{I} - \mathbf{X}^{\mathrm{T}}(\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1}\mathbf{X})\mathbf{Y}^{\mathrm{T}} \tag{30}$$

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 $\mathbf{I} - \mathbf{X}^{\mathrm{T}}(\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1}\mathbf{X}$ extracts

the component of $\mathbf{Y}$ which is orthogonal to the input. 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. So even though α is a free parameter, it doesn't contribute to overfitting.

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

$$\alpha = \frac{md}{\mathrm{tr}(\mathbf{V}^{-1}\mathbf{Y}\mathbf{X}^{\mathrm{T}}(\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1}\mathbf{X}\mathbf{Y}^{\mathrm{T}}) - md} \tag{31}$$

This estimator behaves in a reasonable way: when m increases or when the noise level increases, so does the amount of shrinkage, in order to reduce overfitting. But as N increases, the amount of shrinkage decreases, in order to let the data speak for themselves.

The evidence for linearity is useful for selecting among different linear models, viz. 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 (31) and (27) for each model and see which is largest. Figure 1 has an example of using this rule to select polynomial order. 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 vs. Legendre polynomials or Hermite polynomials (though for MacKay (1992), it did matter). The data is synthetic with $N = 50$ and known variance $\mathbf{V} = 10$.

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.

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

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

$$\bar{\mathbf{y}} = \frac{1}{N}\sum_i \mathbf{y}_i \tag{33}$$

$$\mathbf{S} = (\sum_i \mathbf{y}_i\mathbf{y}_i^{\mathrm{T}}) - \frac{N}{(\alpha+1)}\bar{\mathbf{y}}\bar{\mathbf{y}}^{\mathrm{T}} \tag{34}$$

$$= \mathbf{Y}(\mathbf{I} - \frac{1}{(\alpha+1)N}\mathbf{1}\mathbf{1}^{\mathrm{T}})\mathbf{Y}^{\mathrm{T}} \tag{35}$$

which incorporates shrinkage of the mean. The evidence is maximized when

$$\alpha = \frac{d}{N\bar{\mathbf{y}}^{\mathrm{T}}\mathbf{V}^{-1}\bar{\mathbf{y}} - d} \tag{36}$$

