OPTIMAL GLOBAL ERROR MEASURE APPROACH TO RISK REDUCTION IN MODERN REGRESSION

A Thesis

Submitted to the Faculty

of

Purdue University

by

Hong Pan

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

May 1999

To My Parents

ACKNOWLEDGMENTS

I would like to thank my advisor, Professor Vwani Roychowdhury, for his unending patience, constant encouragement, and financial support. I would also like to thank the Department of Electrical Engineering in the University of California at Los Angeles for their hospitality. I am most grateful to everyone who has encouraged and helped me to get through my education and research endeavors.

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

Pan, Hong, Ph.D., Purdue University, May, 1999. Optimal Global Error Measure Approach to Risk Reduction in Modern Regression. Major Professor: Vwani P. Roychowdhury.

We first review the concepts fundamental to the statistical inference procedures using nonparametric regression models. The global error properties of an estimator over its parameter space are employed to define a general framework that puts various existing optimality criteria and heuristics into a coherent and rigorous perspective. A class of Bayes robust and asymptotically minimax estimator is then constructed by comprehensively considering all the major aspects of their global error measures. This new estimator is shown to have a better risk behavior than the usual Least Squares and other Bayesian procedures, and to be robust with respect to misspecification of the prior assumption on the parameters, among several other desirable properties. Moreover, the related single-run algorithm does not incur extra computational cost, while delivering improved risk performance. As a case study, the prediction performance of the new widely applicable and well-balanced estimation procedure is then evaluated and compared critically on a class of generalized additive regression method, i.e., the feedforward neural network model.

1. Introduction

1.1 Trends in Regression Analysis

1.1.1 Basics of a regression model

The essential part of data analysis is to study various types of relationship between two random variables, a *response variable* Y and an *explanatory variable* (a.k.a. predictor variable, regressor variable) X, based on a training sample of size n taken from a sample space $(\mathcal{X}, \mathcal{Y})$ according to an unknown joint distribution $\mathcal{F}$. A regression model is a statistical tool for summarizing the dependence of the expectation of Y on X, $E(Y|X)$, as a real-valued function of X, say, $f(X)$ so that $f(x) = E(Y|X = x)$ for x in a real-valued interval. In most of applications, one typically has more than one predictor variable in hand, i.e., X is vector-valued as $\boldsymbol{X} = (X_1, ..., X_d)'$. In general, the unknown true conditional expectation $f^* : \mathcal{R}^d \to \mathcal{R}$ is only assumed to be Borel measurable (i.e., $f^* \in \mathcal{B}$). For modeling the response function f^*, one usually starts with a specific class of functions $\mathcal{C}$, a much smaller subset of the all-possible Borel measurable functions in $\mathcal{B}$. For instance, $\mathcal{C}$ is chosen to be the class of linear regression functions $\mathcal{C}_{LM}$ in the form

$$f(\boldsymbol{x}; \boldsymbol{\theta}) = f(x_1, ..., x_d; a, \boldsymbol{b}) = a + \sum_{i=1}^{d} b_i x_i \ , \tag{1.1}$$

or a class of generalized additive models $\mathcal{C}_{GAM}$ in the form

$$f(\boldsymbol{x}; \boldsymbol{\theta}) = f(x_1, ..., x_d; \boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{k=1}^{h} \beta_k g_k(\sum_{i=1}^{d} \alpha_{ki} x_i) \ , \tag{1.2}$$

with certain nonlinear function $g_.(\cdot)$. In both cases, the problem of determining an unknown function is converted into the identification of an unknown parameter vector $\boldsymbol{\theta}$ from a parameter space $\boldsymbol{\Theta}$. Once the presumed class is chosen, the first step in a regression analysis is then to select a function $\hat{f}_n(\boldsymbol{x}) = f(\boldsymbol{x}; \hat{\boldsymbol{\theta}}(D_n))$ from $\mathcal{C}$ with relatively small error measure according to the data $D_n = \{(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_n, y_n)\}$. This problem of determining a suitable *estimate* $\hat{f}_n$ will be the focus throughout this thesis.

Error measure: Goodness-of-fit

In principle, a function from any chosen class, $f(\boldsymbol{X}; \boldsymbol{\theta})$, is defined over both the sample space $\mathcal{X}$ and the parameter space $\boldsymbol{\Theta}$. The specific value of $\boldsymbol{\theta}$ in $\boldsymbol{\Theta}$ and therefore the specific form of f at $\boldsymbol{\theta}$ in $\mathcal{C}$ as the *estimands* remain to be identified in a *point estimation* procedure. Suppose now that the unknown true response function $f = f(\boldsymbol{x}; \boldsymbol{\theta}) = E(Y|\boldsymbol{X} = \boldsymbol{x})$, and is in the class one chooses. From an abstract random observation $S_n = \{(\boldsymbol{X}_1, Y_1), ..., (\boldsymbol{X}_n, Y_n)\}$, an *estimator* of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}(S_n)$, is selected so that $\hat{f} = f(\boldsymbol{x}; \hat{\boldsymbol{\theta}}(S_n))$ is to be close to f. The value $\hat{\boldsymbol{\theta}}(D_n)$ and the resulting $\hat{f}_n = f(\boldsymbol{x}; \hat{\boldsymbol{\theta}}(D_n))$ as the realizations of the estimator at the observed data set D_n are called the *estimates* of $\boldsymbol{\theta}$ and f. Clearly, the estimator $\hat{\boldsymbol{\theta}}(S_n)$ is a random variable defined over the sample space $(\mathcal{X}, \mathcal{Y})$ itself. Because of this fact, an appropriate measure of the closeness of an estimator to f shall be taken in certain average sense. For example, if one chooses a quadratic *loss function*

$$L(\hat{f}, f) = L(\hat{\boldsymbol{\theta}}(S_n), \boldsymbol{\theta}) = [f(\boldsymbol{x}; \boldsymbol{\theta}) - f(\boldsymbol{x}; \hat{\boldsymbol{\theta}}(S_n))]^2 \tag{1.3}$$

to measure the lack-of-closeness of an estimator to its target, then a reasonable estimator $\hat{\boldsymbol{\theta}}(S_n)$ should be the one that minimize the expected loss, namely the *risk function*

$$R(\hat{f}, f) = R(\hat{\boldsymbol{\theta}}(S_n), \boldsymbol{\theta}) = E_{\mathcal{F}} L(\hat{f}, f) = \int [f(\boldsymbol{x}; \boldsymbol{\theta}) - f(\boldsymbol{x}; \hat{\boldsymbol{\theta}}(S_n))]^2 d\mathcal{F} \ . \tag{1.4}$$

After taking the expectation over the whole sample space according to the joint distribution $\mathcal{F}$, the risk function is no longer a random variable but a real-valued function of the parameter vector. In practice, the corresponding sampling version of (1.4), the *empirical risk* (a.k.a. average squared residual (ASR))

$$R_n = \frac{1}{n} \sum_{t=1}^{n} [y_t - f(\boldsymbol{x}_t; \hat{\boldsymbol{\theta}}(D_n))]^2 \ , \tag{1.5}$$

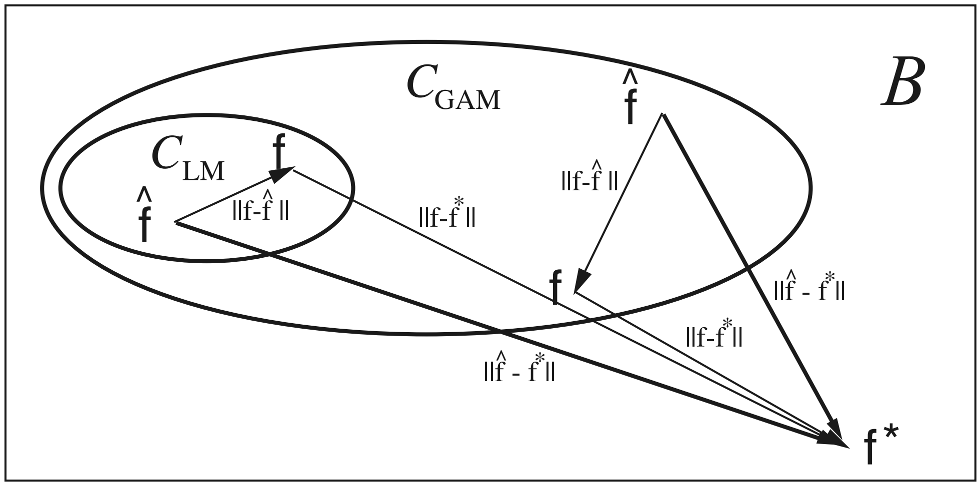


Fig. 1.1. The triangle decomposition of a risk function $||\hat{f} - f^*||_2$: approximation error $||f - f^*||_2$ and estimation error $||f - \hat{f}||_2$.

as an estimate of the risk, is used to serve as the error measure. However, there are no uniformly best estimators that minimize the risk for all values of $\boldsymbol{\theta}$ in the parameter space under (1.4), provided the target function is not constant (see [1], p.5). Clearly, additional optimality criteria are needed to help specify a uniquely determined estimator or substantially smaller subclass of estimators with desirable statistical properties. In this thesis, we shall show that a thorough analysis on the behavior of the risk function over parameter space of a regression model can provide the road map and mathematical machinery needed for this purpose.

One question remains to be answered before we get further into the statistical procedure of point estimation. In the light of data, what is the theoretical ground on which a specific class of regression model is chosen? If, in general, the unknown true response function f^* is not in the class of models one has chosen, the risk function of the selected estimator $\hat{f}$ to the true one f^* (denoted as an L_2 norm $||\hat{f} - f^*||_2$ for quadratic loss) can be decomposed into two parts (see Figure 1.1) in a triangle inequality

$$||\hat{f} - f^*||_2 \leq ||f - f^*||_2 + ||f - \hat{f}||_2 \, , \qquad \forall f \in \mathcal{C} \, . \tag{1.6}$$

The first part, $||f - f^*||_2$, is due to the approximation error originated from the limited capacity of the selected class of functions f's $\in \mathcal{C}$. For example, evidently, a large approximation error is expected when the class of linear models in (1.1) is used to fit a nonlinear relation (see Figures 1.2 and 1.3). Suppose that f is the best possible choice out of the whole class $\mathcal{C}$ such that $||f - f^*||_2$ is minimized if f^* is not in $\mathcal{C}$ and $||f - f^*||_2 = 0$ if $f^* \in \mathcal{C}$. The second part, $||f - \hat{f}||_2$, is the estimation error owing to the limited knowledge of f^* obtained from the finite-sized sample S_n. Obviously, these two parts are only related through one's choice of $\mathcal{C}$, with the first part completely determined by the choice of $\mathcal{C}$. To reduce the overall risk, one must first make an assessment on the capacities of various available regression models.

1.1.2 From parametrics to nonparametrics

There are mainly two classes of regression models: *parametrics* and *nonparametrics*. To use the definition given in [2], an estimator $\hat{f}$ is said to be parametric if $\hat{f} \in \mathcal{C}$ where $\mathcal{C}$ is a collection of Borel measurable functions which can be defined in terms of a finite number of unknown parameters. Otherwise, the estimator $\hat{f}$ is said to be nonparametric.

Parametrics

The most commonly used parametric regression function is the multiple linear regression model in (1.1). The number of parameters in (1.1), $(d+1)$, is predetermined and finite because it only depends on the dimensionality of $\boldsymbol{X}$. In applied statistics, there are many other commonly used parametric regression models that depend on their parameters in nonlinear fashion. For instance, the class of quadratic regression models in the form

$$f(\boldsymbol{x};\boldsymbol{\theta}) = a + \sum_{i=1}^{d} b_i x_i + \sum_{i=1}^{d}\sum_{j=1}^{d} c_{ij} x_i x_j \ ,$$

takes into account of interaction effects among predictors; the class of additive regression model

$$f(\boldsymbol{x};\boldsymbol{\theta}) = a + \sum_{i=1}^{d} g_i(x_i) \ ,$$

with presumed forms for g_i's, is used in analysis of variance; and the hazard model in the form

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \alpha x_0^{\alpha-1} \exp(\alpha \sum_{i=1}^{d} \beta_i x_i) \ ,$$

plays a central role in survival analysis. Parametrics generally have certain advantages: their parameters usually bear some physical meanings which means better interpretability; their exact and explicit formulations make mathematical analysis more tractable; and their statistical estimation procedure is usually efficient, which is one of several additional optimality criteria needed to identify a desirable estimator. In spite of these positive factors, any specific *a priori* formulation assumed for regression may not be adequate for modeling the underlying response surface determined by an arbitrary distribution $\mathcal{F}$, especially when one has no precise knowledge about the form and class of the true response surface. In the terminology of approximation theory, a parametric model cannot serve as a *universal approximator* to an unknown response function f^*. The approximation capacity of parametrics is severely limited in the sense that their approximation error can be arbitrarily large so that the overall risk cannot be reduced, regardless of how many samples might be available and how low the estimation error might be made (see Figures 1.1, 1.2, 1.3 and 1.4). A further look at the bias-variance decomposition of the risk function in the next section will shed more light on this issue (see Figure 1.5).

Nonparametrics

On the other hand, instead of an assumed parametric form, a nonparametric regression model is only defined as an element of some infinite dimensional function space with certain smoothness properties. The generalized additive model in (1.2) is nonparametric, for the number of the additive terms, h, can be any positive integer so that the total number of parameters, $h(d+1)$, is arbitrary. This interesting class of model arises in feedforward neural network regression and projection pursuit regression among several others. With a feedforward neural network, for instance, the $g_k(\cdot)$'s take the logistic (or sigmoidal) form, $g_{\cdot}(u) = e^u/(1+e^u)$, the $\boldsymbol{x}_i$'s are

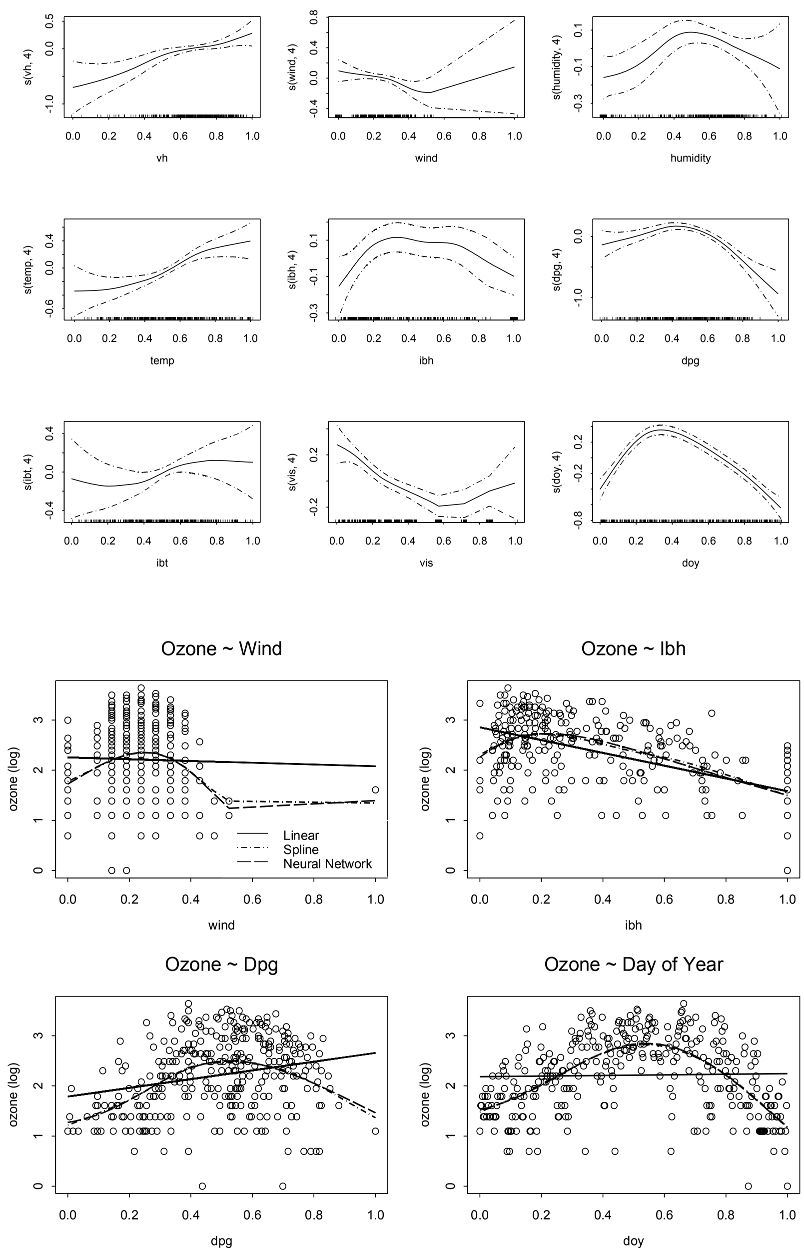
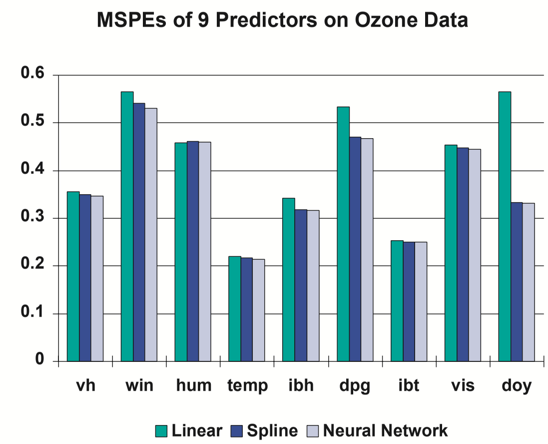


Fig. 1.2. Two commonly used nonparametric regression models, the splines and the neural networks, are used to analyze the ozone data (see Figure A.1 in Appendix A), and compared with the linear regression model. The upper plot shows the separate relations between the ozone measurement and its nine predictors, summarized by a simple fit using univariate spline with pointwise standard-error curve attached. It also shows that the predictors `doy`, `dpg`, `wind` and `ibh` appear convincingly nonlinear. The lower plot shows the estimated response functions for these four particular predictors by these three methods. The nonparametrics provide better fits which lead to significant performance gain, as shown in Figure 1.3.



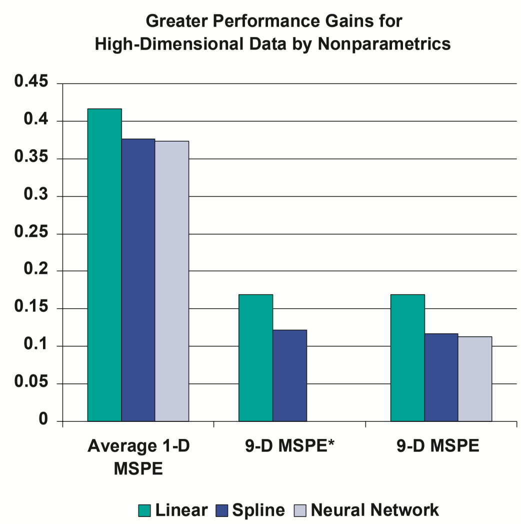


Fig. 1.3. The upper bar plot compares mean squared prediction error (MSPE) for each of the nine one-dimensional predictors. It also shows that the nonparametrics gain significant improvements on the four predictors with evident nonlinearity. The lower bar plot shows the nonparametrics gaining even greater improvement when all nine predictor variables are used to form a multivariate regression model for predicting the response variable. While for one-dimensional predictors the average MSPEs of the spline and the neural network drop 9.44% and 10.30% from the MSPE of the linear regression respectively, the performance gains for the nine-dimensional regression models are increased to 31.10% and 46.56% by each. The middle bars in the lower plot labeled '9-D MSPE*' show that the spline with four apparently nonlinear predictors in nonlinear terms and the rest in linear terms manages a 28.16% improvement over the linear model, which constitutes 90% of performance gain by a full-scale spline shown in the same plot. [Note: The .632 bootstrap estimates of MSPE are used with 1000 resamples for each case. The smoothing additive cubic splines are used to represent the spline method, which has the best performance among several other spline methods reported in [3]. The neural networks with $h = 9$ (i.e., nine hidden units) and no skip layer are also regulated with the decay parameter set to 0.01 for $d = 9$ and 0.025 for $d = 1$.]

rescaled in the d-dimensional unit cube $I = [0,1]^d \in \mathcal{R}^d$, and the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta})' \in \mathcal{R}^q$ with $q = h(d+1)$. If the number of additive terms, h, is permitted to be sufficiently large, the regression model (1.2) is capable of approximating any continuous response surface to any desired degree of accuracy [4, 5, 6, 7, 8]. To be specific, if the target response function f^* is continuously differentiable and the gradient of its Fourier transformation is integrable, then the rate of convergence to zero of the approximation error utilized by the model in (1.2) to f^* is $\mathcal{O}(1/\sqrt{h})$ in a L_2 norm on compacta I in $\mathcal{R}^d$. The resulting response functional form is more heavily relied on and potentially, therefore, more appropriately determined by the given set of data (see Figures 1.2, 1.3, 1.4 and 1.5).