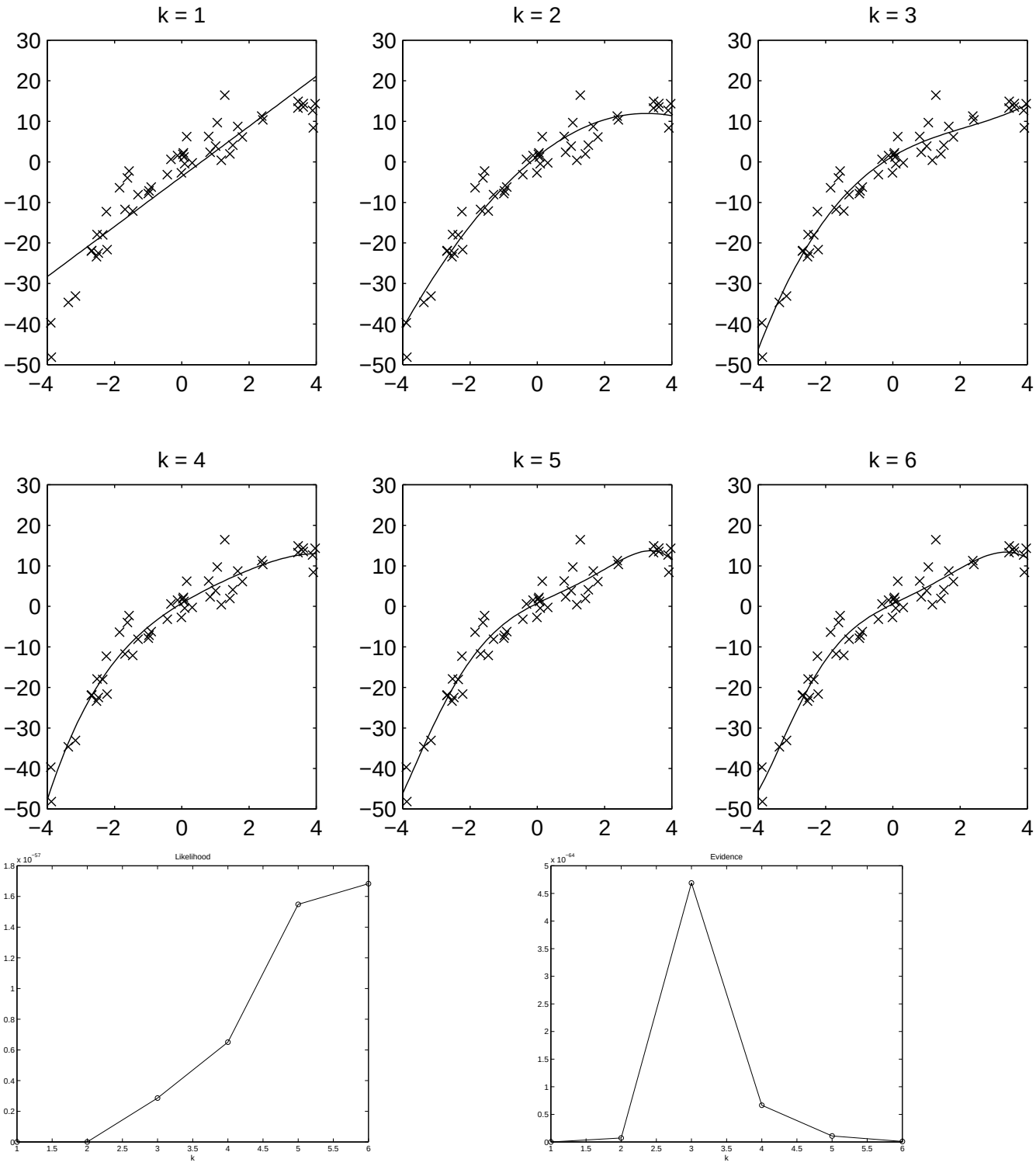

Figure 1: Example of using the evidence to select model order. A synthetic data set is approximated by polynomials of varying order. The likelihood curve always increases with increasing order while the evidence curve has a clear maximum at $k = 3$ (the true order in this case).

2.2 Predicting new outputs

To predict the $\mathbf{y}$ output for a new $\mathbf{x}$ input, consider the augmented data set $D' = \{\mathbf{y}, \mathbf{x}\} \cup D$. Then

$$\begin{aligned}
\mathbf{S}'_{xx} &= \mathbf{S}_{xx} + \mathbf{x}\mathbf{x}^{\mathrm{T}} && (37)\\
\mathbf{S}'_{y|x} &= \mathbf{S}'_{yy} - \mathbf{S}'_{yx}(\mathbf{S}'_{xx})^{-1}(\mathbf{S}'_{yx})^{\mathrm{T}} && (38)\\
&= \mathbf{S}_{yy} - \mathbf{S}_{yx}(\mathbf{S}'_{xx})^{-1}\mathbf{S}_{yx}^{\mathrm{T}} \\
&\quad + \mathbf{y}\mathbf{y}^{\mathrm{T}} - \mathbf{y}\mathbf{x}^{\mathrm{T}}(\mathbf{S}'_{xx})^{-1}\mathbf{x}\mathbf{y}^{\mathrm{T}} - \mathbf{y}\mathbf{x}^{\mathrm{T}}(\mathbf{S}'_{xx})^{-1}\mathbf{S}_{yx}^{\mathrm{T}} - \mathbf{S}_{yx}(\mathbf{S}'_{xx})^{-1}\mathbf{x}\mathbf{y}^{\mathrm{T}} && (39)\\
&= \mathbf{S}_{y|x} + (\mathbf{y} - \mathbf{S}_{yx}(\mathbf{S}'_{xx})^{-1}\mathbf{x}c^{-1})c(\mathbf{y} - \mathbf{S}_{yx}(\mathbf{S}'_{xx})^{-1}\mathbf{x}c^{-1})^{\mathrm{T}} && (40)\\
&= \mathbf{S}_{y|x} + (\mathbf{y} - \mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{x})c(\mathbf{y} - \mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{x})^{\mathrm{T}} && (41)\\
c &= 1 - \mathbf{x}^{\mathrm{T}}(\mathbf{S}'_{xx})^{-1}\mathbf{x} = (1 + \mathbf{x}^{\mathrm{T}}\mathbf{S}_{xx}^{-1}\mathbf{x})^{-1} && (42)
\end{aligned}$$

The invariant prior is now conditional on the augmented $\mathbf{X}$. So

$$\begin{aligned}
p(\mathbf{y}|\mathbf{x}, D, \mathbf{V}) &= p(D'|\mathbf{V})/p(D|\mathbf{V}) && (43)\\
&= \frac{|2\pi\mathbf{V}|^{N/2}\,|\mathbf{S}_{xx}|^{d/2}}{|2\pi\mathbf{V}|^{(N+1)/2}\,|\mathbf{S}'_{xx}|^{d/2}}\exp(-\frac{1}{2}\mathrm{tr}(\mathbf{V}^{-1}(\mathbf{S}'_{y|x} - \mathbf{S}_{y|x})) && (44)\\
&= \frac{1}{|2\pi\mathbf{V}c^{-1}|^{1/2}}\exp(-\frac{c}{2}(\mathbf{y} - \mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{x})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{x})) && (45)\\
&= \mathcal{N}(\mathbf{y}; \mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{x}, \mathbf{V}c^{-1}) && (46)
\end{aligned}$$

Even though we integrated out $\mathbf{A}$ to get this result, the expected value of $\mathbf{y}$ given $\mathbf{x}$ is identical to substituting the posterior mode for $\mathbf{A}$. This makes sense because $E[\mathbf{y}|\mathbf{x}, D] = E[\mathbf{A}|D]\mathbf{x}$. 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 $c^{-1} = \frac{N+1}{N}$ so the predictive density is

$$p(\mathbf{y}|D) \sim \mathcal{N}(\bar{\mathbf{y}}, \frac{N+1}{N}\mathbf{V}) \tag{47}$$

which again incorporates wiggle of the unknown mean.

For predicting K new samples $(\mathbf{Y}', \mathbf{X}')$, we use

$$\begin{aligned}
D' &= \{\mathbf{Y}', \mathbf{X}'\} \cup D && (48)\\
\mathbf{S}'_{xx} &= \mathbf{S}_{xx} + \mathbf{X}'(\mathbf{X}')^{\mathrm{T}} && (49)\\
\mathbf{S}'_{y|x} &= \mathbf{S}'_{yy} - \mathbf{S}'_{yx}(\mathbf{S}'_{xx})^{-1}(\mathbf{S}'_{yx})^{\mathrm{T}} && (50)
\end{aligned}$$

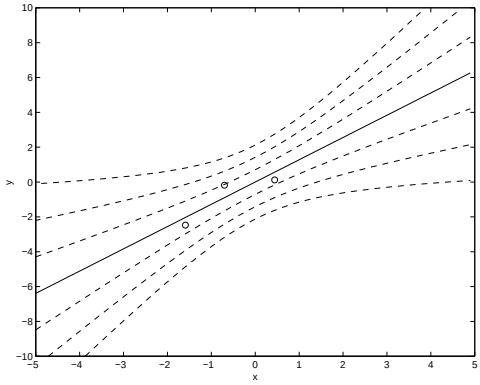

Figure 2: Contours of the predictive density for y, given three training points (circles). The model is $y = ax$ (a line through the origin).