1.1.3 Evolution of nonparametrics

Probability density estimation

Nonparametric regression has its roots in probability density estimation back in the Sixties [9, 10]. The research and development of nonparametric regression methods has been intensified considerably since then, with a huge body of literature mainly devoted to two large classes of conventional nonparametrics called *kernel* and *spline* methods. Indeed, some very basic statistics like the *histogram* can be seen as nonparametrics. For instance, in the one-dimensional case, to estimate an unknown density function $p(x)$ from a sample $\{X_t\}_{t=1}^n$, one first divides the real line into bins

$$B_k = [x_0 + (k-1)h,\ x_0 + kh)\ ,$$

with h the binwidth and x_0 the origin, and count how many data points fall into each bin. The histogram is then defined by

$$\begin{aligned} \hat{p}(x) &= \frac{1}{nh} \sum_{t=1}^{n} \sum_{k} I(X_t \in B_k) I(x \in B_k) \\ &= \frac{1}{n} \cdot \frac{(\text{number of } X_t \text{ in the same bin } B_k \text{ as } x)}{(\text{width of bin containing } x)}\ , \end{aligned} \tag{1.7}$$

with $I(\cdot)$ the indicator function. It is natural to take one step further by defining a kernel function on every data point, so that the averaging of the kernels leads to the

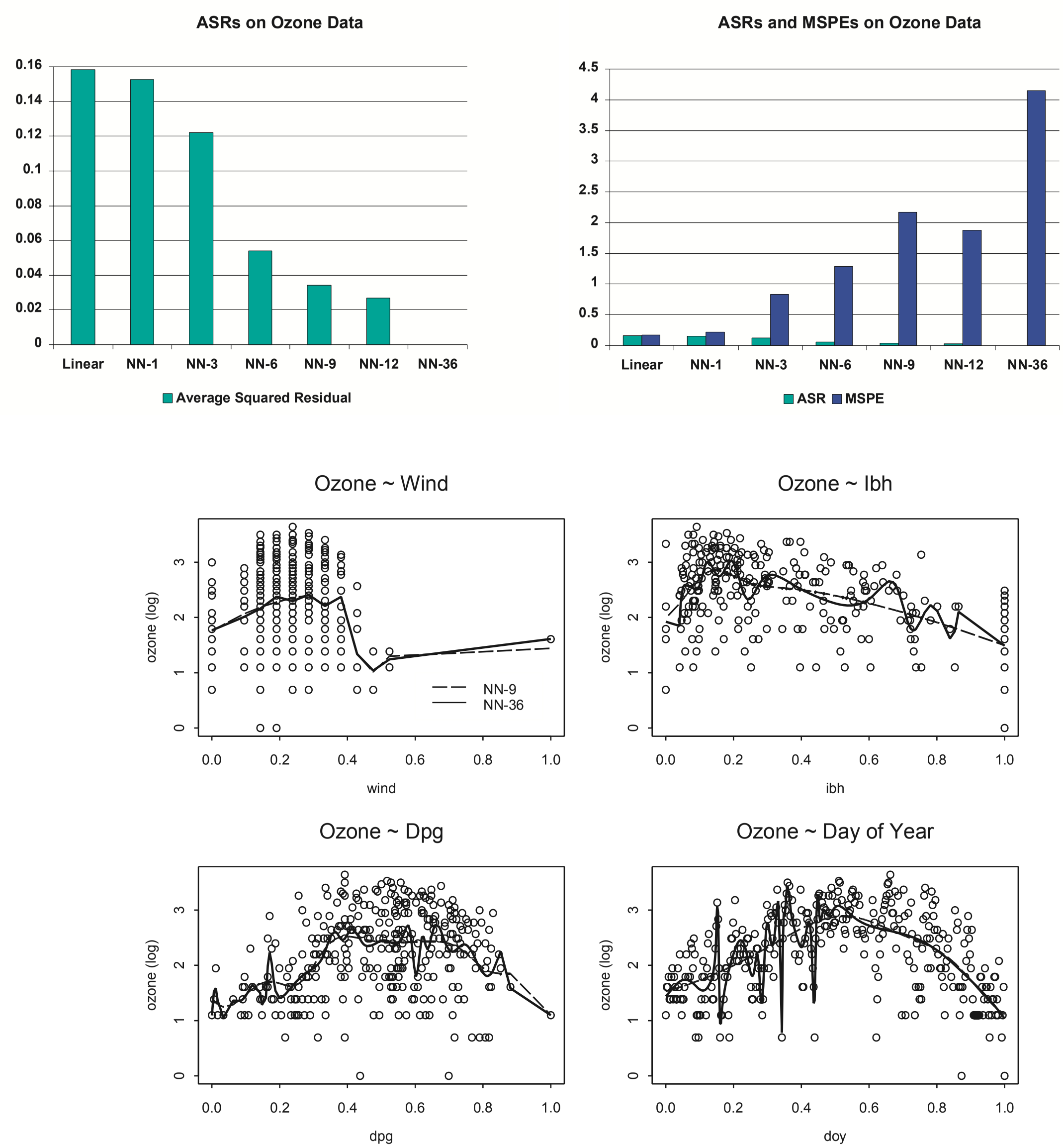


Fig. 1.4. The empirical risks (ASRs) and empirical prediction risks (MSPEs) of multivariate linear regression and six neural network models with $h = 1, 3, 6, 9, 12, 36$ and no skip layers on nine-dimensional ozone data. The average squared residual of the neural network with 36 hidden units is virtually reduced to zero at 2.981857×10^{-7}. However, overfitting and curse-of-dimensionality may severely damage the prediction performance of nonparametrics, as in the case of ozone data which has only 330 data points compared with 12, 34, 67, 100, 133 and 397 parameters in these six neural network models respectively. The overfitting phenomenon is shown graphically in the case of one-dimensional predictors in the lower plot. [Note: Bootstrap estimates of ASRs and MSPEs are used with 1000 resamples for each case. There is no smoothing term added for the neural networks in order to show their universal approximation capacity and overfitting phenomenon.]

kernel density estimator

$$\hat{p}(x) = \frac{1}{nh}\sum_{t=1}^{n} K\Big(\frac{x - X_t}{h}\Big) = \frac{1}{n}\sum_{t=1}^{n} W_h(x - X_t) \ , \tag{1.8}$$

with h the bandwidth. Therefore, the histogram in (1.7) is a kernel estimator with a uniform kernel function $K(u) = I(|u| \leq 1)$ and bandwidth $2h$. It still remains to choose the bandwidth h, that controls the extent to which the data are smoothed according to bias-variance trade-off (see Section 1.2.1). For example, in terms of minimizing the approximate mean integrated square error (see [11]), it has been shown that the ideal kernel function is in the form

$$K(u) = \begin{cases} \frac{3}{4\sqrt{5}}(1 - \frac{1}{5}u^2), & -\sqrt{5} \leq u \leq \sqrt{5} \\ 0, & \text{otherwise} \ . \end{cases}$$

Kernel regression method

In the context of regression, the target unknown function $f(x)$ is the conditional expectation

$$f(x) = E(Y|X = x) = \frac{\int y p(x, y) dy}{p(x)} \ ,$$

with $p(x, y)$ the joint density of (X, Y) and $p(x)$ the marginal density of X. A natural extension from the kernel density estimator in (1.8) leads to the Nadaraya-Watson kernel regression estimator [12, 13]

$$\begin{aligned} \hat{f}(x) &= \frac{\frac{1}{nh}\sum_{t=1}^{n} K\Big(\frac{x - X_t}{h}\Big) Y_t}{\frac{1}{nh}\sum_{t=1}^{n} K\Big(\frac{x - X_t}{h}\Big)} = \frac{1}{n}\sum_{t=1}^{n} \frac{\frac{1}{h}K(\frac{x - X_t}{h})}{\hat{p}(x)} Y_t \\ &= \frac{1}{n}\sum_{t=1}^{n} W_{ht}(x) Y_t = \frac{1}{n}\sum_{t=1}^{n} W_h(x; X_1, ..., X_n) Y_t \ . \end{aligned} \tag{1.9}$$

The only difference between (1.9) and (1.8) is that a weight function $W_h(\cdot)$ is defined for the response variable Y in a neighborhood of x, instead of weighting the frequencies of X itself. Certainly, more flexibility can be utilized if the amount of smoothing is adapted to the local density of data. For example, the variable bandwidth kernel estimator in the form

$$\hat{f}(x) = \frac{\frac{1}{n}\sum_{t=1}^{n} \frac{1}{hd_{t,s}} K\Big(\frac{x - X_t}{hd_{t,s}}\Big) Y_t}{\frac{1}{n}\sum_{t=1}^{n} \frac{1}{hd_{t,s}} K\Big(\frac{x - X_t}{hd_{t,s}}\Big)}$$

with $d_{t,s}$ the distance from X_t to the s-th nearest point, is closely related to the nearest neighbor protocol.

Spline regression method

Basically, there are two different scenarios for constructing a conventional nonparametric model. Firstly, they can be considered as a weighted linear summation of the response variable in a flexible neighborhood of the observed explanatory variables as in (1.9). Secondly, the flexibility of a nonparametric model is controlled by the way that the local density of data is taken into account and then fitted piecewise, as premiered in the histogram. Instead of using a kernel (variable or not), one may divide the bins according to the local density of data and then use a set of well-defined orthogonal basis functions to fit the local data in each bin while satisfying some continuity constraints at the knots (the conjunctions of bins). Splines are piecewise polynomials in the later fashion.

Example 1 (Cubic Spline) Suppose x is in some real-valued interval $[x_{min}, x_{max}]$ which is divided into $h+1$ bins by knots $x_{min} < t_1 < t_2 < ... < t_h < x_{max}$, then a cubic spline is a response function $f(x)$ on $[x_{min}, x_{max}]$ in the form

$$f(x) = d_k(x - t_k)^3 + c_k(x - t_k)^2 + b_k(x - t_k) + a_k \ , \ \forall t_k \leq x < t_{k+1} \ , \qquad (1.10)$$

with the constraints that $f(x)$ and its first and second derivatives are continuous at each knot t_k, $\forall 1 \leq k \leq h$. There are $4(h+1)$ apparent parameters in (1.10) if we define $t_0 = x_{min}$, $t_{h+1} = x_{max}$, and the second and third derivatives of $f(x)$ are zero at t_0 and t_{h+1} so that the spline is linear on $[x_{min}, t_1)$ and $[t_h, x_{max}]$ and

$d_0 = c_0 = d_{h+1} = c_{h+1} = 0$ as in the so-called natural cubic spline. However, the effective number of parameters is only $h+4$, since the maximum continuity conditions specify the following relations

$$\begin{cases} d_k(t_{k+1}-t_k)^3 + c_k(t_{k+1}-t_k)^2 + b_k(t_{k+1}-t_k) + a_k & = a_{k+1} \\ 3d_k(t_{k+1}-t_k)^2 + 2c_k(t_{k+1}-t_k) + b_k & = b_{k+1} \\ 6d_k(t_{k+1}-t_k) + 2c_k & = 2c_{k+1} \ , \end{cases} \tag{1.11}$$

so that there is basically only one effective coefficient for each bin. The cubic spline estimator $\hat{f}$ is defined as a modified least squares estimator that minimizes

$$\frac{1}{n}\sum_{t=1}^{n}(Y_t - f(X_t))^2 + \lambda \int f''(u)^2 du \ , \tag{1.12}$$

with λ the smoothing parameter that controls the trade-off between the residual error and the local variation of $\hat{f}$ (see further detailed discussion in Sections 1.2.1 and 1.2.2).
□

Moreover, it is well known (cf. [14]) that the smoothing spline $\hat{f}$ in (1.12) can be expressed as a weighted linear summation of Y_t's in the form

$$\hat{f}(x) = \frac{1}{n}\sum_{t=1}^{n} W_\lambda(x, X_t)Y_t \ , \tag{1.13}$$

with W_λ the weight function depending on λ. There are further striking relations between kernels and splines in an asymptotic sense (cf. [15]). For large n, small λ and the sample X_t not too close to the boundary, the effective weight function $W_\lambda(x,s) \sim \frac{1}{p(s)}\frac{1}{h(s)}K(\frac{s-x}{h(s)})$ with a effective local bandwidth $h(s) = \lambda^{1/4}p(s)^{-1/4}$ and $p(X)$ the marginal density of X. This fact places the smoothing spline between the fixed kernel (not depending on $p(x)$) and k-nearest-neighbor kernel (with $h(s) \propto p(s)^{-1}$), and rather close to the ideal variable kernel estimator (with $h(s) \propto p(s)^{-1/5}$). It is also shown that a cubic spline estimator can be seen as a variable kernel estimator with the kernel in the form

$$K(u) = \frac{1}{2}\exp(-|u|/\sqrt{2})\sin(|u|/\sqrt{2} + \pi/4) \ ,$$

which is symmetric with exponentially decreasing tails and negative sidelobes. After all, in the terminology of approximation theory, the kernel and spline methods belong to the class of linear integral operators (or so-called linear approximators). It is also well-known that for either case the estimator $\hat{f} \to f$ in probability if the sample size $n \to \infty$, the bandwidth (or effective bandwidth) $h \to 0$ and with other suitable regularity conditions.

So far in this section our discussion has focused on the case for a single predictor. When the predictor variable is vector-valued so that $\boldsymbol{X} \in \mathcal{R}^d$, however, there are some serious problems in choosing the appropriate shape of the kernel and defining the localness in high dimensions, if one wants to adopt the above-mentioned conventional nonparametrics directly. Though there are various generalizations devised (e.g., thin-plate spline and tensor product spline), they are usually not practical for more than two or three predictors (cf. [3], p.32). The major issue here is how to deal with a dismal phenomenon known as the *curse of dimensionality.*

1.1.4 Curse-of-dimensionality and nonlinear approximators

In any high-dimensional sample space, the data points from any practical data set of reasonable size are always not dense enough. For example, in a nine-dimensional unit cube (the same case as the ozone data), a subcube neighborhood containing 1% of the points should have a side length $(0.01)^{1/9} = 0.6$, while it is simply 0.01 for one-dimensional case. This fact has considerable impact on many aspects of regression analysis (see more discussion in Section 1.2.1). For conventional nonparametrics, the increase in dimensionality results in drastic decrease in the rate of convergence to zero in terms of approximation error and estimation error. If the target function is assumed to be in a space of functions with r degree of smoothness (e.g., $r = 2$ in Example 1) with q the number of effective parameters, then the typical rate of convergence for linear approximators is $\mathcal{O}(q^{-r/d})$. The fact that q is typically in the order of h^d does not ease the devastating rate, by two reasons: (1) q is bounded from above by the sample size n in practice; (2) an increase in r will lead to an increase in h accordingly. A similar situation occurs when the estimation error is considered.

The optimal convergence rate in estimation error for conventional nonparametrics is typically $\mathcal{O}(n^{-r/(2r+d)})$ (cf. [16, 2]).

However, the generalized additive models in (1.2) are the very few exceptional nonparametrics that are able to evade (but not break) the curse of dimensionality to a certain extent. Suppose that $C_f = \int |w||\tilde{f}(w)|dw < \infty$, i.e., the first absolute moment of the Fourier magnitude of f is bounded, where $\tilde{f}$ is the Fourier representation of f. Then the approximation error of a generalized additive model f_h in (1.2) with h additive terms, $||f - f_h||_2 \leq \mathcal{O}(C_f/\sqrt{h})$, and the estimation error, $||f_h - \hat{f}_{h,n}||_2 \leq \mathcal{O}((\frac{hd}{n}\log n)^{1/2})$ (cf. [17, 18, 4, 19]). In fact, this class of *nonlinear approximators* eases the problem by projecting the high-dimensional data into low-dimensional subspace. The price to paid is to impose increasingly strict constraint on the smoothness of the target function f as d increases by implicitly setting $r \propto d$ ($r = \lfloor d/2 \rfloor + 2$, for instance) [4, 20]. Nevertheless, the class of function represented in (1.2) is rather large in general. Although C_f is dimension-dependent and may grow exponentially fast in d, the equivalent measure for the conventional methods can be superexponentially large in term d in comparison [4]. While the conventional nonparametrics are not practical even for the cases with moderate dimensionality once $d \geq 3$, the generalized additive model as an archetype of modern regression method is the only viable multivariate tool for the data sets with mild dimensionality so far.

There is an important subclass of the generalized additive model in (1.2) that can be written in the form

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{i=1}^{d} \alpha_i g_i(x_i) \ . \qquad (1.14)$$

If the target function f is genuinely additive in terms of each predictor (and the predictors are not correlated with each other), the conventional nonparametric method can be employed to fit each predictor, and the total summation may be accurate enough for a given application. Though there is no apparent improvement in terms of approximation error, the estimation error of (1.14) is no longer exponentially dependent on d at the attractive $\mathcal{O}(n^{-r/(2r+1)})$ [3, 7, 8]. It is the model in (1.14) that is titled to the name of generalized additive model in statistics, whilst an appropriate name for a neural network model in the terminology of statistics would be generalized additive logistic model. For the sake of convenience, however, we shall use the shorter nomenclature in referring the whole class of model defined in (1.2).

1.1.5 New challenges in regression analysis

Today most of data analysis takes place in the fields outside statistics community. With the amount of data and related applications growing exponentially such as computer automated data collections in science and engineering and commercial data warehouses, there are strong and legitimate demands from all areas of information processing to develop a new generation of automated procedures aimed at discovering patterns and relationships in large complex data sets.

There are two distinctive attributes in this recent surge of activity. First, the data sets tend to be high-dimensional (in hundreds), meanwhile there are usually multiple data sets relating to the same object so that the size of a data set is easily up to mega- or giga-bytes. This fact profoundly increases the level of difficulty in performing data analysis by any data analyst even with the help of most advanced data visualization tools, and indicates the need of developing general-purpose statistical tools that possess superior accuracy yet require less human-machine interaction to the extent that it is possible. Second, due to ever-increasing computing power, many computationally intensive and sophisticated methods are now feasible. The progress made in nonparametrics incarnates this trend and provides potentially widely applicable solutions to this increasingly pressing challenge.

However, many methodologies (such as neural networks) originally proposed in data related fields other than statistics are usually only 'tried-and-true' by simulations over certain data sets or are rationalized by preliminary arguments. An analytic attention based on probabilistic inference is then in order. As we shall demonstrate in this thesis, contrary to popular belief, such an approach will not hamper the progress of the methodology in terms of its applications. In fact, it will only enhance it, and provide algorithms and tools that are computationally feasible and possess characteristics such as reliability, robustness, and improved performance over a wide variety of data sets.

1.2 Statistics of A Regression Model

1.2.1 Bias and variance

The generalized additive regression method has many desirable features regarding its approximation capacity in general. It is not all so clear, on the other hand, how good can one make this method in practice. To answer this question, we shall first take a closer look at the error measure introduced earlier in Section 1.1.1, which is the key statistic of one's regression model.

In practice, the model in (1.2) is to be estimated from a data set of size n, $D_n = \{\boldsymbol{x}_t, y_t\}_{t=1}^n = \{x_{t1}, ..., x_{td}, y_t\}_{t=1}^n$. An estimation error is introduced in this procedure in addition to the approximation error after the number of additive terms, h, is chosen. Conventionally, the model can be formulated with an associated additive residual as in

$$y_t = f(\boldsymbol{x}_t; \boldsymbol{\theta}) + \varepsilon_t = E(Y_t | \boldsymbol{X} = \boldsymbol{x}_t) + \varepsilon_t \ , \ \forall t = 1, ..., n \ . \tag{1.15}$$

The residual random variables, ε_t, are assumed to have a zero mean and a unknown covariance matrix $\boldsymbol{\Sigma}$ (with two notable special cases: the residuals are uncorrelated, $\boldsymbol{\Sigma} = \text{diag}(\sigma^2_{\varepsilon_t})$; the residuals are i.i.d. with a common variance $\sigma^2_\varepsilon = \sigma^2_{Y|\boldsymbol{X}=\boldsymbol{x}}$ (cf. eqn.(1.19)) so that $\boldsymbol{\Sigma} = \sigma^2_\varepsilon \boldsymbol{I}_n$). The residuals encompass all the remainders overlooked by the model, which include the misspecification of the class of model one chooses (that constitutes the approximation error), the unobservable and unaccounted-for predictors, and so on. An ideal estimator of $\boldsymbol{\theta}$ would be the one that minimizes the risk function $R(\hat{f}, f)$ in (1.4). Unfortunately, the risk function is also unknown because it is a function of the unknown $\boldsymbol{\theta}$. In reality, instead of minimizing the risk, the empirical risk in (1.5), an estimate of the risk from D_n, is computed and minimized with respect to $\boldsymbol{\theta}$ to obtain an estimate of the ideal estimator for $\boldsymbol{\theta}$. The usual Least Squares (LS) estimate, $\hat{\boldsymbol{\theta}}_n^{LS}$, of the unknown parameters $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta})'$ is a class of very restricted point estimation procedure in such fashion

$$\hat{\boldsymbol{\theta}}_n^{LS} = \arg\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{t=1}^{n} (y_t - f(\boldsymbol{x}_t; \boldsymbol{\theta}))^2 \ . \tag{1.16}$$

Namely it concerns with the property

$$\sum_{t=1}^{n} (y_t - f(\boldsymbol{x}_t; \boldsymbol{\theta})) \frac{\partial f(\boldsymbol{x}_t; \boldsymbol{\theta})}{\partial \theta_i} \bigg|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_n^{LS}} = 0 \ , \ (i = 1, ..., q) \ , \tag{1.17}$$

that is, the estimators are restricted by the impartiality (or unbiasedness) requirement for all value of $\boldsymbol{\theta}$.

The LS criterion is closely related to another well-known measure of (lack of) performance, the *mean-squared prediction error* (MSPE)

$$MSPE \ = \ \frac{1}{m} \sum_{t=n+1}^{n+m} (y_t - f(\boldsymbol{x}_t; \hat{\boldsymbol{\theta}}_n))^2 \ , \forall m \geq 1 \ , \tag{1.18}$$

which is the quadratic empirical risk of the estimate $\hat{\boldsymbol{\theta}}_n$ evaluated on m new observations. While the quadratic empirical risk in (1.5) is an estimate of the risk in (1.4), the MSPE as the *quadratic empirical prediction risk* is an estimate of the *prediction risk* in the following abstract form

$$\begin{aligned} P(\hat{f}, f) &= P(\hat{\boldsymbol{\theta}}(S_n), \boldsymbol{\theta}) = \int (Y - f(\boldsymbol{x}; \hat{\boldsymbol{\theta}}(S_n)))^2 d\mathcal{F} \\ &= \int (Y - f(\boldsymbol{x}; \boldsymbol{\theta}))^2 d\mathcal{F} + \int (f(\boldsymbol{x}; \boldsymbol{\theta}) - f(\boldsymbol{x}; \hat{\boldsymbol{\theta}}(S_n)))^2 d\mathcal{F} \\ &= E_{\mathcal{F}}(Y - E(Y|\boldsymbol{X} = \boldsymbol{x}))^2 + E_{\mathcal{F}}(f(\boldsymbol{x}; \boldsymbol{\theta}) - f(\boldsymbol{x}; \hat{\boldsymbol{\theta}}(S_n)))^2 \\ &= \sigma^2_{Y|\boldsymbol{X}=\boldsymbol{x}} + R(\hat{f}, f) \\ &= \sigma^2_{Y|\boldsymbol{X}=\boldsymbol{x}} + R(\hat{\boldsymbol{\theta}}(S_n), \boldsymbol{\theta}) \ , \end{aligned} \tag{1.19}$$

where $\sigma^2_{Y|\boldsymbol{X}=\boldsymbol{x}} = \sigma^2_\varepsilon$ is the variance of the residual at $\boldsymbol{x}$ in (1.15), which is the discrepancy between $R(\cdot, \cdot)$ and $P(\cdot, \cdot)$. Ideally, an estimator that minimizes the risk $R(\cdot, \cdot)$ shall also minimize $P(\cdot, \cdot)$ and vice versa, according to (1.19). It is simply not the case for the empirical versions, due to the fact that the empirical risk in its ordinary form (1.5) is not a good estimate of the true risk. When implementing a LS estimator, the resulting parameter estimate and the estimate of the response surface are based solely on the data points in the training set to reduce the model bias over these data points while everything in between is left unregulated. This can be seen clearly in the well-known bias-variance decomposition of the empirical risk.