$$
\begin{aligned}
& = \mathbf{S}_{y|x} + (\mathbf{Y}' - \mathbf{S}_{yx}(\mathbf{S}'_{xx})^{-1}\mathbf{X}'\mathbf{C}^{-1})\mathbf{C}(\mathbf{Y}' - \mathbf{S}_{yx}(\mathbf{S}'_{xx})^{-1}\mathbf{X}'\mathbf{C}^{-1})^{\mathrm{T}} && (51)\\
& = \mathbf{S}_{y|x} + (\mathbf{Y}' - \mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{X}')\mathbf{C}(\mathbf{Y}' - \mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{X}')^{\mathrm{T}} && (52)\\
\mathbf{C} &= \mathbf{I}_K - (\mathbf{X}')^{\mathrm{T}}(\mathbf{S}'_{xx})^{-1}\mathbf{X}' = (\mathbf{I}_K + (\mathbf{X}')^{\mathrm{T}}\mathbf{S}_{xx}^{-1}\mathbf{X}')^{-1} && (53)\\
p(\mathbf{Y}'|\mathbf{X}', D, \mathbf{V}) &= p(D'|\mathbf{V})/p(D|\mathbf{V}) && (54)\\
& = \frac{|2\pi\mathbf{V}|^{N/2}\,|\mathbf{S}_{xx}|^{d/2}}{|2\pi\mathbf{V}|^{(N+K)/2}\,|\mathbf{S}'_{xx}|^{d/2}}\exp(-\frac{1}{2}\mathrm{tr}(\mathbf{S}'_{y|x} - \mathbf{S}_{y|x})) && (55)\\
& = \frac{|\mathbf{C}|^{d/2}}{|2\pi\mathbf{V}|^{K/2}}\exp(-\frac{1}{2}\mathrm{tr}((\mathbf{Y}' - \mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{X}')^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{Y}' - \mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{X}')\mathbf{C})) && (56)\\
& \sim \mathcal{N}(\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{X}', \mathbf{V}, \mathbf{C}) && (57)
\end{aligned}
$$

which is equivalent to (27) after folding D into the prior. In the non-regression case, we have

$$
\begin{aligned}
\mathbf{S}' &= \mathbf{S} + (\mathbf{Y}' - \bar{\mathbf{y}}\mathbf{1}^{\mathrm{T}})\mathbf{C}(\mathbf{Y}' - \bar{\mathbf{y}}\mathbf{1}^{\mathrm{T}})^{\mathrm{T}} && (58)\\
\mathbf{C} &= \mathbf{I}_K - \mathbf{1}\mathbf{1}^{\mathrm{T}}/(N+K) && (59)\\
p(\mathbf{Y}'|\mathbf{Y}, \mathbf{V}) &= p(\mathbf{Y}', \mathbf{Y}|\mathbf{V})/p(\mathbf{Y}|\mathbf{V}) && (60)\\
& = \frac{1}{|2\pi\mathbf{V}|^{K/2}}\left(\frac{N}{N+K}\right)^{d/2}\exp(-\frac{1}{2}\mathrm{tr}((\mathbf{Y}' - \bar{\mathbf{y}}\mathbf{1}^{\mathrm{T}})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{Y}' - \bar{\mathbf{y}}\mathbf{1}^{\mathrm{T}})\mathbf{C})) && (61)\\
& \sim \mathcal{N}(\bar{\mathbf{y}}\mathbf{1}^{\mathrm{T}}, \mathbf{V}, \mathbf{C}) && (62)
\end{aligned}
$$

3 Unknown V

From (27) we see that the posterior for $\mathbf{V}$ is inverse Wishart:

$$p(\mathbf{V}|D) \sim \mathcal{W}^{-1}(\mathbf{S}_{y|x} + \mathbf{S}_0, N) \tag{63}$$

Integrating $\mathbf{A}$ and $\mathbf{V}$ out of (20), or equivalently dividing out (21), gives the evidence for linearity:

$$\begin{aligned}
p(\mathbf{Y}|\mathbf{X}) &= \frac{Z_{(N+N_0)d}}{Z_{N_0 d}} \frac{|\mathbf{K}|^{d/2}}{|\mathbf{S}_{xx}|^{d/2}} \frac{|\mathbf{S}_0|^{N_0/2}}{\pi^{Nd/2} \left|\mathbf{S}_{y|x} + \mathbf{S}_0\right|^{(N+N_0)/2}} && (64)\\
&\sim \mathcal{T}(\mathbf{MX}, \mathbf{S}_0, \mathbf{I} - \mathbf{X}^{\mathrm{T}}\mathbf{S}_{xx}^{-1}\mathbf{X}, N + N_0) && (65)\\
&= \frac{\prod_{i=1}^{d} \Gamma((N + N_0 + 1 - i)/2)}{\prod_{i=1}^{d} \Gamma((N_0 + 1 - i)/2)} \left(\frac{\alpha}{\alpha + 1}\right)^{md/2} (\pi N_0)^{-Nd/2} \left|\frac{\mathbf{S}_{y|x}}{N_0} + \mathbf{I}_d\right|^{-(N+N_0)/2} && (66)
\end{aligned}$$