The quadratic empirical risk R_n is also known simply as the *mean-squared-error* (MSE), that can be further decomposed into two parts as follows in the abstract form

$$\begin{aligned} R(\hat{f}, f) &= \int (f(\boldsymbol{x};\boldsymbol{\theta}) - f(\boldsymbol{x};\hat{\boldsymbol{\theta}}(S_n)))^2 d\mathcal{F} \\ &= [f(\boldsymbol{x};\boldsymbol{\theta}) - E_{\mathcal{F}} f(\boldsymbol{x};\hat{\boldsymbol{\theta}}(S_n))]^2 + E_{\mathcal{F}}[f(\boldsymbol{x};\hat{\boldsymbol{\theta}}(S_n)) - E_{\mathcal{F}} f(\boldsymbol{x};\hat{\boldsymbol{\theta}}(S_n))]^2 \\ &= \mathrm{bias}^2[f(\boldsymbol{x};\hat{\boldsymbol{\theta}}(S_n))] + \mathrm{var}[f(\boldsymbol{x};\hat{\boldsymbol{\theta}}(S_n))] , \end{aligned} \tag{1.20}$$

which is the above-mentioned decomposition of *mean-squared bias* and *model variance*. Nonparametrics are also called exact methods in applied statistics, due to the fact that their model bias can be made as low as one's wish. When h is chosen sufficiently large in (1.2), the resulting response function is able to go through every data point y_t (or nearly so) to satisfy the unbiasedness criterion (see Figure 1.2). However, what is left to be unaccounted-for is the variance component of the estimation error, and consequently causes a phenomenon called overfitting (see Figure 1.4). The imminent consequence is that the trained regression model will reveal not only true but also spurious dependence between the pair of random variables $(\boldsymbol{X}, Y)$, so that the model will perform poorly over the samples other than the training set yet from the same unknown underlying distribution. The 'curse of dimensionality' that stems from the sparsity of the data points in the high-dimensional sample space, makes the situation of overfitting even worse when the sample size is relatively small with respect to d and the effective number of parameters, which is rather common in practice. To reduce the estimation error, an alternative optimality criterion that minimize the average loss has to be defined. A straightforward heuristic is to drop the impartiality restriction for every possible parameter value, take into account of and penalize the variance part of the estimation error. In mathematical statistics literature, mainly two such error reduction methodologies have been considered: minimizing the maximum average loss and minimizing the expected loss weighted by a prior density function of $\boldsymbol{\theta}$, that led to minimax estimator and Bayes estimator respectively. The rationale here is that the over all error will be reduced by allowing a suitable amount of bias in estimation while lowering the variance part to an 'optimal' extent (see Figure 1.5).

1.2.2 Generalities on risk reduction

There is a long list of criteria which address different aspects of the risk function of an estimator or a class of estimators, and are often used to construct or select a single estimator or a substantially small set of estimators in a particular application. We shall hereby introduce the important ones which are most relevant to regression analysis, and employ them in depth in later chapters. For the sake of clarity, the linear regression model is used to showcase the conceptual matters associated with these classical risk properties.

Example 2 Consider a linear regression model in the usual normal-theory setting

$$\boldsymbol{y} = \boldsymbol{B}\boldsymbol{\theta} + \boldsymbol{\varepsilon} \ ,$$

where $\boldsymbol{y} = (y_1, ..., y_n)'$ is the $n \times 1$ vector of response variable, n is the sample size, $\boldsymbol{B} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)'$ is an $n \times q$ matrix with rank q $(n \geq q)$, $\boldsymbol{\theta}$ is the $q \times 1$ vector of unknown regression coefficients, and the residual vector $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma_\varepsilon^2 \boldsymbol{I}_n)$ with σ_ε^2 assumed to be known at this point. Under quadratic loss $L(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})'(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$, the usual LS estimator of $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\theta}}^{LS} = (\boldsymbol{B}'\boldsymbol{B})^{-1}\boldsymbol{B}'\boldsymbol{y} \ , \tag{1.21}$$

which is an unbiased estimator with covariance matrix $\sigma_\varepsilon^2(\boldsymbol{B}'\boldsymbol{B})^{-1}$ as in

$$\hat{\boldsymbol{\theta}}^{LS} \sim \mathcal{N}_q(\boldsymbol{\theta}, \sigma_\varepsilon^2(\boldsymbol{B}'\boldsymbol{B})^{-1}) \ ,$$

and has a constant prediction risk at $(n+q)\sigma_\varepsilon^2$, hence a constant risk at $q\sigma_\varepsilon^2$.

However, there are two deficiencies in $\hat{\boldsymbol{\theta}}^{LS}$:

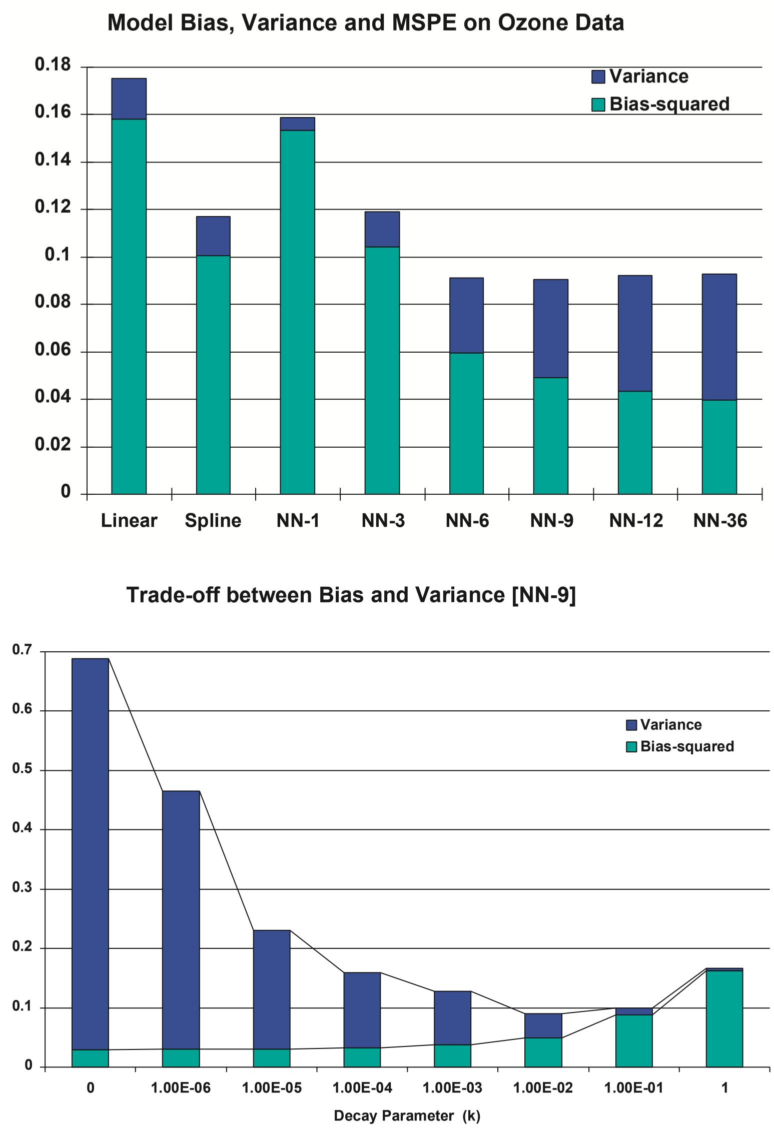


Fig. 1.5. The bias-variance decomposition of regression models on ozone data (a detailed look at the left bars in the lower plot in Figure 1.3). [Note: Bootstrap estimates of model bias and variance are used with 1000 resamples for each case. The smoothing additive cubic splines are used to represent the spline method. The neural networks (with no skip layer) are also regulated with the decay parameter set to 0.1 for $h = 1, 3$, 0.01 for $h = 6, 9, 12, 36$. Comparing with the performance by unregulated neural networks in the top plots in Figure 1.4, the regulated neural networks (using single-prior Bayes estimators discussed in Chapter 2) achieve superior overall prediction performances by allowing certain amount of model bias while drastically reducing model variability. The trade-off between model bias and variance is obtained by finely tuning decay parameter to an appropriate value (k=0.01 for the case of neural network with 9 hidden units) as shown in the lower plot.]

1. From computational perspective, $\hat{\boldsymbol{\theta}}^{LS}$ will be *unstable* in the sense that a nearly singular $(\boldsymbol{B}'\boldsymbol{B})$ will yield an inverse with inflated diagonal values so that small changes in $\boldsymbol{y}$ may produce large changes in $\hat{\boldsymbol{\theta}}^{LS}$.

2. From the perspective of global error property, $\hat{\boldsymbol{\theta}}^{LS}$ is *inadmissible* (when $q > 2$) in the sense that there are other classes of estimators whose risk are lower than or equal to that of $\hat{\boldsymbol{\theta}}^{LS}$ for all possible $\boldsymbol{\theta}$.

Motivated to correct the first deficiency in $\hat{\boldsymbol{\theta}}^{LS}$, Hoerl and Kennard [21, 22, 23] proposed a *ridge estimator* in the form

$$\hat{\boldsymbol{\theta}}(k) = (\boldsymbol{B}'\boldsymbol{B} + k\boldsymbol{I}_q)^{-1}\boldsymbol{B}'\boldsymbol{y} \ , \ \text{with } k > 0 \ . \tag{1.22}$$

It is evident that the ridge estimator is numerically stabilized by adding a positive constant to the diagonal elements of the design matrix $\boldsymbol{B}'\boldsymbol{B}$, so that the later is averted from singularity. The resulting estimator in (1.22) is a biased estimator with shrunk magnitude as shown in the form

$$\hat{\boldsymbol{\theta}}(k) \sim \mathcal{N}_q((\boldsymbol{B}'\boldsymbol{B} + k\boldsymbol{I}_q)^{-1}\boldsymbol{B}'\boldsymbol{B}\boldsymbol{\theta}, \sigma_\varepsilon^2(\boldsymbol{B}'\boldsymbol{B} + k\boldsymbol{I}_q)^{-1}\boldsymbol{B}'\boldsymbol{B}(\boldsymbol{B}'\boldsymbol{B} + k\boldsymbol{I}_q)^{-1}) \ . \tag{1.23}$$

This is also reasonable, since the expected magnitude of $\hat{\boldsymbol{\theta}}^{LS}$ is always higher than the true length as shown in

$$\boldsymbol{E}[(\hat{\boldsymbol{\theta}}^{LS} - \boldsymbol{\theta})'(\hat{\boldsymbol{\theta}}^{LS} - \boldsymbol{\theta})] = \boldsymbol{E}((\hat{\boldsymbol{\theta}}^{LS})'\hat{\boldsymbol{\theta}}^{LS}) - \boldsymbol{\theta}'\boldsymbol{\theta} = \sigma_\varepsilon^2(\boldsymbol{B}'\boldsymbol{B})^{-1} > 0 \ .$$

Moreover, it can be shown that for a fixed parameter point $\boldsymbol{\theta}_0$ (and fixed $(\boldsymbol{B}'\boldsymbol{B})$), there exists a $k > 0$ depending on $\boldsymbol{\theta}_0$, for which the risk of $\hat{\boldsymbol{\theta}}(k)$ is smaller than the risk of $\hat{\boldsymbol{\theta}}^{LS}$ (see Figure 1.6).

The ridge estimator $\hat{\boldsymbol{\theta}}(k)$ can be seen as a *single-prior Bayes* estimator with respect to a prior density of the parameter vector $\boldsymbol{\theta}$, $\pi(\boldsymbol{\theta}) \sim \mathcal{N}_q(\boldsymbol{0}, \frac{\sigma_\varepsilon^2}{k}\boldsymbol{I}_q)$, over the parameter space $\boldsymbol{\Theta}$. For a Bayesian point of view, the ridge estimator is the posterior mean of the distribution of the parameter given data, $p(\boldsymbol{\theta}|D_n)$, as presented in (1.23). In fact, any estimator of parameters in a regression model is a random variable itself as

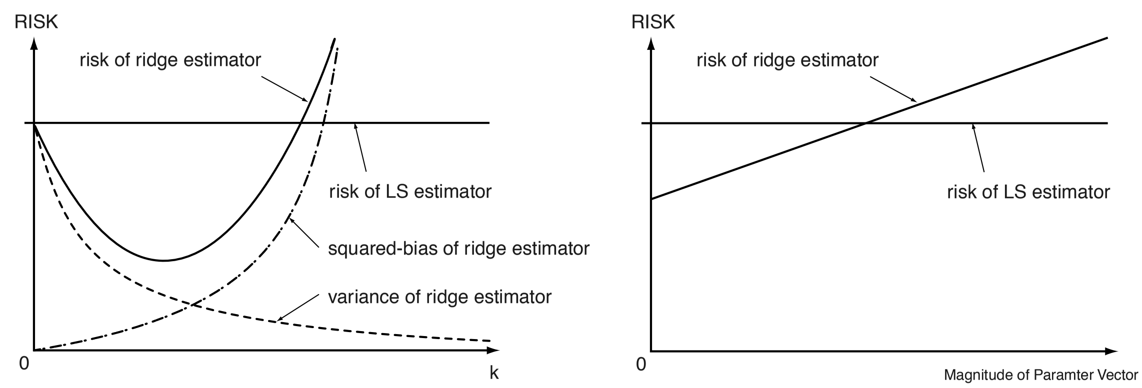


Fig. 1.6. The Diagram of risk behavior of an ordinary ridge regression model.

mentioned in Section 1.1.1. According to Bayesian philosophy, it is natural to treat any unknown parameter in one's model as a random variable and to assign a prior density function $\pi(\boldsymbol{\theta})$ over the parameter space. The Bayes expected loss (*Bayes risk*)

$$r(\pi, \hat{\boldsymbol{\theta}}) = \int R(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$$

is then sought to remove the randomness in the risk function.

Definition 1 (Bayes estimator) An estimator, $\hat{\boldsymbol{\theta}}^{\pi}$, which minimizes $r(\pi, \hat{\boldsymbol{\theta}})$ is called a *Bayes estimator* with respect to $\pi(\boldsymbol{\theta})$. □

However, a further examination by global property analysis will show that no fixed (chosen) $k > 0$ can dominate the LS estimator for all possible $\boldsymbol{\theta}$ and $(\boldsymbol{B}'\boldsymbol{B})$. The risk function of the ridge estimator has the following bias-variance decomposition

$$\begin{aligned} R(\hat{\boldsymbol{\theta}}(k)) &= \boldsymbol{E}||\boldsymbol{f}(\hat{\boldsymbol{\theta}}(k)) - \boldsymbol{f}(\boldsymbol{\theta})||^2 = (\boldsymbol{B}\hat{\boldsymbol{\theta}}(k) - \boldsymbol{B}\boldsymbol{\theta})'(\boldsymbol{B}\hat{\boldsymbol{\theta}}(k) - \boldsymbol{B}\boldsymbol{\theta}) \\ &= k^2\boldsymbol{\theta}'(\boldsymbol{B}'\boldsymbol{B} + k\boldsymbol{I}_q)^{-1}\boldsymbol{B}'\boldsymbol{B}(\boldsymbol{B}'\boldsymbol{B} + k\boldsymbol{I}_q)^{-1}\boldsymbol{\theta} \\ &\quad + \sigma_\varepsilon^2 \mathrm{tr}[\boldsymbol{B}(\boldsymbol{B}'\boldsymbol{B} + k\boldsymbol{I}_q)^{-1}\boldsymbol{B}'\boldsymbol{B}(\boldsymbol{B}'\boldsymbol{B} + k\boldsymbol{I}_q)^{-1}\boldsymbol{B}'] \\ &= k^2 \sum_{i=1}^{q} \frac{\gamma_i^2 \lambda_i}{(\lambda_i + k)^2} + \sigma_\varepsilon^2 \sum_{i=1}^{q} \frac{\lambda_i^2}{(\lambda_i + k)^2} \\ &= \mathrm{bias}^2[\boldsymbol{f}(\hat{\boldsymbol{\theta}}(k))] + \mathrm{variance}[\boldsymbol{f}(\hat{\boldsymbol{\theta}}(k))] \ , \end{aligned}$$

where a canonical reduction is performed by letting $\boldsymbol{B}'\boldsymbol{B} = (\boldsymbol{G}^{-1})'\boldsymbol{\Lambda}\boldsymbol{G}^{-1}$, $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_i)$, and $\boldsymbol{\gamma} = \boldsymbol{G}^{-1}\boldsymbol{\theta}$. It is then obvious that the risk of $\hat{\boldsymbol{\theta}}(k)$ becomes unbounded when either the magnitude of the true $\boldsymbol{\theta}$ or k increases (see Figure 1.6). An ordinary ridge estimator has bad tail risk behavior, though it is admissible with respect to the LS estimator (i.e., its risk can be lower than that of the LS estimator somewhere over the parameter space but not everywhere). It can easily shown that a ridge estimator (i.e., with a normal prior assumed) would have infinite Bayes risk if the true prior were Cauchy [24].

A conservative approach to risk reduction is then devoted to minimize the maximum risk of a possible estimator.

Definition 2 (Minimax estimator) An estimator, $\hat{\boldsymbol{\theta}}^M$, which satisfies

$$\sup_{\boldsymbol{\theta}} R(\hat{\boldsymbol{\theta}}^M, \boldsymbol{\theta}) = \inf_{\hat{\boldsymbol{\theta}}} \sup_{\boldsymbol{\theta}} R(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) ,$$

is called a *minimax estimator.* □

It is clear that the LS estimator (1.21) is a minimax estimator itself. Though there is no simple direct method to construct a minimax estimator, the class of shrinkage estimators originated from James-Stein estimator [25, 26, 27] possesses very attractive classical risk properties. For instance, the Berger-Hudson estimator [28, 29] in the canonical form

$$\hat{\gamma}_i^{BH} = \left(1 - \frac{(q-2)\sigma_\varepsilon^2 \lambda_i}{\sum_{i=1}^q \lambda_i^2 \hat{\gamma}_i^2}\right) \hat{\gamma}_i \ , \forall i = 1, ..., q \ , \qquad (1.24)$$

with $\hat{\boldsymbol{\gamma}} = \boldsymbol{G}^{-1}\hat{\boldsymbol{\theta}}^{LS}$ and $\hat{\boldsymbol{\theta}}^{BH} = \boldsymbol{G}\hat{\boldsymbol{\gamma}}^{BH}$, is a minimax estimator (see Figure 1.7). Compared with the left diagram in Figure 1.6, the risk function of $\hat{\boldsymbol{\theta}}^{BH}$ has a nice tail behavior (i.e., its risk never exceeds the risk of LS estimator). It can be shown that a Berger-Hudson estimator can be seen as a generalized ridge regression model in the form

$$\hat{\boldsymbol{\theta}}(\boldsymbol{C}) = (\boldsymbol{B}'\boldsymbol{B} + \boldsymbol{C})^{-1}\boldsymbol{B}'\boldsymbol{y} ,$$

with the matrix $\boldsymbol{C}$ having the elements

$$c_{ij} = -b_{ij} + \delta_{ij} / \left(1 - \frac{(q-2)\sigma_\varepsilon^2}{(\boldsymbol{B}\hat{\boldsymbol{\theta}}^{LS})'\boldsymbol{B}\hat{\boldsymbol{\theta}}^{LS}}\right) ,$$

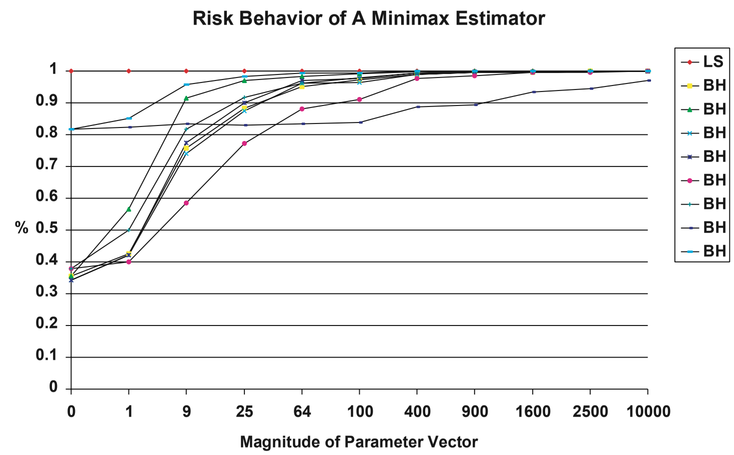


Fig. 1.7. The risk behavior of a Berger-Hudson minimax estimator in (1.24). [Note: An example from [30] is plotted here to show the minimaxity of Berger-Hudson estimator. For various types of the eigenvalue spectrum of $\boldsymbol{B}'\boldsymbol{B}$ and true values of the regression coefficients in a linear regression model with $q = 6$ and $\sigma_\varepsilon^2 = 1$, the risks of $\hat{\boldsymbol{\theta}}^{BH}$ is illustrated as percentages of the constant risk of $\hat{\boldsymbol{\theta}}^{LS}$.]

where $b_{ij} = \{\boldsymbol{B}'\boldsymbol{B}\}_{ij}$ and δ_{ij} is 0 if $i \neq i$ and 1 if $i = j$. Hence, the minimax estimator in (1.24) shares the numerical stability possessed by the ridge procedures, and is admissible at the same time. However, the build-in nature of the shrinkage factor in a James-Stein estimator limits its usage, because it makes difficult (if not entirely impossible) to directly incorporate any prior knowledge. When certain prior knowledge (even if it is rather vague) is variable, a Bayesian treatment is definitely in order to enable one to finely tune the significantly improved regions on parameter space for a particular application. In fact, Bayesian procedure is one of the methods for constructively generating estimators with optimal frequentist properties such as minimaxity.

In general, all alternatives of the ordinary LS estimator improve their performance only over certain regions of the parameter space. Outside these regions, their risks are either essentially equal to or worse than that of the LS estimator. An reasonable thought is that such region should be somehow determined by the data in hand when the estimate is sought. Hence it is natural to adopt the Bayesian approach in such

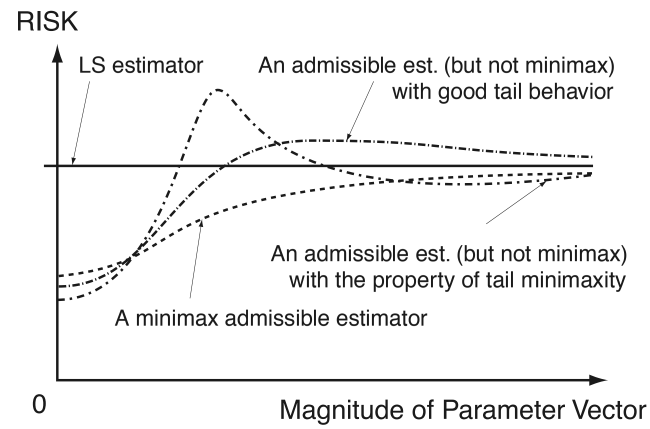


Fig. 1.8. Possible risk behaviors of various improved alternative estimators.

direction so that one can incorporate any prior knowledge to take advantage of the parameter regions of significant improvement in risk. For example, for the linear regression model, a vague prior believe of $\boldsymbol{\theta}$ can be phrased as follows: the regression coefficients should reflect the order of magnitude of the response variable. According to this prior believe, [31, 32] suggested to use $\hat{k} = \hat{\sigma}_\varepsilon^2/[1/q(\hat{\boldsymbol{\theta}}^{LS})'\hat{\boldsymbol{\theta}}^{LS}]$ in the place of k in (1.22). The rationale here is to consider a class of prior instead of fixed one, and then use the data set in hand to select the most probable one which reflects the preferred regions on parameter space. Empirical Bayes, hierarchical Bayes and robust Bayes estimators can all be seen as variants in this direction, and the resulting estimators can be viewed as certain trade-offs among single-prior Bayes, minimax and LS estimators (see a conceptual diagram of possible risk properties of various estimators in Figure 1.8 and further details in Chapter 2). □

Finally, we summarize a set of desirable properties of an alternative improved estimator from the perspective of its global error measure in 1-4 and others from other practical considerations in 5-6 as follows:

1. $\hat{\boldsymbol{\theta}}$ and $f(\boldsymbol{x}, \hat{\boldsymbol{\theta}})$ should be ready for incorporating prior information on the residuals and the parameters, and should possess Bayesian robustness, i.e., be robust

with respect to misspecification of prior information.

2. $\hat{\boldsymbol{\theta}}$ and $f(\boldsymbol{x}, \hat{\boldsymbol{\theta}})$ should keep classical risk properties of Stein-like shrinkage estimators, i.e., should be asymptotically minimax and admissible with respect to the Least Squares estimators (or nearly so).

3. $\hat{\boldsymbol{\theta}}$ should have improved good confidence regions for $f(\boldsymbol{x}, \hat{\boldsymbol{\theta}})$.

4. $\hat{\boldsymbol{\theta}}$ and $f(\boldsymbol{x}, \hat{\boldsymbol{\theta}})$ should be $\sqrt{n}$-consistent and asymptotically efficient.

5. $\hat{\boldsymbol{\theta}}$ and $f(\boldsymbol{x}, \hat{\boldsymbol{\theta}})$ should preserve the numerical stability of the ridge procedures.

6. $\hat{\boldsymbol{\theta}}$ and $f(\boldsymbol{x}, \hat{\boldsymbol{\theta}})$ should be in an explicit closed formulation that is easy for both numerical implementation and theoretical analysis of unintuitive risk properties.

The goal of this thesis is to utilize the above list in constructing improved algorithm for the generalized additive model in (1.2), which has been widely used in application fields such as machine learning, pattern recognition, neural computation, signal and image processing, data and knowledge engineering, econometrics, applied statistics, and other areas of information processing.

1.3 Contributions

State-of-the-art

The feedforward neural network model was related to nonparametric statistical inference in a tutorial by Geman, Bienenstock and Doursat [33], and was catalogued as a state-of-the-art statistical method for high-dimensional data analysis. The bias-variance dilemma was also explained for this model and other commonly used nonparametric models such as kernels and nearest neighbor methods. It was also rightfully pointed out that the asymptotic consistency of ML (LS) estimation shared by all nonparametric methods does not provide clue on how to balance bias and variance for training samples of finite size. A justified conclusion drawn from experiments with complex data is that the identification of carefully designed biases are the more fundamental and difficult research tasks. In most applications, the bias is designed by hand for each particular problem while giving up generality, therefore there is no guarantee that the improved prediction performance will sustain when data situation changes.

Nonparametric statistics in general has matured in the last two decades. It is not surprising that many 'generalist' techniques and tools developed for kernels and splines methods have been increasingly used to address the issue of risk reduction in neural network models. A *'generalist'* approach is to index the model with hyperparameters such as the number of hidden units h or the ridge coefficient k in (1.22), then adjust it to a proper value according to the sample to deliver a good bias-variance trade-off, without other ad hoc assumptions on the data or the model. The most commonly used method in this venue is *cross-validation* [34], which can be seen as a simple version of Monte Carlo assessment of estimation performance. Let $D_n^{(t)} = \{(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_{t-1}, y_{t-1}), (\boldsymbol{x}_{t+1}, y_{t+1}), ..., (\boldsymbol{x}_n, y_n)\}$ be the 'leave-one-out' data set excluding the tth data point $(\boldsymbol{x}_t, y_t)$. Then for each fixed value of k, n neural network models are trained based on n 'leave-one-out' data sets $D_n^{(t)}, \forall t = 1, ..., n$, and a total prediction error measured on the left-out data points under this particular choice of hyperparameter k

$$CV_n(k) = \frac{1}{n}\sum_{t=1}^{n}[y_t - f(\boldsymbol{x}_t; \hat{\boldsymbol{\theta}}(k, D_n^{(t)})]^2$$

is calculated. The cross-validated hyperparameter k^* minimizes $CV_n(k)$ so that the resulting estimator of the parameters is $\hat{\boldsymbol{\theta}}(k^*, D_n)$. The main problem with this approach is its high computational cost and relatively weak analytic and numerical vindication that it indeed reaches an optimal bias-variance trade-off.

In searching for a 'standard' single-run algorithm of training a neural network regression model, the Bayesian paradigm (or regularization method in the terminology of approximation theory) is naturally sought, due to the possibility of closed formulation. The following Bayesian approaches have been exploited for training 'generalist' neural network models.

1. The single-prior Bayes method assumes a prior density for the parameter vector

$\boldsymbol{\theta}$ as in

$$\pi(\boldsymbol{\theta}) \sim \mathcal{N}_q(\mathbf{0}, \frac{\sigma_\varepsilon^2}{k}\boldsymbol{I}_q) \ , \tag{1.25}$$

with $k > 0$ fixed. Quasi-Newton or conjugate gradient algorithm of neural network with fixed k as an option has been coded and included in well-established statistics packages such as SAS and S-PLUS [35, 36, 37]. However, there is no additional mechanism provided in the softwares to balance the bias and variance, so that the user must resort to cross-validation or a single guess of a suitable k.

2. The empirical Bayes method was investigated by MacKay and Neal [38, 39] among several others, though it was not stated explicitly to be empirical Bayesian. MacKay's work is basically to devise the type-II ML choice of k in the adaptive ridge fashion for neural network model, which is also pointed out by Ripley [36]. Neal (and MacKay later as well) seeks an exact calculation under the same approach by using Markov Chain Monte Carlo (MCMC) techniques, with a much higher computational cost than the cross-validation for even very small examples. There was no risk analysis performed on this approach to see to what extent the adaptive k helps or hurts the bias-variance trade-off and how much performance gain is indeed utilized by iterating k.

3. The hierarchical Bayes method was introduced to train a neural network regression model by Müller and Neal [40, 39]. By using the standard conjugate prior hierarchy in (2.39) and completing numerical integration through MCMC, the results on small examples are mainly used for the purpose of showcasing the potential multimodality of posteriors of the parameters, rather than a feasible standard training algorithm. Again there is no risk analysis available to show the possible benefit obtained from this approach for the bias-variance trade-off.

Overall, although the several Bayesian approaches were introduced for the purpose of balancing the model bias and variance, theoretical and numerical verifications have yet to be carried out with the following open questions in mind.

1. Is there a rigorous framework for risk reduction for neural network regression model as for the canonical case of multivariate normal mean vector and linear regression model?

2. How can one derive the existing estimators from various approaches such as Bayesian and regularization methods by this framework, instead of in an ad hoc fashion?

3. How can one evaluate various alternative estimators in the light of their 'true' prediction performance, instead of only running a few small example simulations?

4. Is it possible to go beyond the existing methods and devise a standard single-run algorithm for the 'generalist' neural network regression model?

5. Will the neural network regression model with carefully designed model bias indeed show better prediction performance than the best conventional nonparametric method? [The answer is yes at least for the ozone data (see Figure 1.5)]

Tools and techniques

As a necessity for all nonlinear regression models in general [41, 42], large-sample asymptotics based on linear approximation of the response surface are used in our risk analysis. An asymptotic squared-bias and variance decomposition can be then derived to motivate and evaluate new general-purpose algorithms that possess good risk behaviors. An estimator is likely to be reasonably good if it can be shown that its asymptotic bias and variance are under control. At the same time, further verification is needed for any claim in the large-sample sense. This is usually done by evaluating the measure of curvature at the estimated parameter point and taking into account of other small-sample effects, which is one of the open topics in neural network regression. The ultimate goal is to develop a predictable and verifiable risk analysis and evaluation procedure for nonlinear regression at large. Our results will show that this approach to risk reduction in neural network regression model is rather effective. The conditions for the validity of the approximations used here appear to be similar to the conditions under which the nonlinear regression model was handled in general [41] and the methodologies was originally developed. Our presentation shall focus more on the practical side of various approaches, and leave more precise mathematical treatments for future work.

In the past, for the numerical evaluation of prediction performance of a neural network training algorithm, a test data set is usually used to calculate the total prediction risk, and is usually compared with the LS estimator to show that an improvement is achieved. This can be rather misleading for two reasons: the test set like the training set is usually too small for the purpose at hand; a lower risk than the LS estimator does not mean a rightful bias-variance trade-off is established. Only a rather accurate estimation of both bias and variance of the trained models on the data set can properly justify if a better bias-variance balance is indeed obtained and no more risk reduction can be utilized under the current scenario. We resort to the well-developed bootstrap method [43] from statistics to evaluate the model bias and variance of various training algorithms. Since the number of runs in bootstrap method is not upper bounded by the sample size as in the case of cross-validation and can be set arbitrarily large, we can ensure the accuracy of the evaluation by making the standard error of an estimation of bias, variance or the total prediction error 10 factors lower than the estimation itself. The number of bootstrap runs is set to 1000 in our experiments with satisfactory results.

Contributions

There is two major aspects in our work on developing new estimation procedures that possess the set of desirable risk properties that we have emphasized.

First, for the application fields, the statistical analysis based on risk properties over parameter space is introduced for the first time to the neural network regression model. Although several Bayes methods have been adopted to address the issues latently related to risk behavior of neural network training algorithms, it is generally unaware of the potential drawbacks that come with these methods and the possible solutions to overcome their weakness while keeping their strength. We developed a framework for asymptotic risk analysis based on linear approximation formulation of the response surface. We first clarify how existing Bayesian methods for the neural network training are derived. Then we employ the framework to evaluate them analytically and numerically with highlights on the following major drawbacks:

1. For the single-prior Bayes method, any predetermined value of the hyperparameter in prior can lead to an unbounded resulting risk, that is due to the potentially unbounded model bias.

2. For the empirical Bayes method based on the type-II ML method, the adaptive hyperparameter tends to converge to a value which is too high so that with high probability the parameters are shrunk too much. This contributes to a undesirable high model bias, and also lowers the coverage probability of its corresponding confidence intervals.

Overall, the deficiencies in the above two approaches come with the potential misspecification of the prior density (be it adaptive or not) and lack of Bayesian robustness when the light-tailed prior is used. To overcome the weakness of previously employed Bayesian methods in neural network training, we use hierarchical Bayesian methodology to develop a new robust Bayesian estimator for neural network regression model. The concept of Bayesian robustness is introduced to these application fields for the first time through our work. We show that this concept and associated methodology can be extended to nonlinear regression models such as neural network rather well. And we suggest a scenario under which a Newton-Raphson iterative optimization procedure is derived and coded to show how this new estimator improves on the capacities of existing algorithms. The resulting estimator is in the form

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{\tau+1} &= \hat{\boldsymbol{\theta}}_{\tau} + (\hat{\boldsymbol{F}}'_{\tau}\hat{\boldsymbol{F}}_{\tau} + \hat{k}_{\tau}\boldsymbol{I}_q)^{-1} \\ &\quad \{[\boldsymbol{I}_q + (1-\hat{r}_{q\tau})\hat{k}_{\tau}(\hat{\boldsymbol{F}}'_{\tau}\hat{\boldsymbol{F}}_{\tau})^{-1}]\hat{\boldsymbol{F}}_{\tau}\hat{\boldsymbol{\varepsilon}}_{\tau} - \hat{r}_{q\tau}\hat{k}_{\tau}\hat{\boldsymbol{\theta}}_{\tau}\} \; , \qquad (1.26)\end{aligned}$$

where $\hat{k}_\tau$ plays the same role of residual variance as in the single-prior Bayes and empirical Bayes methods (cf. eqn. (1.25)), $\hat{r}_{q\tau}$ is a new function emerged from a more complicated prior hierarchy and will play a crucial role in robustifying the Bayesian procedure and balancing bias and variance (cf. eqns. (3.3) and (3.16)). It demonstrates an improved overall prediction performance in the sense that (cf. Figures 3.1, 3.2 and 3.3):

1. When the guessed hyperparameter is wrong (too high or too low), the new estimator delivers a lower prediction error than that of the single-prior Bayes method, showing the effect of the Bayesian robustness it possesses.

2. If the guessed hyperparameter is too low, the new estimator bears more character of the empirical Bayes method with a higher model bias but lower model variance and a lower prediction risk.

3. If the guessed hyperparameter is too high, the new estimator shows more character of a Least Squares estimator with a lower model bias but higher model variance and a lower prediction risk that levels off at the level of an empirical Bayes estimator and never goes unbounded like in the case of the single-prior Bayes method.

4. When the guessed hyperparameter is about right, the new estimator delivers virtually the same good performance as the single-prior Bayes method, and avoids a higher prediction risk as expected for the empirical Bayes method by not overshrinking the parameters.

Furthermore, as a default, 'standard' and single-run algorithm, the new estimator shows consistent performance gain over a wide variety of data settings illustrated by synthetic data sets.

Second, for the fields of analytic and applied statistics, our work ratchets up the performance gain delivered by the neural network nonparametric regression model and the theoretical understanding of the underlying mechanism which makes this model more appealing and acceptable as a new kit in statistical toolbox. The applicability of many well-developed methodologies from statistics has yet been verified in the application fields such as neural computation. Often the statistical methodologies are adopted into applications without care of the conditions under which the method can be used, and without theoretically sound justification that the weakness of a specific method would not become a major shortcoming in model performance. By adopting various techniques from statistics and related fields, we first clarify some major flaws of existing methods used in neural network training, then propose a much refined approach inspired by mathematical statistical study of the rather basic case in canonical form. Our approach finally results in an algorithm for neural network training that has various desirable statistical characteristics such as Bayesian robustness, numerical stability from ridge procedure, asymptotic minimaxity, improved prediction confidence intervals. Moreover, the algorithm is in an explicit closed formulation that is easy to program and does not require significant extra computational cost than the ordinary Least Squares estimation.

Overview of thesis

The thesis is organized as follows. After an extensive survey over major statistical regression models and methodologies from the global error measure perspective in Chapter 1, we prepare theoretical setups for general nonlinear regression analysis necessary to our presentation in Section 2.1. The single-prior Bayes method is shown as a Bayesian implementation of the ridge procedure with a potentially unbounded risk in Section 2.2.1. The empirical Bayes method is shown to be an adaptive version of the ridge procedure with a high probability of overshrinking the parameters in Section 2.2.2. All the above are corroborated with simulation results on real and synthetic data sets. The hierarchical Bayesian methodology is introduced in Section 2.2.3. Using this methodology, we develop the new robust Bayes estimator in Chapter 3, and propose a plausible default version and implement it numerically to show its overall improved prediction performance. In the final chapter, we reflect on various aspects of the approach characterized by balanced consideration in risk behavior of a potential estimator, and the open questions remained in this field.

2. Approaches Based On Global Error Properties

2.1 Preparations

2.1.1 Likelihood function, Newton-Raphson method and one-step approximation

Typically, for the generalized additive model in (1.2), a usual LS estimate of the parameter vector represented in (1.16) and (1.17) has multiple roots. The most commonly used approach is to use iterative optimization methods such as Newton-Raphson to obtain an approximate solution of (1.17).

Under the assumption of i.i.d. normal residuals, the ordinary LS estimator is the same as the maximum likelihood (ML) estimator. The normal-theory regression model in (1.15) implies that

$$\boldsymbol{y} \sim \mathcal{N}_n(\boldsymbol{f}(\boldsymbol{\theta}^*), \sigma_\varepsilon^2 \boldsymbol{I}_n) \ ,$$

where $\boldsymbol{\theta}^*$, the true value of $\boldsymbol{\theta}$, belongs to $\boldsymbol{\Theta} \subset \mathcal{R}^q$. We shall use the notation $f_t(\boldsymbol{\theta}) = f(\boldsymbol{x}_t; \boldsymbol{\theta})$, $n \times 1$ vectors $\boldsymbol{y} = (y_1, y_2, ..., y_n)'$ and $\boldsymbol{f}(\boldsymbol{\theta}) = (f_1(\boldsymbol{\theta}), f_2(\boldsymbol{\theta}), ..., f_n(\boldsymbol{\theta}))'$, and an $n \times q$ matrix $\boldsymbol{F}(\boldsymbol{\theta}) = \nabla \boldsymbol{f}(\boldsymbol{\theta}) = \frac{\partial \boldsymbol{f}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \left[\left(\frac{\partial f_t(\boldsymbol{\theta})}{\partial \theta_i}\right)\right]$ with $n > q$, $\text{rank}(\boldsymbol{F}(\boldsymbol{\theta})) = q$, and $\text{rank}(\boldsymbol{F}(\boldsymbol{\theta})'\boldsymbol{F}(\boldsymbol{\theta})) = q$. The *likelihood function* of a regression model can be written in a conditional density form

$$p(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\theta}, \sigma_\varepsilon^2) = (2\pi\sigma_\varepsilon^2)^{-n/2} \exp\left[-\frac{1}{2\sigma_\varepsilon^2} \parallel \boldsymbol{y} - \boldsymbol{f}(\boldsymbol{\theta}) \parallel^2\right] \ , \tag{2.1}$$

with $\parallel \boldsymbol{y} - \boldsymbol{f}(\boldsymbol{\theta}) \parallel^2 = \sum_{t=1}^n (y_t - f(\boldsymbol{x}_t; \boldsymbol{\theta}))^2$, and the log-likelihood function as

$$l(\boldsymbol{\theta}, \sigma_\varepsilon^2) = \log p(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\theta}, \sigma_\varepsilon^2) = -\frac{n}{2}\log(2\pi\sigma_\varepsilon^2) - \frac{1}{2\sigma_\varepsilon^2} \parallel \boldsymbol{y} - \boldsymbol{f}(\boldsymbol{\theta}) \parallel^2 \ . \tag{2.2}$$

An ML estimate of the parameters is the one that maximizes the (log-)likelihood function by solving the *likelihood equation*

$$l'(\boldsymbol{\theta}, \sigma_\varepsilon^2) = \frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta}, \sigma_\varepsilon^2) = 0 \ , \tag{2.3}$$

and $\frac{\partial}{\partial \sigma_\varepsilon^2} l(\boldsymbol{\theta}, \sigma_\varepsilon^2) = 0$. Note that in a small neighborhood of a root $\boldsymbol{\theta}^*$ of (2.3), the linear Taylor expansion of $\boldsymbol{f}(\boldsymbol{\theta})$ is

$$f_t(\boldsymbol{\theta}) \approx f_t(\boldsymbol{\theta}^*) + \sum_{i=1}^{q} \left. \frac{\partial f_t(\boldsymbol{\theta})}{\partial \theta_i} \right|_{\boldsymbol{\theta}^*} (\theta_i - \theta_i^*) \ ,$$

or

$$\boldsymbol{f}(\boldsymbol{\theta}) \approx \boldsymbol{f}(\boldsymbol{\theta}^*) + \boldsymbol{F}(\boldsymbol{\theta}^*)(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \ . \tag{2.4}$$

Hence, the log-likelihood function becomes

$$\begin{aligned} l(\boldsymbol{\theta}, \sigma_\varepsilon^2) &\approx -\frac{n}{2} \log(2\pi\sigma_\varepsilon^2) - \frac{1}{2\sigma_\varepsilon^2} \parallel \boldsymbol{y} - \boldsymbol{f}(\boldsymbol{\theta}^*) - \boldsymbol{F}(\boldsymbol{\theta}^*)(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \parallel^2 \\ &= -\frac{n}{2} \log(2\pi\sigma_\varepsilon^2) - \frac{1}{2\sigma_\varepsilon^2} \parallel \boldsymbol{\varepsilon} - \boldsymbol{F}(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \parallel^2 \ , \end{aligned} \tag{2.5}$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)' = \boldsymbol{y} - \boldsymbol{f}(\boldsymbol{\theta}^*)$ and $\boldsymbol{F} = \boldsymbol{F}(\boldsymbol{\theta}^*)$. The likelihood equation (2.3) is solved approximately when $\boldsymbol{\theta}$ and σ_ε^2 are substituted by

$$\hat{\boldsymbol{\theta}} \approx \boldsymbol{\theta}^* + (\boldsymbol{F}'\boldsymbol{F})^{-1} \boldsymbol{F}' \boldsymbol{\varepsilon} \ , \tag{2.6}$$

and

$$\hat{\sigma_\varepsilon^2} = \frac{1}{n} \parallel \boldsymbol{y} - \boldsymbol{f}(\hat{\boldsymbol{\theta}}) \parallel^2 \ , \tag{2.7}$$

where $\hat{\sigma_\varepsilon^2}$ is usually replaced by its unbiased version, $\hat{\sigma_\varepsilon^2} = \dfrac{1}{n-q} \parallel \boldsymbol{y} - \boldsymbol{f}(\hat{\boldsymbol{\theta}}) \parallel^2$.

In practice, the unknown root $\boldsymbol{\theta}^*$ must be replaced by an approximate one itself, that leads to the *Newton-Raphson* iterative procedure. If $\tilde{\boldsymbol{\theta}}$ is the approximate solution, then a linear Taylor expansion of the left side of (2.3) about $\tilde{\boldsymbol{\theta}}$ leads to the approximation

$$0 = l'(\hat{\boldsymbol{\theta}}) = l'(\tilde{\boldsymbol{\theta}}) + l''(\tilde{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}) \ ,$$

and this gives us

$$\hat{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}} - \frac{l'(\tilde{\boldsymbol{\theta}})}{l''(\tilde{\boldsymbol{\theta}})} \ , \tag{2.8}$$

which is equivalent to (2.6) with $\boldsymbol{\theta}^*$ replaced by $\tilde{\boldsymbol{\theta}}$ and $l'(\tilde{\boldsymbol{\theta}}) = -2\tilde{\boldsymbol{F}}(\boldsymbol{y} - \boldsymbol{f}(\tilde{\boldsymbol{\theta}}))$, $l''(\tilde{\boldsymbol{\theta}}) = 2\tilde{\boldsymbol{F}}'\tilde{\boldsymbol{F}}$. The resulting sequence of estimates $\{\hat{\boldsymbol{\theta}}_n\}$ and its substitution estimates

$\{f(\hat{\boldsymbol{\theta}}_n)\}$ (as n increases) can be made consistent, asymptotically normal and asymptotically efficient under certain regularity conditions as presented in [44, 1, 41, 45]. Throughout the rest of the thesis, we are mainly concerned with the risk behaviors of the *one-step* approximation of various estimators of $\boldsymbol{\theta}$ in (1.2) and using them as indications for developing new estimators in the Newton-Raphson fashion with desirable global error properties.