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

$$\begin{aligned}
\hat{\mathbf{V}} &= (\mathbf{S}_{y|x} + N_0\mathbf{I}_d)/(N + N_0) && (67)\\
\alpha &= \frac{md}{\mathrm{tr}(\hat{\mathbf{V}}^{-1}\mathbf{Y}\mathbf{X}^{\mathrm{T}}(\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1}\mathbf{X}\mathbf{Y}^{\mathrm{T}}) - md} && (68)\\
N_0^{new} &= N_0 \frac{\sum_{i=1}^{d} \Psi((N + N_0 + 1 - i)/2) - \Psi((N_0 + 1 - i)/2)}{\log\left|\frac{\mathbf{S}_{y|x}}{N_0} + \mathbf{I}_d\right| + \mathrm{tr}(\hat{\mathbf{V}}^{-1}) - d} && (69)
\end{aligned}$$

As $N_0 \to 0$, the posterior predictive distribution is

$$\begin{aligned}
p(\mathbf{y}|\mathbf{x}, D) &= p(D')/p(D) && (70)\\
&= \frac{Z_{(N+1)d} \left|\mathbf{S}_{xx}\right|^{d/2} \left|\pi\mathbf{S}_{y|x}\right|^{N/2}}{Z_{Nd} \left|\mathbf{S}'_{xx}\right|^{d/2} \left|\pi\mathbf{S}'_{y|x}\right|^{(N+1)/2}} && (71)\\
&= \frac{\Gamma((N+1)/2)}{\Gamma((N+1-d)/2)} \left|\pi\mathbf{S}_{y|x}c^{-1}\right|^{-1/2} \\
&\quad \left((\mathbf{y} - \mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{x})^{\mathrm{T}}\mathbf{S}_{y|x}^{-1}c(\mathbf{y} - \mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{x}) + 1\right)^{-(N+1)/2} && (72)\\
&\sim \mathcal{T}(\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{x}, \mathbf{S}_{y|x}c^{-1}, N + 1) && (73)
\end{aligned}$$

In the non-regression case, we had $c^{-1} = \frac{N+1}{N}$ so

$$p(\mathbf{y}|D) \sim \mathcal{T}(\bar{\mathbf{y}}, \frac{N+1}{N}\mathbf{S}, N + 1) \tag{74}$$

To predict K new samples, integrate $\mathbf{V}$ out of (56) times (63) to get

$$p(\mathbf{Y}'|\mathbf{X}', D) \sim \mathcal{T}(\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{X}', \mathbf{S}_{y|x}, \mathbf{C}, N+K) \tag{75}$$

which is equivalent to (64) after folding D into the prior. In the non-regression case, this is

$$p(\mathbf{Y}'|D) \sim \mathcal{T}(\bar{\mathbf{y}}\mathbf{1}^{\mathrm{T}}, \mathbf{S}, \mathbf{C}, N+K) \tag{76}$$

4 Piecewise regression

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

$$p(\mathbf{y}_i|\mathbf{x}_i, \mathbf{A}_1, \mathbf{V}_1, \mathbf{A}_2, \mathbf{V}_2) \sim \begin{cases} \mathcal{N}(\mathbf{A}_1\mathbf{x}_i, \mathbf{V}_1) & \text{if } i < t \\ \mathcal{N}(\mathbf{A}_2\mathbf{x}_i, \mathbf{V}_2) & \text{if } i \geq t \end{cases} \tag{77}$$

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

$$p(t|D) \propto p(t)p(\mathbf{y}_1..\mathbf{y}_{t-1}|\mathbf{x}_1..\mathbf{x}_{t-1})p(\mathbf{y}_t..\mathbf{y}_N|\mathbf{x}_t..\mathbf{x}_N) \tag{78}$$

where the last two terms are given by separate applications of (64). The normalizing constant is the evidence for the existence of a changepoint:

$$p(D|\text{changepoint}) = \sum_{t=1}^{N} p(t)p(\mathbf{y}_1..\mathbf{y}_{t-1}|\mathbf{x}_1..\mathbf{x}_{t-1})p(\mathbf{y}_t..\mathbf{y}_N|\mathbf{x}_t..\mathbf{x}_N) \tag{79}$$

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 ((79) divided by (64)) 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 (64).