2.1.2 The usual Least Squares estimation and statistical inference

When n is large, the usual LS estimate $\hat{\boldsymbol{\theta}}$ is in a small neighborhood of $\boldsymbol{\theta}^*$ as in (2.6), and

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \sim \mathcal{N}_q(\mathbf{0}, \frac{1}{n}\sigma_\varepsilon^2(\boldsymbol{F}'\boldsymbol{F})^{-1}) \ , \tag{2.9}$$

where the approximation holds to $\boldsymbol{o}_p(n^{-1/2})$ (i.e., $\forall\delta > 0, \lim_{n\to\infty} \Pr(\sqrt{n}||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*|| \leq \delta) = 1$). Consequently, the substitution estimate

$$\boldsymbol{f}(\hat{\boldsymbol{\theta}}) \approx \boldsymbol{f}(\boldsymbol{\theta}^*) + \boldsymbol{F}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \approx \boldsymbol{f}(\boldsymbol{\theta}^*) + \boldsymbol{F}(\boldsymbol{F}'\boldsymbol{F})^{-1}\boldsymbol{F}'\boldsymbol{\varepsilon} = \boldsymbol{f}(\boldsymbol{\theta}^*) + \boldsymbol{P_F}\boldsymbol{\varepsilon} \ , \tag{2.10}$$

or written as $\boldsymbol{f}(\hat{\boldsymbol{\theta}}) \sim \mathcal{N}_n(\boldsymbol{f}(\boldsymbol{\theta}^*), \sigma_\varepsilon^2\boldsymbol{P_F})$, that leads to

$$\boldsymbol{y} - \boldsymbol{f}(\hat{\boldsymbol{\theta}}) \approx \boldsymbol{y} - \boldsymbol{f}(\boldsymbol{\theta}^*) - \boldsymbol{F}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \approx \boldsymbol{\varepsilon} - \boldsymbol{P_F}\boldsymbol{\varepsilon} = (\boldsymbol{I}_n - \boldsymbol{P_F})\boldsymbol{\varepsilon} \ , \tag{2.11}$$

with $\boldsymbol{P_F} = \boldsymbol{F}(\boldsymbol{F}'\boldsymbol{F})^{-1}\boldsymbol{F}'$ and $(\boldsymbol{I}_n - \boldsymbol{P_F})$ symmetric and idempotent (i.e., $\boldsymbol{P_F}^2 = \boldsymbol{P_F}$ and hence $(\boldsymbol{I}_n - \boldsymbol{P_F})^2 = \boldsymbol{I}_n - 2\boldsymbol{P_F} + \boldsymbol{P_F}^2 = \boldsymbol{I}_n - \boldsymbol{P_F}$). The loss function of a predictive action by the corresponding LS estimator is given by

$$L(\hat{\boldsymbol{\theta}}) = (n-q)s^2 = \| \boldsymbol{y} - \boldsymbol{f}(\hat{\boldsymbol{\theta}}) \|^2 \approx ||(\boldsymbol{I}_n - \boldsymbol{P_F})\boldsymbol{\varepsilon}||^2 = \boldsymbol{\varepsilon}'(\boldsymbol{I}_n - \boldsymbol{P_F})\boldsymbol{\varepsilon} \ , \tag{2.12}$$

where the approximation holds to $o_p(1)$, $s^2 = \hat{\sigma}_\varepsilon^2$ is the unbiased estimate of σ_ε^2 to the order of $1/n$ and asymptotically independent of $\hat{\boldsymbol{\theta}}$, and $(n-q)s^2/\sigma_\varepsilon^2 \approx \boldsymbol{\varepsilon}'(\boldsymbol{I}_n -$

$\boldsymbol{P_F})\boldsymbol{\varepsilon}/\sigma_\varepsilon^2 \sim \chi_{n-q}^2$. The prediction risk of the usual LS estimator is

$$\begin{aligned} P(\hat{\boldsymbol{\theta}}) &= \boldsymbol{E} \parallel y - \boldsymbol{f}(\boldsymbol{\theta}^*) + \boldsymbol{f}(\boldsymbol{\theta}^*) - \boldsymbol{f}(\hat{\boldsymbol{\theta}}) \parallel^2 \\ &= \boldsymbol{E} \parallel \boldsymbol{\varepsilon} \parallel^2 + \boldsymbol{E} \parallel \boldsymbol{f}(\boldsymbol{\theta}^*) - \boldsymbol{f}(\hat{\boldsymbol{\theta}}) \parallel^2 \\ &= n\sigma_\varepsilon^2 + R(\hat{\boldsymbol{\theta}}) \\ &\approx n\sigma_\varepsilon^2 + \boldsymbol{E}||\boldsymbol{P_F}\boldsymbol{\varepsilon}||^2 \\ &= n\sigma_\varepsilon^2 + \sigma_\varepsilon^2 \mathrm{tr}(\boldsymbol{F}(\boldsymbol{F}'\boldsymbol{F})^{-1}\boldsymbol{F}') \\ &= (n+q)\sigma_\varepsilon^2 . \end{aligned} \tag{2.13}$$

In practice, the above asymptotic results need to be further validated by two statistical measures: confidence region and relative curvature. Since as $n \to \infty$,

$$\frac{[L(\boldsymbol{\theta}^*) - L(\hat{\boldsymbol{\theta}})]/q}{L(\hat{\boldsymbol{\theta}})/(n-q)} \approx \frac{\boldsymbol{\varepsilon}'\boldsymbol{P_F}\boldsymbol{\varepsilon}}{\boldsymbol{\varepsilon}'(\boldsymbol{I}_n - \boldsymbol{P_F})\boldsymbol{\varepsilon}} \frac{n-q}{q} = \frac{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)'\boldsymbol{F}'\boldsymbol{F}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)}{qs^2} \sim F_{q,n-q} , \tag{2.14}$$

a commonly used *approximate* $100(1-\alpha)\%$ *confidence region* for $\boldsymbol{\theta}$ is

$$\{\boldsymbol{\theta} : \ (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})'\hat{\boldsymbol{F}}'\hat{\boldsymbol{F}}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \leq qs^2 F_{q,n-q}^\alpha\} , \tag{2.15}$$

where $\hat{\boldsymbol{F}} = \boldsymbol{F}(\hat{\boldsymbol{\theta}})$ and $F_{q,n-q}^\alpha$ is the upper α critical value of the $F_{q,n-q}$ distribution. Another confidence region with practical importance is the *prediction confidence interval* (a.k.a. error bars) for y_t at $\boldsymbol{x} = \boldsymbol{x}_t$, $\forall t = 1, ..., n$. Since $y_t - \hat{y}_t$ is asymptotically $\mathcal{N}(0, \sigma_\varepsilon^2[1 + \boldsymbol{f}_t'(\boldsymbol{F}'\boldsymbol{F})^{-1}\boldsymbol{f}_t])$ and s^2 is asymptotically independent of $y_t - \hat{y}_t$, it is asymptotically true that

$$\frac{y_t - \hat{y}_t}{s\sqrt{1 + \boldsymbol{f}_t'(\boldsymbol{F}'\boldsymbol{F})^{-1}\boldsymbol{f}_t}} \sim t_{n-q} ,$$

hence an approximate $100(1-\alpha)\%$ confidence interval for y_t is

$$\hat{y}_t \pm t_{n-q}^{\alpha/2} s[1 + \hat{\boldsymbol{f}}_t'(\hat{\boldsymbol{F}}'\hat{\boldsymbol{F}})^{-1}\hat{\boldsymbol{f}}_t]^{1/2} , \tag{2.16}$$

where t_{n-q} is the t-distribution with $n-q$ degrees of freedom, and $\boldsymbol{f}_t$ is the t-th q-dimensional row vector of $\boldsymbol{F}$ $(t = 1, ..., n)$.

The confidence interval of (2.16) can be further validated by examining higher-order Taylor approximation in the neighborhood of $\hat{\boldsymbol{\theta}}$:

$$\begin{aligned} \boldsymbol{f}(\boldsymbol{\theta}) - \boldsymbol{f}(\hat{\boldsymbol{\theta}}) &\approx \hat{\boldsymbol{F}}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})'\hat{\boldsymbol{H}}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \\ &= \hat{\boldsymbol{F}}\boldsymbol{\delta} + \frac{1}{2}\boldsymbol{\delta}'\hat{\boldsymbol{H}}\boldsymbol{\delta} , \end{aligned} \tag{2.17}$$

where $\hat{\boldsymbol{H}} = \left[\left(\frac{\partial^2 \boldsymbol{f}(\boldsymbol{\theta})}{\partial\theta_r\partial\theta_s}]\right)\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = [(\hat{\boldsymbol{f}}_{rs})]$ is a $q \times q$ array of n-dimensional vectors $\hat{\boldsymbol{f}}_{rs}$. Evidently, the validity of the linear approximation starting from (2.6) depends on the relative magnitude of the quadratic term $\boldsymbol{\delta}'\hat{\boldsymbol{H}}\boldsymbol{\delta}$ with respect to the linear term $\hat{\boldsymbol{F}}\boldsymbol{\delta}$. Based on the work of Bates and Watts [46] and the concepts from differential geometry, the response surface (a.k.a. expectation surface), $\Omega = \{\boldsymbol{f}(\boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$, can be seen as a q-dimensional surface in the n-dimensional sample space. The column space of $\hat{\boldsymbol{F}}$ in the linear approximation in (2.10) is the tangent plane to the response surface at the point $\hat{\boldsymbol{\theta}}$. By taking the linear approximation in (2.10), one assumes that the response surface can be replaced locally and uniform-coordinately by the tangent plane. To verify this, the quadratic term (i.e., the curvature) in (2.17) needs to be small, compared with the linear term. Bates and Watts first decompose $\hat{\boldsymbol{H}}$ into two components orthogonal to each other, by using the projection matrix $\hat{\boldsymbol{P}}_{\boldsymbol{F}}$:

$$\begin{aligned} \hat{\boldsymbol{H}} &= [(\hat{\boldsymbol{P}}_{\boldsymbol{F}}\hat{\boldsymbol{f}}_{rs})] + [((\boldsymbol{I}_n - \hat{\boldsymbol{P}}_{\boldsymbol{F}})\hat{\boldsymbol{f}}_{rs})] \\ &= \hat{\boldsymbol{H}}^T + \hat{\boldsymbol{H}}^N , \end{aligned} \tag{2.18}$$

and then define two measures of curvature: the tangential *parameter-effects curvature*

$$K_{\boldsymbol{\delta}}^T = \frac{||\boldsymbol{\delta}'\hat{\boldsymbol{H}}^T\boldsymbol{\delta}||}{||\hat{\boldsymbol{F}}\boldsymbol{\delta}||^2},$$

for it depends on the particular parameterization used in $\boldsymbol{f}(\boldsymbol{\theta})$; and the normal *intrinsic curvature*

$$K_{\boldsymbol{\delta}}^N = \frac{||\boldsymbol{\delta}'\hat{\boldsymbol{H}}^N\boldsymbol{\delta}||}{||\hat{\boldsymbol{F}}\boldsymbol{\delta}||^2},$$

for it only reflects the property of the response surface. Notice that the approximate $100(1-\alpha)\%$ confidence region for $\boldsymbol{\theta}$

$$(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})'\hat{\boldsymbol{F}}'\hat{\boldsymbol{F}}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \leq qs^2 F_{q,n-q}^{\alpha}$$

is an ellipsoid centered at $\hat{\boldsymbol{\theta}}$, with a $s\sqrt{qF_{q,n-q}^{\alpha}}$ radius (i.e., $1/(s\sqrt{qF_{q,n-q}^{\alpha}})$ the curvature of the ellipsoid). The curvature measures can be made scale-free by standardizing all the quantities involved with the standard radius $\rho = s\sqrt{q}$ so that the standardized tangential curvature $\kappa_{\boldsymbol{\delta}}^T = K_{\boldsymbol{\delta}}^T\rho$, the standardized normal curvature $\kappa_{\boldsymbol{\delta}}^N = K_{\boldsymbol{\delta}}^N\rho$, and the standardized curvature of the ellipsoid $1/\sqrt{F^{\alpha}_{q,n-q}}$. Bates and Watts [46] suggested that the linear approximation would be tenable if $\kappa^T_{\max}$ is close to zero and $\kappa^N_{\max} < 1/2\sqrt{F^{\alpha}_{q,n-q}}$. For instance, at the level $\alpha = 0.05$, the above inequality allows only a less than 14% departure from the tangent plane approximation.

2.2 Bayesian Approaches: Average Risk Optimality

2.2.1 Single-Prior Bayes and Ordinary Ridge Regression

Given σ^2_ε, we first consider the prior distribution of $\boldsymbol{\theta}$

$$\pi(\boldsymbol{\theta}) \sim \mathcal{N}_q(\mathbf{0}, \frac{\sigma^2_\varepsilon}{k}\boldsymbol{I}) \qquad \text{with} \qquad k > 0 \; . \tag{2.19}$$

With the likelihood function in (2.1), the posterior distribution of $\boldsymbol{\theta}$ given the data is in the form

$$p(\boldsymbol{\theta}|D_n) \propto p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \; . \tag{2.20}$$

A Bayesian estimate of $\boldsymbol{\theta}$ is then obtained by maximizing a posteriori (MAP), i.e., solving the new likelihood equation

$$\begin{aligned}\hat{\boldsymbol{\theta}}(k) &= \arg\max_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}|D_n) \\ &= \arg\max_{\boldsymbol{\theta}}[-\frac{1}{2\sigma^2_\varepsilon}||\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x};\boldsymbol{\theta})||^2 - \frac{k}{2\sigma^2_\varepsilon}||\boldsymbol{\theta}||^2 + \text{constants}] \\ &= \arg\min_{\boldsymbol{\theta}}[||\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x};\boldsymbol{\theta})||^2 + k||\boldsymbol{\theta}||^2] \; ,\end{aligned} \tag{2.21}$$

which is equivalent to add a penalty term to the loss function with k in the place of the smoothing parameter λ in (1.12). Then the Bayes estimator of $\boldsymbol{\theta}$ can be written in the Newton-Raphson form

$$\hat{\boldsymbol{\theta}}(k) \approx \boldsymbol{\theta}^* + (\boldsymbol{F}'\boldsymbol{F} + k\boldsymbol{I}_q)^{-1}[\boldsymbol{F}'\boldsymbol{\varepsilon} - k\boldsymbol{\theta}^*] \; , \tag{2.22}$$

i.e., the normal approximation of the posterior density of $\hat{\boldsymbol{\theta}}(k)$ given the data is

$$\begin{aligned}p(\hat{\boldsymbol{\theta}}(k)|D_n) &\sim \mathcal{N}_q([\boldsymbol{I}_q - k(\boldsymbol{F}'\boldsymbol{F} + k\boldsymbol{I}_q)^{-1}]\boldsymbol{\theta}^*, \boldsymbol{V}(k)) \\ &\sim \mathcal{N}_q((\boldsymbol{F}'\boldsymbol{F} + k\boldsymbol{I}_q)^{-1}\boldsymbol{F}'\boldsymbol{F}\boldsymbol{\theta}^*, \boldsymbol{V}(k)) \; ,\end{aligned}$$

where $\boldsymbol{V}(k) = \sigma_\varepsilon^2(\boldsymbol{F}'\boldsymbol{F} + k\boldsymbol{I}_q)^{-1}\boldsymbol{F}'\boldsymbol{F}(\boldsymbol{F}'\boldsymbol{F} + k\boldsymbol{I}_q)^{-1} + k^2(\boldsymbol{F}'\boldsymbol{F} + k\boldsymbol{I}_q)^{-1}\boldsymbol{\theta}^*(\boldsymbol{\theta}^*)'(\boldsymbol{F}'\boldsymbol{F} + k\boldsymbol{I}_q)^{-1}$. The estimator $\hat{\boldsymbol{\theta}}(k)$ shrinks the ML estimator $\hat{\boldsymbol{\theta}}$ toward the priori zero mean, and follows a similar fashion of Hoerl and Kennard's ordinary ridge estimator for linear regression model in (1.22) and (1.23).

The risk behavior of $\hat{\boldsymbol{\theta}}(k)$ shares the similar characteristic of an ordinary ridge procedure as well. From (2.4) with $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}(k)$, we have

$$\begin{aligned}
\boldsymbol{f}(\hat{\boldsymbol{\theta}}(k)) - \boldsymbol{f}(\boldsymbol{\theta}^*) &\approx \boldsymbol{F}(\hat{\boldsymbol{\theta}}(k) - \boldsymbol{\theta}^*) \\
&= \boldsymbol{F}(\hat{\boldsymbol{\theta}}(k) - \boldsymbol{E}\hat{\boldsymbol{\theta}}(k)) + \boldsymbol{F}(\boldsymbol{E}\hat{\boldsymbol{\theta}}(k) - \boldsymbol{\theta}^*) \\
&= \boldsymbol{F}(\hat{\boldsymbol{\theta}}(k) - \boldsymbol{E}\hat{\boldsymbol{\theta}}(k)) - k\boldsymbol{F}(\boldsymbol{F}'\boldsymbol{F} + k\boldsymbol{I}_q)^{-1}\boldsymbol{\theta}^* \ . \qquad (2.23)
\end{aligned}$$

Hence, the prediction risk of the Bayes estimator $\hat{\boldsymbol{\theta}}(k)$ can be written in the form of typical bias-variance decomposition for mean-squared prediction error (MSPE) as

$$\begin{aligned}
P(\hat{\boldsymbol{\theta}}(k)) &= \boldsymbol{E}[l(\hat{\boldsymbol{\theta}}(k))] = \boldsymbol{E} \parallel \boldsymbol{y} - \boldsymbol{f}(\hat{\boldsymbol{\theta}}(k)) \parallel^2 \\
&= \boldsymbol{E} \parallel \boldsymbol{\varepsilon} \parallel^2 + \boldsymbol{E} \parallel \boldsymbol{f}(\boldsymbol{\theta}^*) - \boldsymbol{f}(\hat{\boldsymbol{\theta}}(k)) \parallel^2 \\
&= n\sigma_\varepsilon^2 + \boldsymbol{E}\left[(\hat{\boldsymbol{\theta}}(k) - \boldsymbol{\theta}^*)'\boldsymbol{F}'\boldsymbol{F}(\hat{\boldsymbol{\theta}}(k) - \boldsymbol{\theta}^*)\right] + o_p(1) \\
&= n\sigma_\varepsilon^2 + \Big\{\boldsymbol{E}\left[(\hat{\boldsymbol{\theta}}(k) - \boldsymbol{E}\hat{\boldsymbol{\theta}}(k))'\boldsymbol{F}'\boldsymbol{F}(\hat{\boldsymbol{\theta}}(k) - \boldsymbol{E}\hat{\boldsymbol{\theta}}(k))\right] + \\
&\quad k^2\boldsymbol{\theta}^{*'}(\boldsymbol{F}'\boldsymbol{F} + k\boldsymbol{I}_q)^{-1}\boldsymbol{F}'\boldsymbol{F}(\boldsymbol{F}'\boldsymbol{F} + k\boldsymbol{I}_q)^{-1}\boldsymbol{\theta}^*\Big\} + o_p(1) \\
&= n\sigma_\varepsilon^2 + \sigma_\varepsilon^2\mathrm{tr}\left(\boldsymbol{F}(\boldsymbol{F}'\boldsymbol{F} + k\boldsymbol{I}_q)^{-1}\boldsymbol{F}'\boldsymbol{F}(\boldsymbol{F}'\boldsymbol{F} + k\boldsymbol{I}_q)^{-1}\boldsymbol{F}'\right) + \\
&\quad k^2\boldsymbol{\theta}^{*'}(\boldsymbol{F}'\boldsymbol{F} + k\boldsymbol{I}_q)^{-1}\boldsymbol{F}'\boldsymbol{F}(\boldsymbol{F}'\boldsymbol{F} + k\boldsymbol{I}_q)^{-1}\boldsymbol{\theta}^* + o_p(1) \\
&= n\sigma_\varepsilon^2 + \mathrm{var}^*[\hat{\boldsymbol{\theta}}(k)] + \mathrm{bias}^{*2}[\hat{\boldsymbol{\theta}}(k)] + o_p(1) \\
&= n\sigma_\varepsilon^2 + R^*(\hat{\boldsymbol{\theta}}(k)) + o_p(1) \\
&= n\sigma_\varepsilon^2 + R(\hat{\boldsymbol{\theta}}(k)) \ , \qquad (2.24)
\end{aligned}$$

with bias^{*2}, var^* and $R^*(\hat{\boldsymbol{\theta}}(k))$ the model squared bias, variance and risk in the asymptotic sense ($n \to \infty$). Let $\boldsymbol{\Lambda}$ be the diagonal matrix of eigenvalues of $\boldsymbol{F}'\boldsymbol{F}$ and $\boldsymbol{G}$ be the orthogonal transformation such that $\boldsymbol{F}'\boldsymbol{F} = (\boldsymbol{G}^{-1})'\boldsymbol{\Lambda}\boldsymbol{G}^{-1}$, $(\boldsymbol{G}^{-1})'\boldsymbol{G}^{-1} = \boldsymbol{I}_n$, $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_i)$, $\lambda_{max} = \lambda_1 \geq \lambda_2 \geq ... \geq \lambda_q = \lambda_{min} > 0$, and $\boldsymbol{\gamma} = \boldsymbol{G}^{-1}\boldsymbol{\theta}^*$. Then the asymptotic risk is given by

$$R^*(\hat{\boldsymbol{\theta}}(k)) = \sigma_\varepsilon^2 \sum_{i=1}^{q} \frac{\lambda_i^2}{(\lambda_i + k)^2} + k^2 \sum_{i=1}^{q} \frac{\gamma_i^2 \lambda_i}{(\lambda_i + k)^2} \ . \tag{2.25}$$

The properties of the asymptotic variance and squared bias terms from the asymptotic risk are instantly followed (cf. Figure 1.6):

1. For the variance term:

   (a) $\text{var}^*[\hat{\boldsymbol{\theta}}(0)] = q\sigma_\varepsilon^2$;

   (b) $\lim_{k\to\infty} \text{var}^*[\hat{\boldsymbol{\theta}}(k)] = 0$;

   (c) $\frac{d}{dk}(\text{var}^*[\hat{\boldsymbol{\theta}}(k)]) = -2\sigma_\varepsilon^2 \sum_{i=1}^{q} \frac{\lambda_i^2}{(\lambda_i + k)^3} < 0$, so $\text{var}^*[\hat{\boldsymbol{\theta}}(k)]$ is a continuous, monotone decreasing function for $k \geq 0$.

2. For the squared bias term:

   (a) $\text{bias}^{*2}[\hat{\boldsymbol{\theta}}(0)] = 0$;

   (b) $\lim_{k\to\infty} \text{bias}^{*2}[\hat{\boldsymbol{\theta}}(k)] = (\boldsymbol{\theta}^*)'\boldsymbol{\theta}^*$;

   (c) $\lim_{||\boldsymbol{\theta}^*||^2\to\infty} \text{bias}^{*2}[\hat{\boldsymbol{\theta}}(k)] \to \infty$;

   (d) $\frac{d}{dk}(\text{bias}^{*2}[\hat{\boldsymbol{\theta}}(k)]) = 2k \sum_{i=1}^{q} \frac{\gamma_i^2 \lambda_i}{(\lambda_i + k)^3} > 0$, so $\text{bias}^{*2}[\hat{\boldsymbol{\theta}}(k)]$ is a continuous, monotone increasing function for $k \geq 0$.

3. There always exists a $k > 0$ such that $R^*(\hat{\boldsymbol{\theta}}(k)) < R^*(\hat{\boldsymbol{\theta}})$. Since $\text{var}^*[\hat{\boldsymbol{\theta}}(0)] = q\sigma_\varepsilon^2$, $\text{bias}^{*2}[\hat{\boldsymbol{\theta}}(0)] = 0$, $\text{var}^*[\hat{\boldsymbol{\theta}}(k)]$ and $\text{bias}^{*2}[\hat{\boldsymbol{\theta}}(k)]$ are monotonically decreasing and increasing for $k > 0$ respectively, we only have to show that $\exists k > 0$ such that $\frac{d}{dk}R^*(\hat{\boldsymbol{\theta}}(k)) < 0$, i.e.,

$$\begin{aligned} \frac{d}{dk}R^*(\hat{\boldsymbol{\theta}}(k)) &= \frac{d}{dk}(\text{var}^*[\hat{\boldsymbol{\theta}}(k)]) + \frac{d}{dk}(\text{bias}^{*2}\hat{\boldsymbol{\theta}}(k)]) \\ &= -2\sigma_\varepsilon^2 \sum_{i=1}^{q} \frac{\lambda_i^2}{(\lambda_i + k)^3} + 2k \sum_{i=1}^{q} \frac{\gamma_i^2 \lambda_i}{(\lambda_i + k)^3} < 0 \ . \end{aligned}$$

   Hence, the condition is given by

$$k < \frac{\sigma_\varepsilon^2 \lambda_{min}^2}{\lambda_{max} \gamma_{max}^2} \ . \tag{2.26}$$

The properties of $R^*(\hat{\boldsymbol{\theta}}(k))$ show that it will go through a minimum (cf. Figures 1.5 and 2.2). Since $\lim_{k\to\infty} \text{bias}^{*2}[\hat{\boldsymbol{\theta}}(k)] = \boldsymbol{\theta}^{*'}\boldsymbol{\theta}^*$, this minimum will move toward $k = 0$ as the squared length of the unknown regression coefficients increases. Although for fixed $\boldsymbol{x}_t$'s and $\boldsymbol{\theta}$ there is a neighborhood of zero for k within which $\hat{\boldsymbol{\theta}}(k)$ has smaller asymptotic risk, no fixed $k > 0$ can dominate the ML (LS) estimator $\hat{\boldsymbol{\theta}}$ for all possible $\boldsymbol{x}_t$'s and $\boldsymbol{\theta}$. It is also impossible to determine an optimal choice of k before a Newton-Raphson iteration is implemented. Furthermore, the confidence intervals are only improved when the right choice of k is made, that can be seen in the following formulation

$$\begin{aligned}&[\hat{y}_t(k) + k\hat{\boldsymbol{f}}_t(\hat{\boldsymbol{F}}'\hat{\boldsymbol{F}} + k\boldsymbol{I}_q)^{-1}\hat{\boldsymbol{\theta}}(k)]\\ &\pm t_{n-q}^{\alpha/2} s[1 + \hat{\boldsymbol{f}}_t'(\hat{\boldsymbol{F}}'\hat{\boldsymbol{F}} + k\boldsymbol{I}_q)^{-1}\hat{\boldsymbol{F}}'\hat{\boldsymbol{F}}(\hat{\boldsymbol{F}}'\hat{\boldsymbol{F}} + k\boldsymbol{I}_q)^{-1}\hat{\boldsymbol{f}}_t\\ &+(k\hat{\boldsymbol{f}}_t'(\hat{\boldsymbol{F}}'\hat{\boldsymbol{F}} + k\boldsymbol{I}_q)^{-1}\hat{\boldsymbol{\theta}}(k))^2]^{1/2}\ ,\end{aligned} \tag{2.27}$$

where all the $\hat{\cdot}$'s are evaluated at $\hat{\boldsymbol{\theta}}(k)$. For example, the 95% confidence intervals of a single-prior Bayes neural network model with $h = 9$ and $k = 0.01$ covers 98.18% of data points, that is higher than 97.27% from the LS method. Nevertheless, it is also numerically stabilized with the matrix singularity reduced by the hyperparameter (see Figure 2.1).

A reasonable approach in practice is to train a model by the data set in hand under different choices of k and to evaluate the prediction performance of the trained models through either cross-validation or bootstrap resampling methods, then to settle for a trained model with the lowest prediction error (as shown in Figure 1.5). The advantages of this simplistic approach is that it keeps the numerical stability of the ridge procedure and reaches a relatively good bias-variance trade-off with moderate computational cost. However, there are also a few drawbacks: it does not possess Bayesian robustness generically, the accuracy of the 'optimal' k is often compromised by consideration of computational cost, and there is no closed formulation for the approach as a whole that makes it difficult to both analytic verification of its optimality and numerical implementation as a default 'standard' algorithm.

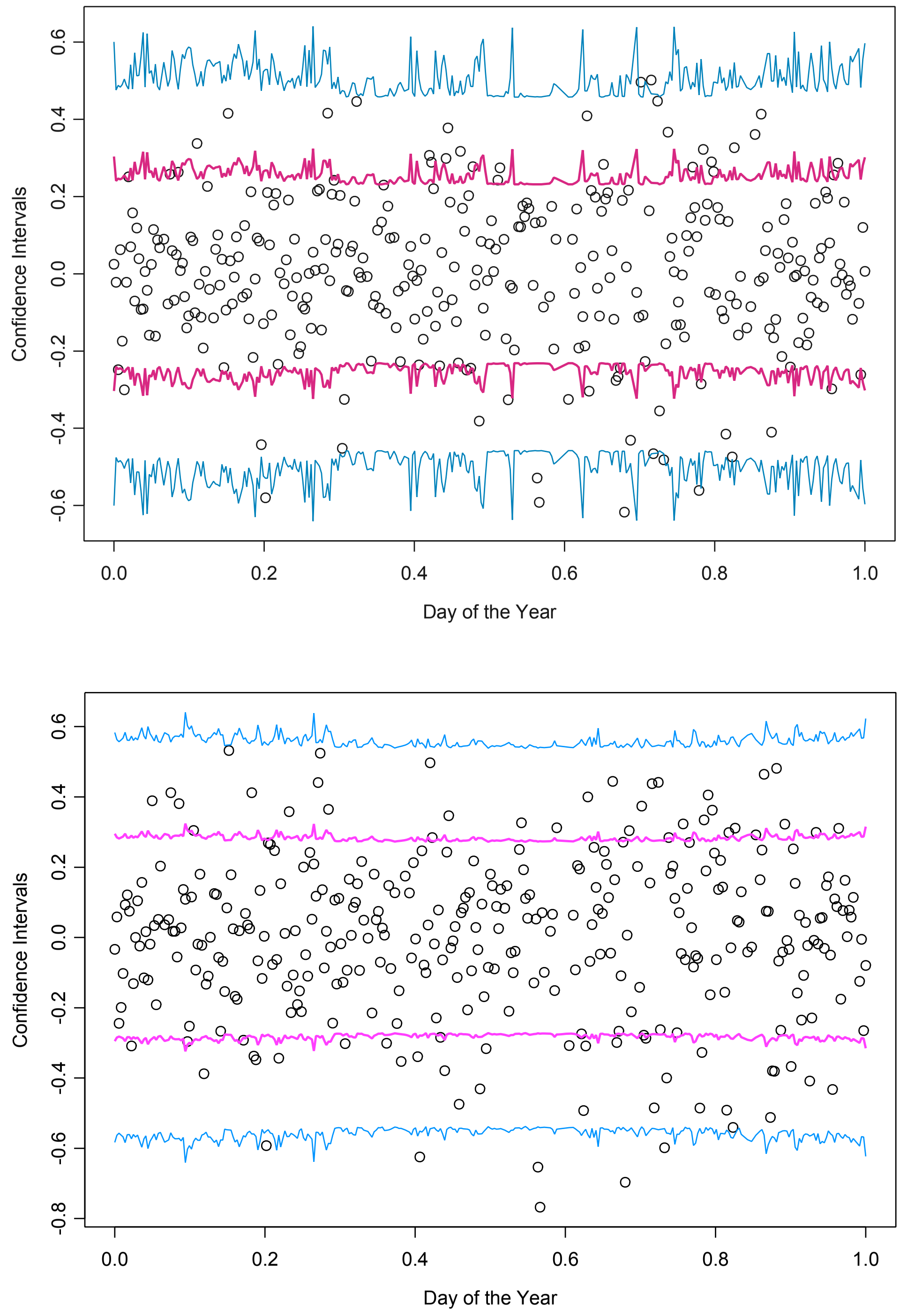


Fig. 2.1. The confidence intervals defined in (2.16) and (2.27) are plotted with the observed data points (y_t's) centered around the fitted response values ($\hat{y}_t$'s), against `doy` (one of nine predictors). The 68% confidence region is located between the two thick lines and the 95% confidence region is between the thin lines. The upper plot is from LS method, and the lower plot is the result of the single-prior Bayes method. Evidently, the latter is stabilized numerically through the ridge procedure.

2.2.2 Empirical Bayes

A natural generalization of single-prior Bayes estimation is to treat the ridge parameter k as unknown along with σ_ε^2 as well, so that one deals with a class of prior density with varying mean vector and variance-covariance matrix instead of fixed ones. The focus is once again on one's treatment of the prior density function. In general, one can assume the prior distribution of $\boldsymbol{\theta}$, $\pi(\boldsymbol{\theta})$, to be modeled either by parametric densities with unknown hyperparameters (e.g., a normal density with unknown k and σ_ε^2 is used in (2.19) or by nonparametrics. There are two basic strategies to solve the new uncertainty in prior.

Firstly, one can pick the most probable prior by estimating the hyperparameters based on the marginal distribution of the data. Suppose that the likelihood function $p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta},\sigma_\varepsilon^2)$ be the same as in (2.1) and a prior density of $\boldsymbol{\theta}$, $\pi(\boldsymbol{\theta}|k;\sigma_\varepsilon^2)$ in (2.19) with k unknown, then the marginal density of the data is

$$m(\boldsymbol{y}|\boldsymbol{x};\pi) = \int p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta},\sigma_\varepsilon^2)\pi(\boldsymbol{\theta}|k;\sigma_\varepsilon^2)d\boldsymbol{\theta} \ .$$

The marginal distribution $m(\boldsymbol{y}|\boldsymbol{x};\pi)$ can be considered as a likelihood function for the prior $\pi(\boldsymbol{\theta}|k;\sigma_\varepsilon^2)$ indexed by the unknown k and σ_ε^2. $m(\boldsymbol{y}|\boldsymbol{x};\pi(k_1)) > m(\boldsymbol{y}|\boldsymbol{x};\pi(k_2))$ indicates that the data provides more support for the choice of k_1 than k_2. A type-II maximum likelihood prior (ML-II prior) is the one that satisfies

$$\hat{\pi} : m(\boldsymbol{y}|\boldsymbol{x};\hat{\pi}) = \arg\max_{k,\sigma_\varepsilon^2} m(\boldsymbol{y}|\boldsymbol{x};\pi) \ .$$

A ML estimate of the hyperparameter k (the same for σ_ε^2) can then be obtained by solving the new likelihood equation

$$\frac{\partial}{\partial k} m(\boldsymbol{y}|\boldsymbol{x};\pi(k)) = 0 \ .$$

Once the empirical prior is chosen, the rest of analysis can be carried out in a typical Bayesian fashion.

Secondly, the uncertainty in prior can be carried over into the posterior distribution of the parameters given data, and the hyperparameters in the prior are treated as part of model parameterization and optimized together with other model parameters. Assume a more general prior of $\boldsymbol{\theta}$ in the form

$$\mathcal{N}_q(\mathbf{0}, \sigma_\varepsilon^2 \boldsymbol{K}^{-1}) \; , \tag{2.28}$$

where $\boldsymbol{K} = \mathrm{diag}(k_i)$ (i.e., each parameter in the model is assigned a different hyperparameter of variance). In the Newton-Raphson formulation, this leads to the following iteration

$$\hat{\boldsymbol{\theta}}_{\tau+1} = \hat{\boldsymbol{\theta}}_\tau + (\hat{\boldsymbol{F}}_\tau' \hat{\boldsymbol{F}}_\tau + \hat{\boldsymbol{K}}_\tau)^{-1} (\hat{\boldsymbol{F}}_\tau' \hat{\boldsymbol{\varepsilon}}_\tau - \hat{\boldsymbol{K}}_\tau \hat{\boldsymbol{\theta}}_\tau) \; , \tag{2.29}$$

where $\hat{\boldsymbol{F}}_\tau = \boldsymbol{F}(\hat{\boldsymbol{\theta}}_\tau)$, $\hat{\boldsymbol{\varepsilon}}_\tau = \boldsymbol{y} - \boldsymbol{f}(\hat{\boldsymbol{\theta}}_\tau)$ and $\hat{\boldsymbol{K}}_\tau = \mathrm{diag}(\hat{k}_{\tau i})$. The posterior distribution of the Bayesian estimator $\hat{\boldsymbol{\theta}}_{\tau+1}$ at the $(\tau+1)$-th step of iteration given the data can be seen as

$$\hat{\boldsymbol{\theta}}_{\tau+1} \sim \mathcal{N}_q((\hat{\boldsymbol{F}}_\tau' \hat{\boldsymbol{F}}_\tau + \hat{\boldsymbol{K}}_\tau)^{-1} \hat{\boldsymbol{F}}_\tau' \hat{\boldsymbol{F}}_\tau \boldsymbol{\theta}^*, \hat{\boldsymbol{V}}_\tau) \; , \tag{2.30}$$

where

$$\begin{aligned} \hat{\boldsymbol{V}}_\tau &= \hat{\sigma}_\varepsilon^2 (\hat{\boldsymbol{F}}_\tau' \hat{\boldsymbol{F}}_\tau + \hat{\boldsymbol{K}}_\tau)^{-1} \hat{\boldsymbol{F}}_\tau' \hat{\boldsymbol{F}}_\tau (\hat{\boldsymbol{F}}_\tau' \hat{\boldsymbol{F}}_\tau + \hat{\boldsymbol{K}}_\tau)^{-1} + \\ & \quad (\hat{\boldsymbol{F}}_\tau' \hat{\boldsymbol{F}}_\tau + \hat{\boldsymbol{K}}_\tau)^{-1} \hat{\boldsymbol{K}}_\tau \boldsymbol{\theta}^* (\boldsymbol{\theta}^*)' \hat{\boldsymbol{K}}_\tau (\hat{\boldsymbol{F}}_\tau' \hat{\boldsymbol{F}}_\tau + \hat{\boldsymbol{K}}_\tau)^{-1} \; . \end{aligned}$$

With a similar canonical reduction such that $\hat{\boldsymbol{F}}_\tau' \hat{\boldsymbol{F}}_\tau = (\hat{\boldsymbol{G}}_\tau^{-1})' \hat{\boldsymbol{\Lambda}}_\tau \hat{\boldsymbol{G}}_\tau^{-1}$, $(\hat{\boldsymbol{G}}_\tau^{-1})' \hat{\boldsymbol{G}}_\tau^{-1} = \boldsymbol{I}_n$, $\hat{\boldsymbol{\Lambda}}_\tau = \mathrm{diag}(\hat{\lambda}_{\tau i})$, $\hat{\boldsymbol{\gamma}}_\tau = \hat{\boldsymbol{G}}_\tau^{-1} \hat{\boldsymbol{\theta}}_\tau$ and $\hat{\boldsymbol{\gamma}}_{\tau 0} = \hat{\boldsymbol{G}}_\tau^{-1} \boldsymbol{\theta}^*$, the asymptotic risk of $\hat{\boldsymbol{\theta}}_{\tau+1}$ is sought to be minimized to obtain the 'optimal' choice of $\hat{\boldsymbol{K}}_\tau = \mathrm{diag}(\hat{k}_{\tau i})$ for the coming $(\tau+1)$-th iteration.

$$\begin{aligned} \frac{\partial}{\partial \hat{k}_{\tau i}} R^*(\hat{\boldsymbol{\theta}}_{\tau+1}) &= \frac{\partial}{\partial \hat{k}_{\tau i}} [\sum_{i=1}^{q} \frac{\hat{\sigma}_\varepsilon^2 \hat{\lambda}_{\tau i}^2 + \hat{\gamma}_{\tau i}^2 \hat{\lambda}_{\tau i} \hat{k}_{\tau i}^2}{(\hat{\lambda}_{\tau i} + \hat{k}_{\tau i})^2}] \\ &= \sum_{i=1}^{q} \frac{2 \hat{\lambda}_{\tau i}^2 (\hat{\lambda}_{\tau i} + \hat{k}_{\tau i})(\hat{\gamma}_{\tau i}^2 \hat{k}_{\tau i} - \hat{\sigma}_\varepsilon^2)}{(\hat{\lambda}_{\tau i} + \hat{k}_{\tau i})^4} = 0 \; , \end{aligned}$$

which yields that

$$\hat{k}_{\tau i} = \frac{\hat{\sigma}_\varepsilon^2}{\hat{\gamma}_{\tau i}^2} = \frac{\hat{\sigma}_\varepsilon^2}{\hat{\theta}_{\tau i}^2} \; , \tag{2.31}$$

with $\hat{\sigma}_\varepsilon^2$ replaced by either a fixed estimate $\hat{\sigma}_\varepsilon^2$ from a trained LS model or an iterative version $\hat{\sigma}_{\varepsilon\tau}^2 = \frac{1}{n-q}||\boldsymbol{y} - \boldsymbol{f}(\hat{\boldsymbol{\theta}}_\tau)||^2$. This approach is equivalent to the adaptive ridge procedure for linear regression (cf. [47, 48, 49]) and is advocated by MacKay [38] and Neal [39] for neural network regression model in (1.2).

It is easy to show that the above two empirical Bayesian approaches deliver the essentially same adaptive version of ridge procedure if a normal prior is assumed. However, it also can be shown analytically and corroborated numerically that this kind of simultaneous adaptive iteration of both the parameters and hyperparameters tends to shrink the parameters too much toward the prior mean (usually set to zero), because $\hat{k}_{\tau i}$ is often too large. This can result in poor prediction performance with the bias-variance compromise heavily tilted toward a very low model variance but a way too high bias component (see Figure 2.2). The coverage probability of its corresponding confidence intervals is also lowered (see Figure 2.3).

Adopting the technique introduced in [50] with the above-mentioned canonical reduction and the adaptive choice of $\boldsymbol{K}$ from (2.31), the adaptive procedure in (2.29) and (2.30) is approximately equivalent to

$$\hat{\boldsymbol{\Gamma}}_{\tau+1} := [\hat{\boldsymbol{\Lambda}}_\tau + \hat{\sigma}_{\varepsilon\tau}^2 \hat{\boldsymbol{\Gamma}}_\tau^{-2}]^{-1} \hat{\boldsymbol{\Lambda}}_\tau \hat{\boldsymbol{\Gamma}}_{\tau 0} = [\boldsymbol{I}_q + \hat{\sigma}_{\varepsilon\tau}^2 \hat{\boldsymbol{\Lambda}}_\tau^{-1} \hat{\boldsymbol{\Gamma}}_\tau^{-2}]^{-1} \hat{\boldsymbol{\Gamma}}_{\tau 0} \ , \tag{2.32}$$

where ':=' means that the right side is the mean vector of the posterior distribution of the left side given the data, $\hat{\boldsymbol{\Lambda}}_\tau = \hat{\boldsymbol{G}}_\tau^{-1} \hat{\boldsymbol{F}}_\tau' \hat{\boldsymbol{F}}_\tau (\hat{\boldsymbol{G}}_\tau^{-1})'$, and the q-vectors $\hat{\boldsymbol{\gamma}}_\tau$ and $\hat{\boldsymbol{\gamma}}_{\tau 0}$ are represented as diagonal matrices

$$\hat{\boldsymbol{\Gamma}}_\tau = \begin{bmatrix} \hat{\gamma}_{\tau 1} & 0 & \cdots & 0 \\ 0 & \hat{\gamma}_{\tau 2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{\gamma}_{\tau q} \end{bmatrix} ,$$

and

$$\hat{\boldsymbol{\Gamma}}_{\tau 0} = \begin{bmatrix} [\hat{\boldsymbol{G}}_\tau^{-1}\boldsymbol{\theta}^*]_1 & 0 & \cdots & 0 \\ 0 & [\hat{\boldsymbol{G}}_\tau^{-1}\boldsymbol{\theta}^*]_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & [\hat{\boldsymbol{G}}_\tau^{-1}\boldsymbol{\theta}^*]_q \end{bmatrix} .$$

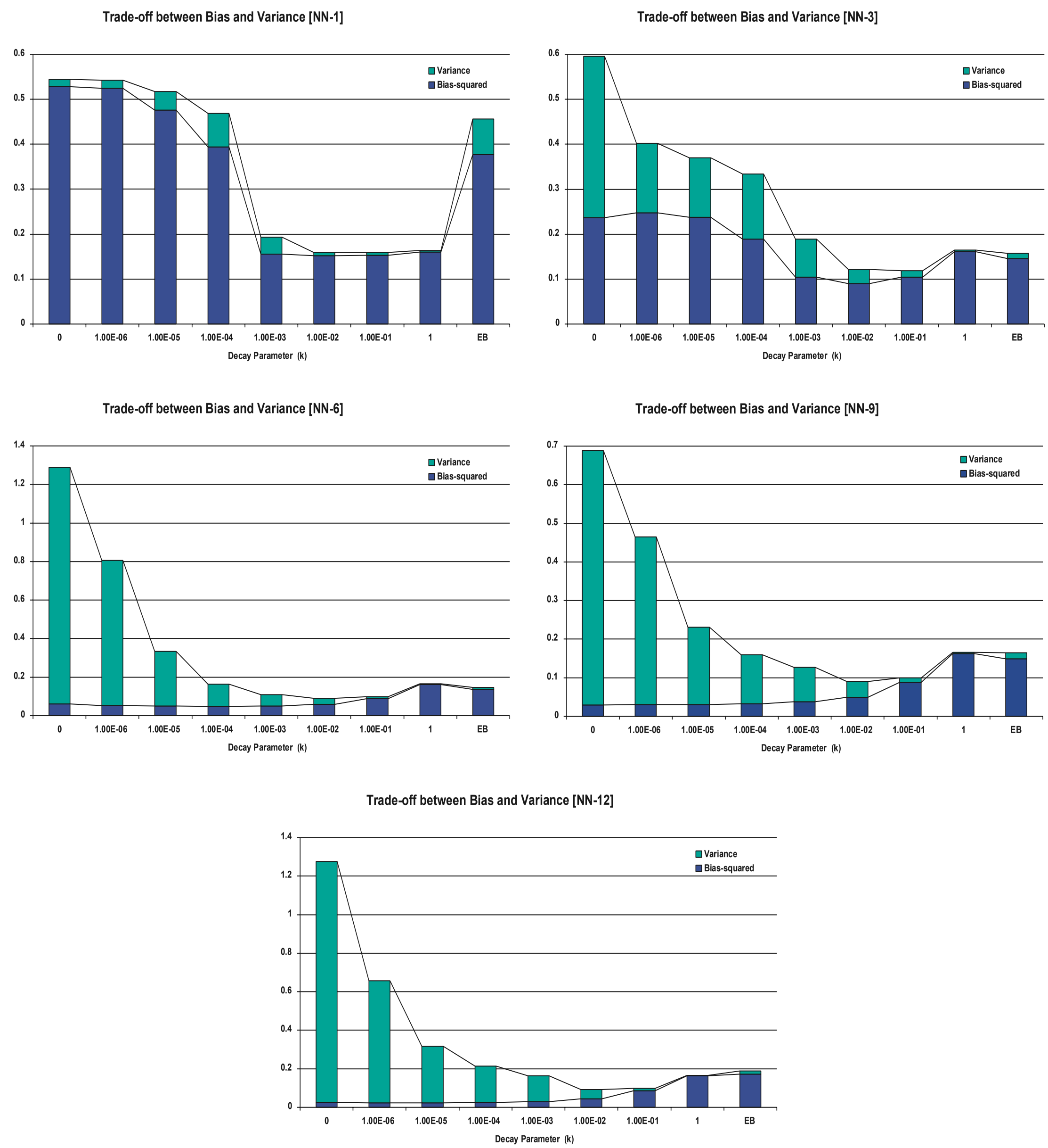


Fig. 2.2. The empirical Bayesian approach (EB) is compared with the single-prior Bayes methods used in neural network regression models with various number of hidden units (h=1,3,6,9,12) on ozone data, in terms of prediction performance (MSPE) and its bias-variance decomposition. Notice that the MSPEs of the models trained by empirical Bayesian method discussed in Section 2.2.2 have a high bias-squared components, which prevent them from reaching 'optimal' bias-variance trade-off in all occasions, even though EB method usually delivers an improved performance over ML (LS) method. The high model bias seems originated from the over-shrunk parameter estimates (see Figures 2.5). [Note: Bootstrap estimates of model bias and variance are used with 1000 resamples for each case. The bias-squared components also include the contribution from the model residuals, so that they can be rather high when the models are not adequate as for the cases of $h = 1, 3$.]