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

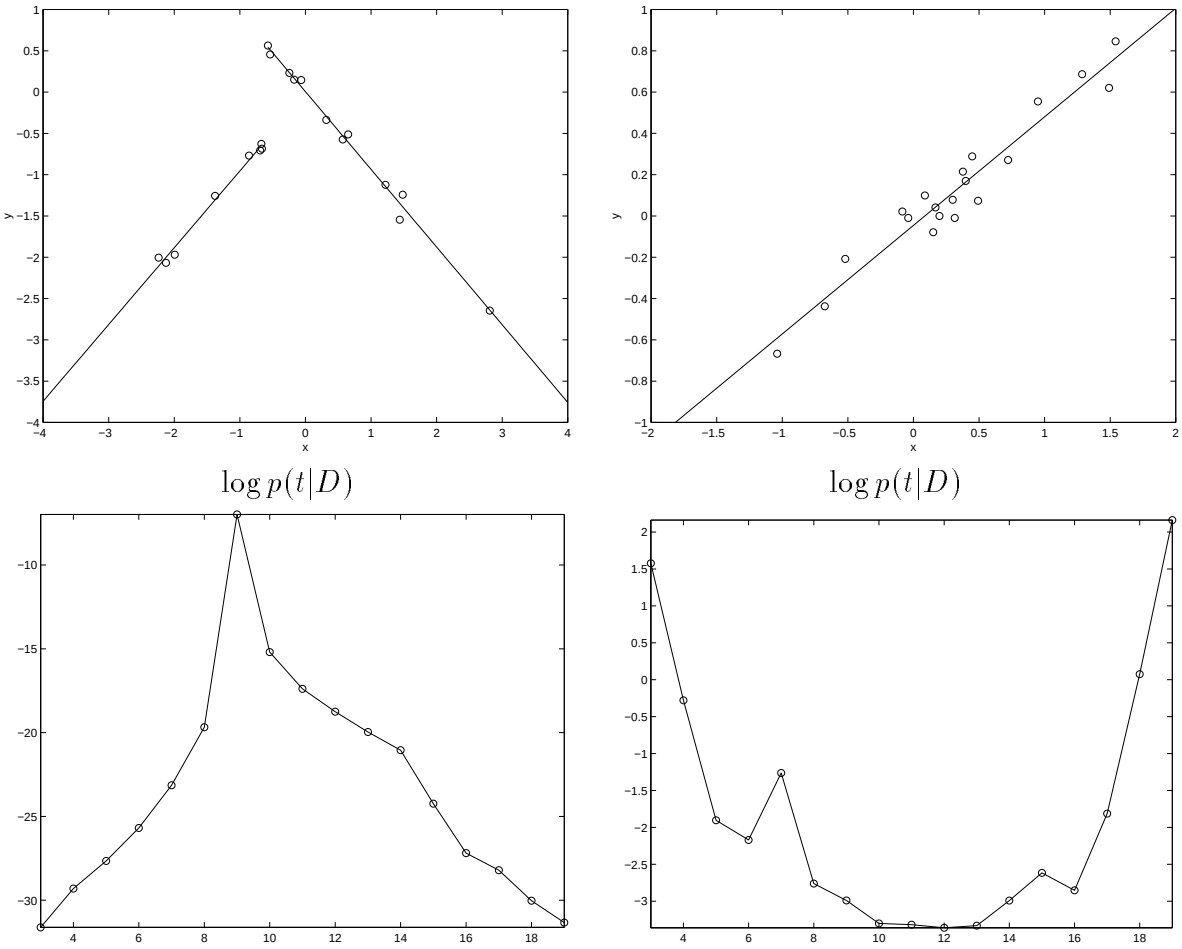

Figure 3: Example of changepoint analysis. On the left, the changepoint at $t = 9$ is correctly found. On the right, there is no changepoint.

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 5.

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).

5 Basis function regression

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

$$x_i = f_i(\mathbf{z}) \tag{80}$$

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.