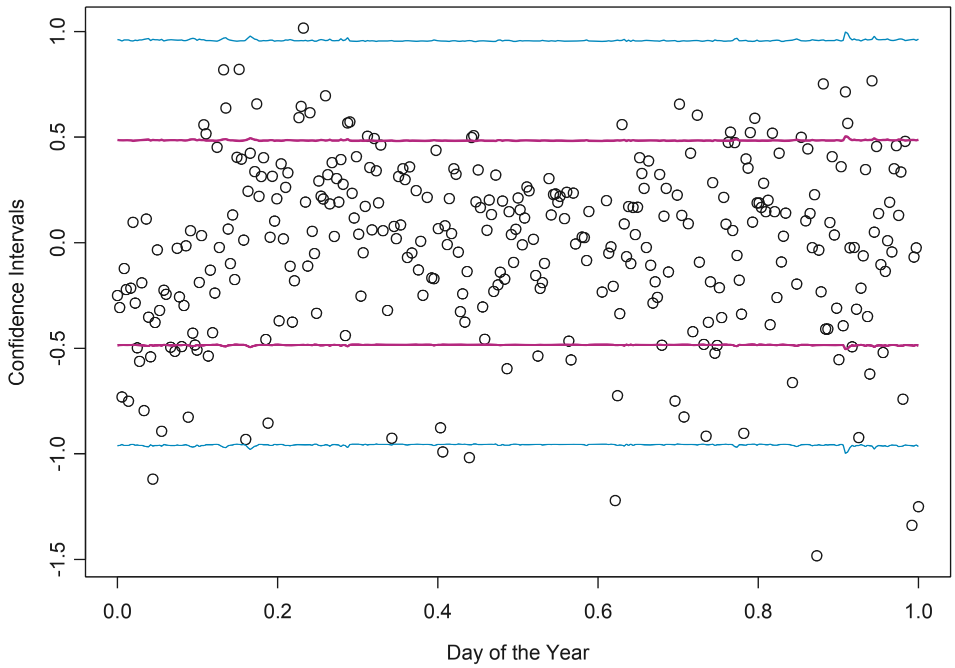


Fig. 2.3. The confidence intervals defined in (2.27) are plotted with the observed data points (y_t's) centered around the fitted response values ($\hat{y}_t$'s), and k replaced by $\hat{k}$ which is the value when the iteration procedure is ended. Compared with Figure 2.1, the smooth contours indicate the adaptive choice of k is too large. The coverage probability of 95% confidence region is 97.57%, down from 98.18% of the single-prior Bayes method with $k = 0.01$.

Let $\hat{\boldsymbol{A}}_\tau = \hat{\boldsymbol{\Lambda}}_\tau / \hat{\sigma}^2_{\varepsilon\tau}$, then (2.32) becomes

$$\hat{\boldsymbol{\Gamma}}_{\tau+1} := [\boldsymbol{I}_q + \hat{\boldsymbol{A}}_\tau^{-1} \hat{\boldsymbol{\Gamma}}_\tau^{-2}]^{-1} \hat{\boldsymbol{\Gamma}}_{\tau 0} \,.$$

Since both matrices $[\boldsymbol{I}_q + \hat{\boldsymbol{A}}_\tau^{-1} \hat{\boldsymbol{\Gamma}}_\tau^{-2}]$ and $\hat{\boldsymbol{\Gamma}}_{\tau 0}$ are diagonal and commute,

$$\begin{aligned} \hat{\boldsymbol{\Gamma}}^{-2}_{\tau+1} &:= \hat{\boldsymbol{\Gamma}}^{-1}_{\tau 0} [\boldsymbol{I}_q + \hat{\boldsymbol{A}}_\tau^{-1} \hat{\boldsymbol{\Gamma}}_\tau^{-2}] \hat{\boldsymbol{\Gamma}}^{-1}_{\tau 0} [\boldsymbol{I}_q + \hat{\boldsymbol{A}}_\tau^{-1} \hat{\boldsymbol{\Gamma}}_\tau^{-2}] \\ &= \hat{\boldsymbol{\Gamma}}^{-2}_{\tau 0} [\boldsymbol{I}_q + \hat{\boldsymbol{A}}_\tau^{-1} \hat{\boldsymbol{\Gamma}}_\tau^{-2}]^2 \,. \end{aligned}$$

Multiplying both sides by $\hat{\boldsymbol{A}}^{-1}_{\tau+1}$ gives

$$\hat{\boldsymbol{A}}^{-1}_{\tau+1} \hat{\boldsymbol{\Gamma}}^{-2}_{\tau+1} := \hat{\boldsymbol{A}}^{-1}_{\tau+1} \hat{\boldsymbol{\Gamma}}^{-2}_{\tau 0} [\boldsymbol{I}_q + \hat{\boldsymbol{A}}_\tau^{-1} \hat{\boldsymbol{\Gamma}}_\tau^{-2}]^2 \,. \tag{2.33}$$

Assume that the iterative procedure in (2.33) converges in the sense that $\lim_{\tau\to\infty} \boldsymbol{\Gamma}_\tau^{-2} = \hat{\boldsymbol{\Gamma}}^{\star -2}$. Then it is evident that all other quantities involved are convergent as well,

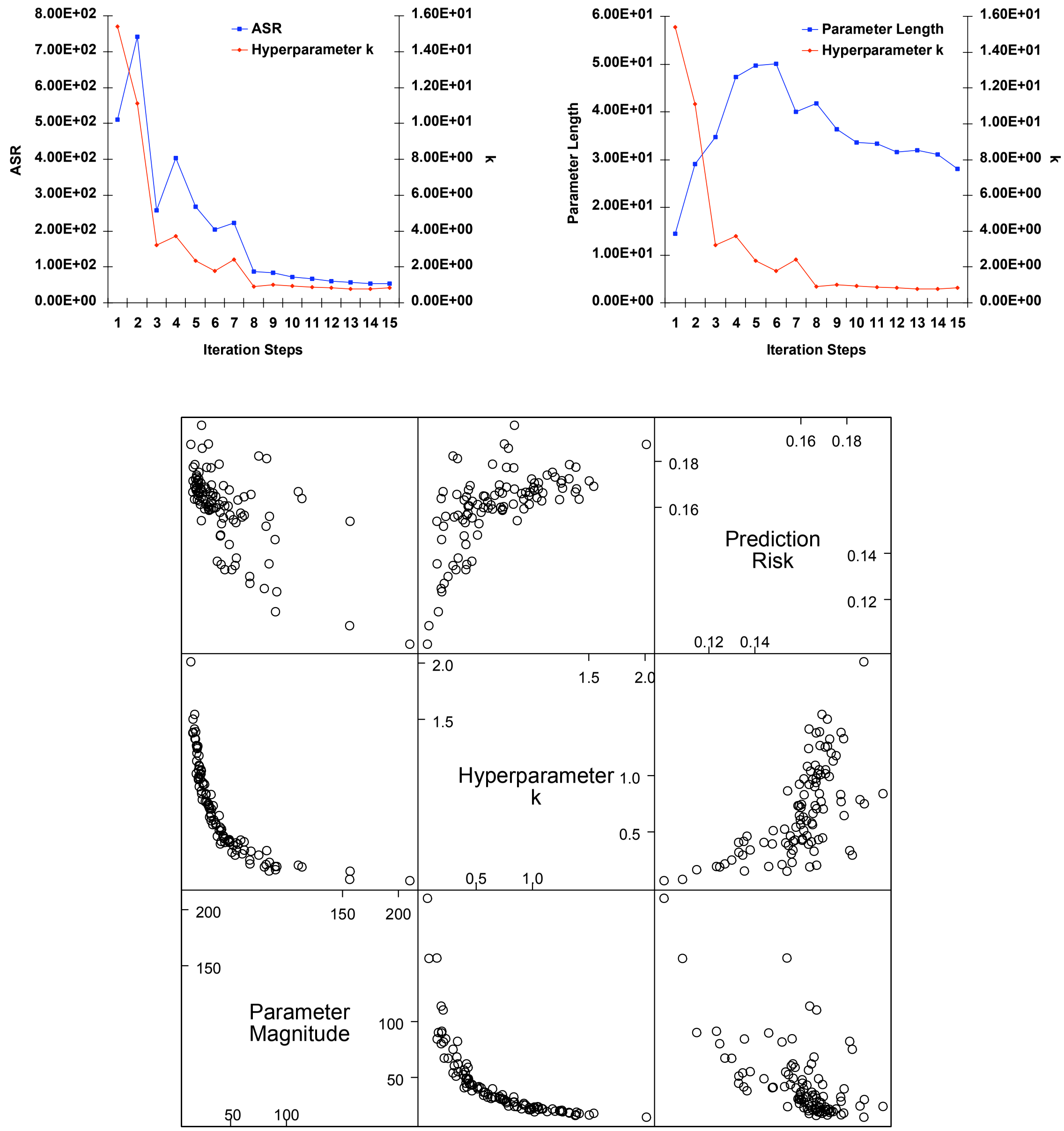


Fig. 2.4. The upper panel provides a look at the iterative changes of three quantities (hyperparameter k, parameter magnitude $||\boldsymbol{\theta}||^2$ and ASR) in a Newton-Raphson procedure when the empirical Bayesian approach is used in a neural network ($h = 9$) on ozone data. The iteratively updated hyperparameter k's are way higher (in the neighborhood of 1) than the 'optimal' choice (in the neighborhood of 0.01) as indicated by the simulations on the single-prior Bayesian method. The relations among estimated parameter magnitude, hyperparameter k and prediction risk are illustrated in the lower panel by 100 neural network models trained by empirical Bayes method with 9 hidden units on ozone data. All three quantities take their values at the end of Newton-Raphson iteration. Evidently, when hyperparameter increases, the parameter length decreases as the result of more shrinkage, and the prediction risk increases due to the higher model bias.

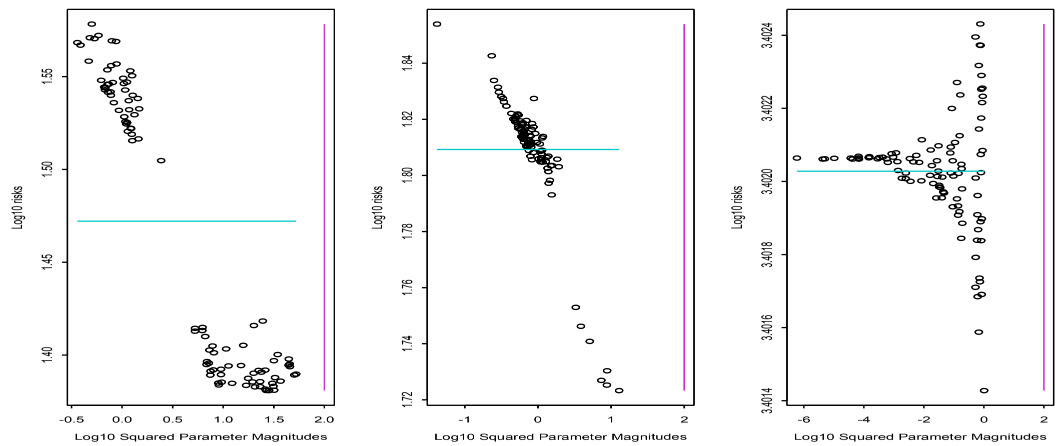


Fig. 2.5. The relation between shrinkage and risk is shown in terms of prediction performance (MSPE) and corresponding estimated parameter magnitude. The empirical Bayesian algorithm examined in Section 2.2.2 is used to train neural network models on synthetic data with a 'true' parameter magnitude at 100. The signal-to-noise ratio decreases from the left to the right at 100, 1, and 0.01. The horizontal lines in each plot are the mean values of prediction risks. The vertical line indicates the location of the 'true' parameter magnitude. The EB algorithm tends to take more liberties with shrinkage as the SNR decreases. When it does that, the prediction risk of the resulting model tends to return to the level of the LS estimator.

and denoted as $\sigma_\varepsilon^{\star 2}$, $\boldsymbol{F}^\star$, $\boldsymbol{G}^\star$, $\boldsymbol{\Lambda}^\star$, $\boldsymbol{A}^\star$ and $\boldsymbol{\Gamma}_0^\star$ respectively. Let $\boldsymbol{B}^\star = \boldsymbol{A}^\star \boldsymbol{\Gamma}^{\star -2}$ and $\boldsymbol{B}_0^\star = \boldsymbol{A}^\star \boldsymbol{\Gamma}_0^{\star -2}$, the equilibrium equation from (2.33) is

$$\boldsymbol{B}^\star = \boldsymbol{B}_0^\star [\boldsymbol{I}_q + \boldsymbol{B}^\star]^2 \ , \tag{2.34}$$

i.e.,

$$\boldsymbol{B}_0^\star \boldsymbol{B}^{\star 2} + (2\boldsymbol{B}_0^\star - 1)\boldsymbol{B}^\star + \boldsymbol{B}_0^\star = 0 \ .$$

Since all matrices in (2.34) are diagonal, the equilibrium equation is a system of q equations of the form

$$b_0^\star b^{\star 2} + (2b_0^\star - 1)b^\star + b_0^\star = 0 \ , \tag{2.35}$$

where $b_0^\star$ and $b^\star$ stand for any one of the diagonal elements $b_{0i}^\star$ and $b_i^\star$ $(i = 1, ..., q)$ respectively. Solving (2.35) for $b^\star$, one has

$$b^\star = \frac{(1 - 2b_0^\star) \pm \sqrt{(1 - 4b_0^\star)}}{2b_0^\star} \ . \tag{2.36}$$

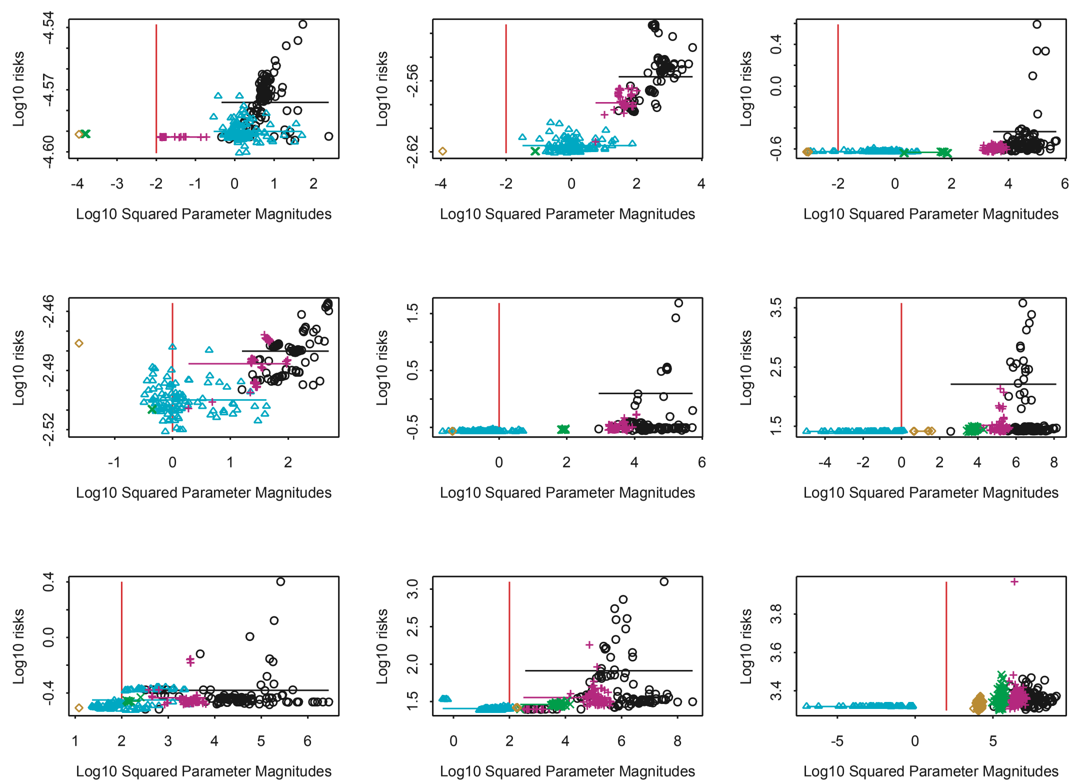


Fig. 2.6. The empirical Bayesian approach (EB) ('△') is compared with the LS estimator ('○') and the single-prior Bayes estimators with $k = 0.0001$ ('+'), $k = 0.01$ ('×') and $k = 1$ ('◇') on synthetic data set I in Appendix A, in terms of prediction performance (MSPE) and corresponding estimated parameter magnitude based on 100 runs per case. Each row represents a different 'true' parameter magnitude at the values of 0.01, 1 and 100 for a neural network model with $d = 3$ and $h = 3$. Each column represents a different signal-to-noise ratio (SNR) at the levels of 100, 1, 0.01 respectively. The 9 plots as a panel are used to show the relations between the prediction risk and the parameter magnitude for the 9 different data settings defined by the 'true' parameter magnitude and SNR. See Figure 2.7 for the corresponding boxplots of prediction risks.