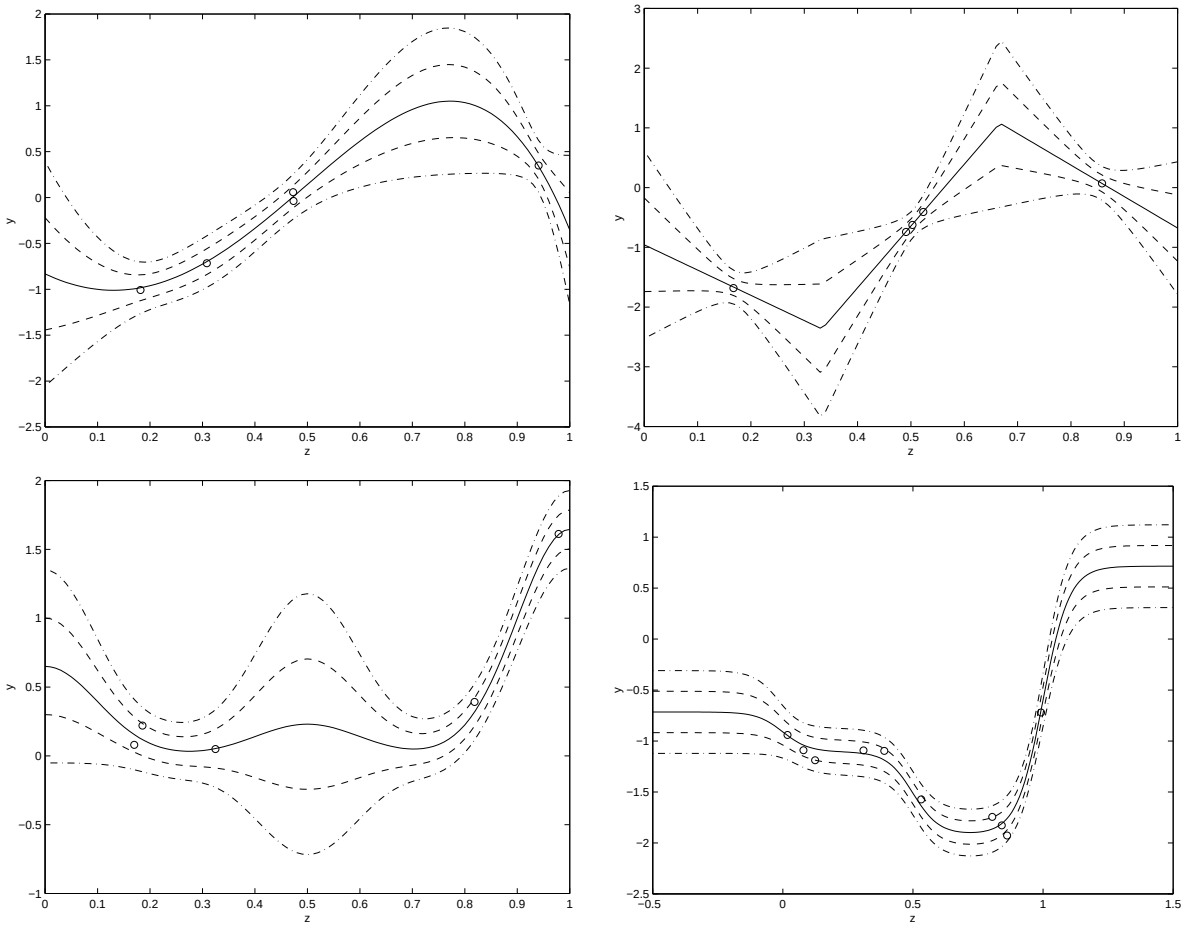

Figure 4: Example of basis function regression. The contours of the predictive density are shown for a polynomial basis, a piecewise linear basis, a Gaussian basis, and a tanh basis.

The evidence formula (64) 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. MLPs are usually trained via least-squares, without marginalizing over $\mathbf{A}$, which can lead to overfitting. Training the hidden units via the evidence formula instead of least-squares should help avoid this. Indeed, some of the modified training rules proposed in the literature have this flavor.

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.

[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.