For $b_0^\star > 1/4$, the iterative procedure in (2.34) diverges since the radicand $(1 - 4b_0^\star)$ is negative. $\gamma_i^\star = 0$ (i.e., θ_i has been shrunk to zero) for all i with $b_{0i}^\star > 1/4$, because $b_i^\star = \frac{\sigma_\varepsilon^{\star 2}}{\lambda_i^\star (\gamma_i^\star)^2}$. Though there might exist other $\gamma_i^\star$'s remaining at no-zero values when the iteration converges for the case of $0 < b_0^\star \leq 1/4$, the chance of divergence is overwhelmingly high. Since

$$\Pr(\gamma_i^\star = 0) = \Pr(b_{0i}^\star > 1/4) = \Pr(1/b_{0i}^\star < 4)$$

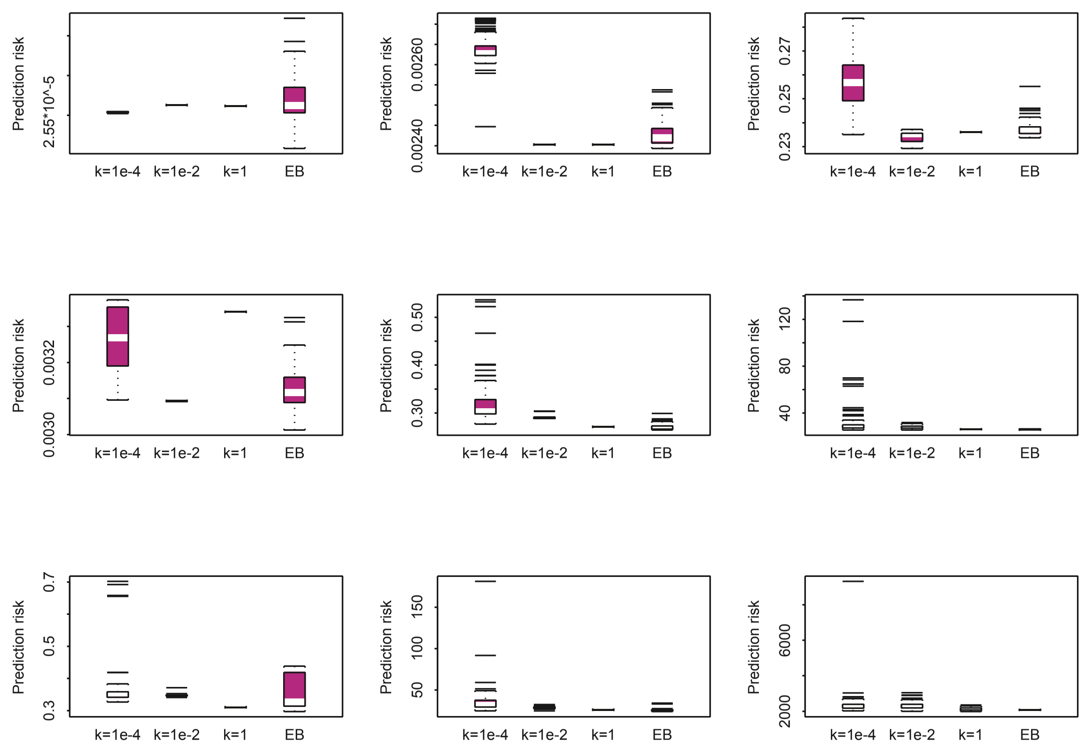


Fig. 2.7. The corresponding boxplots of prediction risks from Figure 2.6 (without the LS estimator).

and

$$\frac{1}{b_{0i}^{\star}} = \frac{\lambda_i^{\star}(\gamma_{0i}^{\star})^2}{\sigma_{\varepsilon}^{\star 2}} ,$$

one can define a null hypothesis $H_0 : \gamma_i^* = [\boldsymbol{G}^{\star -1}\boldsymbol{\theta}^*]_i = 0$, that reflects the prior believe of the parameters in (2.28). Referring to (2.14), under H_0,

$$\frac{1}{b_{0i}^{\star}} \sim F_{1,n-q-1} .$$

So the probability of $\gamma_i^{\star}$ being shrunk to zero becomes

$$\Pr(\gamma_i^{\star} = 0 | H_0) = \Pr(F_{1,n-q-1} < 4) ,$$

which increases with $(n - q - 1)$ (the difference between the sample size and the number of parameters). For instance, $\Pr(\gamma_i^{\star} = 0|H_0) = 0.953318$ for a 9-hidden-unit

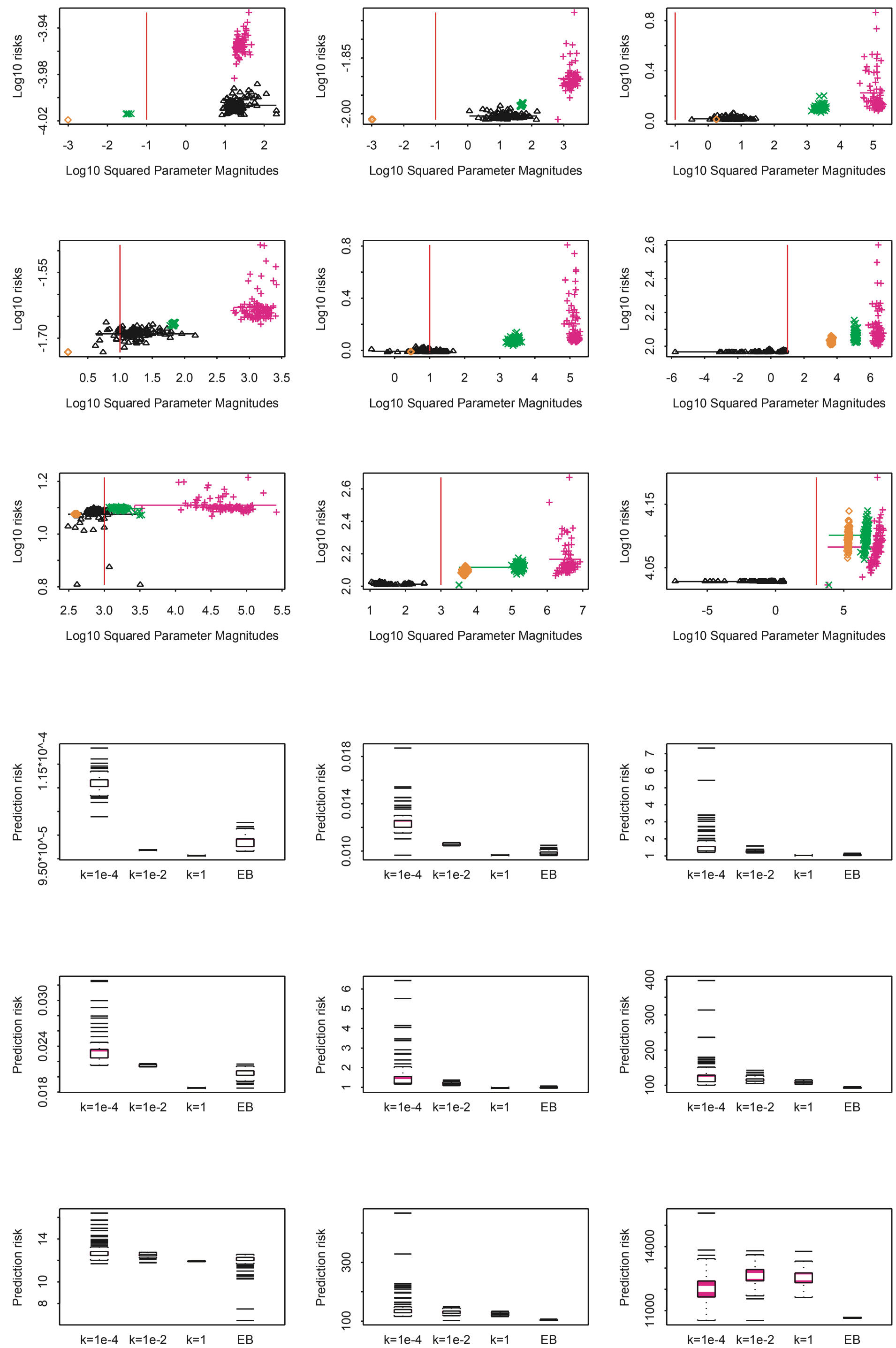


Fig. 2.8. A similar simulation on synthetic data set II with $d = 9$ and $h = 9$, where each row represents a different 'true' parameter magnitude at the values of 0.1, 10 and 1000 respectively. The EB method does better when SNR is low and the 'true' parameter magnitude is high, while the single-prior Bayes method with a right choice of the hyperparameter has a better performance with the opposite data situations.

neural network model of the ozone data (where $n = 330, q = h(d+1) + (h+1) = 9(9+1) + (9+1) = 100, (n-q-1) = 229, F_{1,229}^{1-0.953318} \approx 4$).

The EB algorithm discussed here can be improved by either formulation using higher-order asymptotic approximation or exact integral calculation by Markov Chain Monte Carlo (MCMC) techniques. But both improvements come with higher computational cost, especially the latter. For example, a MCMC implementation of a very small example with 2 response variables and 2 predictors needs about 20 hours of computation time [39], while a run of single-prior Bayes or EB estimation lasts only seconds. Furthermode, the EB method is not generically Bayesian robust, since the prior has tails that are of the same form as the likelihood function and hence they will work only when the likelihood function is concentrated in the central portion of the prior. Although the ML-II empirical Bayesian approach strove to be a default algorithm for the model in (1.2), some intrinsic flaws in approximating and evaluating the prior assumption prevent it from reaching the optimality it aimed for.

2.2.3 Hierarchical Bayes

Rather than specifying the prior as a single function, the hyperparameters of the prior distribution used in hierarchical Bayes methods are further modeled by other hyperprior distributions. For instance of the model in (1.2),

$$y_t = \sum_{k=1}^{h} \beta_k g(\boldsymbol{x}_t' \boldsymbol{\alpha}_k) + \varepsilon_t \ , \ \ t = 1, ..., n \ , \tag{2.37}$$

the standard conjugate normal hierarchy assigns the following set of priors and hyperpriors to the parameters $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta})'$:

$$\begin{aligned}
\pi(\beta_k|\mu_\beta, \sigma_\beta^2) \ &\sim \ \mathcal{N}(\mu_\beta, \sigma_\beta^2) \ , k = 1, ..., h. \\
\pi(\boldsymbol{\alpha}_k|\boldsymbol{\mu}_\alpha, \boldsymbol{\Sigma}_\alpha) \ &\sim \ \mathcal{N}_d(\boldsymbol{\mu}_\alpha, \boldsymbol{\Sigma}_\alpha) \ , k = 1, ..., h. \\
\pi(\mu_\beta) \sim \mathcal{N}(a_1, a_2) \ , \qquad & \pi(\sigma_\beta^{-1}) \sim Gamma(a_3, a_4) \\
\pi(\boldsymbol{\mu}_\alpha) \sim \mathcal{N}_d(\boldsymbol{b}_1, \boldsymbol{B}_2) \ , \qquad & \pi(\boldsymbol{\Sigma}_\alpha^{-2}) \sim Wishart(\boldsymbol{b}_3, \boldsymbol{B}_4) \\
\pi(\sigma_\varepsilon^{-2}) \ &\sim \ Gamma(c_1, c_2) \ ,
\end{aligned} \tag{2.38}$$

where all the a's, b's and c's are assumed known. If $\boldsymbol{\xi}$ denotes all the hyperparameters, $\boldsymbol{\xi} = (\mu_\beta, \sigma_\beta, \boldsymbol{\mu}_\alpha, \boldsymbol{\Sigma}_\alpha, \sigma_\varepsilon)'$, the hierarchy can be summarized in three levels:

$$\begin{aligned} D_n|\boldsymbol{\theta} &\sim p(D_n|\boldsymbol{\theta}) \ , \text{ likelihood} \\ \boldsymbol{\Theta}|\boldsymbol{\xi} &\sim \pi(\boldsymbol{\theta}|\boldsymbol{\xi}) \ , \text{ prior} \\ \boldsymbol{\Xi} &\sim \pi(\boldsymbol{\xi}) \ , \text{ hyperprior} \ . \end{aligned} \tag{2.39}$$

The Bayesian inference is then based on the marginal posterior distribution of the parameter vector

$$\pi(\boldsymbol{\theta}|D_n) = \int \pi(\boldsymbol{\theta}, \boldsymbol{\xi}|D_n) d\boldsymbol{\xi} \ , \tag{2.40}$$

where the posterior distribution of all parameters

$$\pi(\boldsymbol{\theta}, \boldsymbol{\xi}|D_n) = \frac{p(D_n|\boldsymbol{\theta}, \boldsymbol{\xi})\pi(\boldsymbol{\theta}, \boldsymbol{\xi})}{\int p(D_n|\boldsymbol{\theta}, \boldsymbol{\xi})\pi(\boldsymbol{\theta}, \boldsymbol{\xi}) d\boldsymbol{\theta} d\boldsymbol{\xi}} \ , \tag{2.41}$$

with $p(D_n|\boldsymbol{\theta}, \boldsymbol{\xi}) \propto \exp(\frac{-1}{2\sigma_\varepsilon^2} \sum_{t=1}^n (y_t - f(\boldsymbol{x}_t; \boldsymbol{\theta}))^2)$ the likelihood function and $\pi(\boldsymbol{\theta}, \boldsymbol{\xi}) = \pi(\boldsymbol{\theta}|\boldsymbol{\xi})\pi(\boldsymbol{\xi})$ the prior. The final outcome of the inference is based on the predictive distribution

$$p(y_{n+1}|D_n, \boldsymbol{x}_{n+1}) = \int p(y_{n+1}|\boldsymbol{x}_{n+1}; \boldsymbol{\theta}, \boldsymbol{\xi})\pi(\boldsymbol{\theta}, \boldsymbol{\xi}|D_n) d\boldsymbol{\theta} d\boldsymbol{\xi} \ , \tag{2.42}$$

where $p(y|\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\xi})$ designates the conditional distribution of response variable y given $\boldsymbol{x}$ and the parameters, which is $\mathcal{N}(f(\boldsymbol{x}; \boldsymbol{\theta}), \sigma_\varepsilon^2)$ in a normal-theory setting. The resulting estimates of the parameter vector and the response surface are typically the posterior mean and associated confidence regions (estimated from the posterior variance-covariance) obtained from (2.40) and (2.42). For example, the hierarchical Bayes estimator of $\boldsymbol{\theta}$ is

$$\begin{aligned} \hat{\boldsymbol{\theta}} = E(\boldsymbol{\Theta}|D_n) &= \int \boldsymbol{\theta}\pi(\boldsymbol{\theta}|D_n) d\boldsymbol{\theta} \\ &= \iint \boldsymbol{\theta}\pi(\boldsymbol{\theta}, \boldsymbol{\xi}|D_n) d\boldsymbol{\theta} d\boldsymbol{\xi} \\ &= \iint \boldsymbol{\theta}\pi(\boldsymbol{\theta}|D_n, \boldsymbol{\xi}) d\boldsymbol{\theta} \pi(\boldsymbol{\xi}|D_n) d\boldsymbol{\xi} \\ &= E^{\pi(\boldsymbol{\xi}|D_n)}[E(\boldsymbol{\Theta}|D_n, \boldsymbol{\xi})] \ , \end{aligned} \tag{2.43}$$

which is the expectation of a single-prior Bayes estimator over the hyperprior density $\pi(\boldsymbol{\xi}|D_n)$, and can be considered as a limit of simpler estimators. As we shall see in Chapter 3, (2.43) is usually not in a closed form, but numerical calculation is relatively simple. Another advantage of this approach is that it is usually Bayesian robust with desirable classical frequentist risk properties, since one can obtain prior distributions with flatter tails through certain hyperpriors (cf. Chapter 3). Therefore, it is often the case that the hierarchical Bayesian methodology serves as an effective way to construct estimators appealing to both Bayesians and frequentists. For the model in (2.37), however, only a MCMC implementation is tried on some very small examples [40] with very high computational cost, and there is no easy-to-use default algorithm developed so far.

3. Approach Based On Robust Bayesian Steinization

3.1 A Robust Bayes and Asymptotically Minimax Estimator

When mathematical statisticians develop an estimator, the Bayes robustness over the class of prior being used might not be the apparent objective, but rather the minimaxity of the estimator serves as the organizing theme [51]. However, the hierarchical Bayesian methodology often plays an operational role in constructing reasonable estimators with classical frequentist risk properties such as minimaxity or near minimaxity. Adopting a research line presented in [52, 53, 54, 24] addressing the minimax estimation of a multivariate normal mean, we develop a robust Bayes and asymptotically minimax estimator in Newton-Raphson form and its resulting confidence intervals for the regression model in (1.2).

3.1.1 The hierarchy

Consider the following hierarchy:

$$D_n|\boldsymbol{\theta} \sim p(D_n|\boldsymbol{\theta}) \propto \exp\{-\frac{1}{2\sigma_\varepsilon^2}\sum_{t=1}^{n}[y - f(\boldsymbol{x}_t;\boldsymbol{\theta})]^2\} \text{ , likelihood} \tag{3.1}$$

$$\pi(\boldsymbol{\theta}|\xi) \sim \mathcal{N}_q(\boldsymbol{\mu}, \boldsymbol{B}(\xi)) \text{ , prior} \tag{3.2}$$

$$\pi(\xi) \sim \frac{1}{2}\xi^{-1/2} \text{ on } (0,1) \text{ , hyperprior .} \tag{3.3}$$

The likelihood level can be replaced by the maximum likelihood estimator itself

$$p(\boldsymbol{\delta}_0|\boldsymbol{\theta}) \sim \mathcal{N}_q(\boldsymbol{\theta}, \boldsymbol{\Sigma}) \text{ , the estimator that shall be improved upon ,} \tag{3.4}$$

which is asymptotically an unbiased estimator with a variance-covariance matrix

$$\boldsymbol{\Sigma} = \sigma_\varepsilon^2(\boldsymbol{F}'\boldsymbol{F})^{-1} \text{ ,} \tag{3.5}$$

In Newton-Raphson formulation, the iterative ML estimate is in the form

$$\hat{\boldsymbol{\theta}}_{\tau+1} = \hat{\boldsymbol{\theta}}_\tau + (\hat{\boldsymbol{F}}_\tau'\hat{\boldsymbol{F}}_\tau)^{-1}\hat{\boldsymbol{F}}_\tau'\boldsymbol{\varepsilon}_\tau \text{ ,} \tag{3.6}$$

where $\hat{\boldsymbol{F}}_\tau = \boldsymbol{F}(\hat{\boldsymbol{\theta}}_\tau)$, $\hat{\boldsymbol{\varepsilon}}_\tau = \boldsymbol{y} - \boldsymbol{f}(\hat{\boldsymbol{\theta}}_\tau)$. Therefore, for the calculation of $\hat{\boldsymbol{\theta}}_{\tau+1}$, $\hat{\boldsymbol{F}}_\tau$'s and hence $\boldsymbol{\Sigma}_\tau = \hat{\sigma}^2_{\varepsilon\tau}(\hat{\boldsymbol{F}}'_\tau\hat{\boldsymbol{F}}_\tau)^{-1}$ are known. In the asymptotic sense ($n \to \infty$), especially when $\hat{\boldsymbol{\theta}}_\tau$ is close to converge, the iterative ML estimate can be approximately summarized as $\mathcal{N}_q(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ with a constant variance-covariance matrix $\boldsymbol{\Sigma}$.

3.1.2 The prior

In the prior level, let

$$\boldsymbol{B}(\xi) = \rho\xi^{-1}(\boldsymbol{\Sigma} + \boldsymbol{A}) - \boldsymbol{\Sigma} \ ,$$

where $\rho = \frac{q+1}{q+3}$ and $\boldsymbol{A}$ is the variance-covariance matrix reflecting the accuracy of one's prior belief in $\boldsymbol{\mu}$. The free parameters, $\boldsymbol{\mu}$ and $\boldsymbol{A}$, in the prior, are devised for two purposes. First, this makes the new estimator ready to incorporate very simple prior inputs that are summarized in the first and second moments. Second, this allows the practitioners to easily locate certain parameter region over which the risk behavior is to be improved. For instance, $\boldsymbol{\mu}$ and $\boldsymbol{A}$ specify an ellipsoid

$$\{\boldsymbol{\theta} : (\boldsymbol{\theta} - \boldsymbol{\mu})'\boldsymbol{A}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu}) \leq q - 0.6\} \ ,$$

which has probability of $\frac{1}{2}$ for containing $\boldsymbol{\theta}$. For the sake of making impartial comparison with other methodologies in Chapter 2, $\boldsymbol{\mu} = \boldsymbol{0}$ will be assumed in this chapter, while $\boldsymbol{\mu}$ is not difficult to be included in the subsequent calculations.

To robustify the prior hierarchy, a resulting prior around $\boldsymbol{\mu}$ with flat tail is desired. By assuming the hierarchical prior defined in (3.3), the prior density of the parameter vector $\boldsymbol{\theta}$ is

$$\begin{aligned} \pi(\boldsymbol{\theta}) &= \int \pi(\boldsymbol{\theta}|\xi)\pi(\xi)d\xi \\ &= \int_0^1 [\det \boldsymbol{B}(\xi)]^{-1/2} \exp\{-\boldsymbol{\theta}'\boldsymbol{B}(\xi)^{-1}\boldsymbol{\theta}/2\}\xi^{-1/2}/2d\xi \ . \end{aligned} \qquad (3.7)$$

For large $||\boldsymbol{\theta}||^2$, $\pi(\boldsymbol{\theta}) \propto \{\boldsymbol{\theta}'(\boldsymbol{\Sigma} + \boldsymbol{A})^{-1}\boldsymbol{\theta}\}^{-(q+1)/2}$ that indicates a flatter (polynomial) tail, comparing with the exponentially decreasing tails of a normal density as in the case of the likelihood function. It also can be shown that $\pi(\boldsymbol{\theta})$ has finite mass, and is proper when $\lambda_{\max}(\boldsymbol{A}^{-1}\boldsymbol{\Sigma}) \leq (q+1)/2$.

3.1.3 The abstract version

With the prior in (3.7), the abstract form of the new estimator is the posterior mean of $\pi(\boldsymbol{\theta}|D_n)$, i.e.,

$$\hat{\boldsymbol{\theta}} = \frac{\int \boldsymbol{\theta} \exp\{-(\boldsymbol{\delta}_0 - \boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\delta}_0 - \boldsymbol{\theta})/2\}\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int \exp\{-(\boldsymbol{\delta}_0 - \boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\delta}_0 - \boldsymbol{\theta})/2\}\pi(\boldsymbol{\theta})d\boldsymbol{\theta}} \ . \tag{3.8}$$

The derivation of the abstract version in this subsection follows the results in [54, 24], and is included for the sake of completeness. Since $\pi(\boldsymbol{\theta})$ is finite in any compacta of zero and bounded outside the compacta, it is allowed to interchange the order of the integration in (3.8), and its numerator becomes

$$\begin{aligned} & \int \boldsymbol{\theta} \exp\{-(\boldsymbol{\delta}_0 - \boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\delta}_0 - \boldsymbol{\theta})/2\}\pi(\boldsymbol{\theta})d\boldsymbol{\theta} \\ = & \int_0^1 \int \boldsymbol{\theta} \exp\{-[(\boldsymbol{\delta}_0 - \boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\delta}_0 - \boldsymbol{\theta}) + \boldsymbol{\theta}'\boldsymbol{B}(\xi)^{-1}\boldsymbol{\theta}]/2\}d\boldsymbol{\theta} \\ & \times[\det \boldsymbol{B}(\xi)]^{-1/2}\xi^{-1/2}/2d\xi \ . \end{aligned} \tag{3.9}$$

By completing squares and integrating out over $\boldsymbol{\theta}$, the numerator is equal to

$$\begin{aligned} & \int_0^1 \int \boldsymbol{\theta} \exp\{-[\boldsymbol{\theta} - (\boldsymbol{\Sigma}^{-1} + \boldsymbol{B}(\xi)^{-1})^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}_0]'(\boldsymbol{\Sigma}^{-1} + \boldsymbol{B}(\xi)^{-1}) \\ & \times[\boldsymbol{\theta} - (\boldsymbol{\Sigma}^{-1} + \boldsymbol{B}(\xi)^{-1})^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}_0]/2\}d\boldsymbol{\theta} \\ & \times \exp\{-[\boldsymbol{\delta}_0'\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}_0 - \boldsymbol{\delta}_0'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma}^{-1} + \boldsymbol{B}(\xi)^{-1})^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}_0]/2\} \\ & \times[\det \boldsymbol{B}(\xi)]^{-1/2}\xi^{-1/2}/2d\xi \\ = & \int_0^1 (\boldsymbol{\Sigma}^{-1} + \boldsymbol{B}(\xi)^{-1})^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}_0 \\ & \times \exp\{-[\boldsymbol{\delta}_0'\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}_0 - \boldsymbol{\delta}_0'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma}^{-1} + \boldsymbol{B}(\xi)^{-1})^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}_0]/2\} \\ & \times[\det(\boldsymbol{\Sigma}^{-1} + \boldsymbol{B}(\xi)^{-1})]^{-1/2}[\det \boldsymbol{B}(\xi)]^{-1/2}\xi^{-1/2}/2d\xi \ . \end{aligned} \tag{3.10}$$

Since

$$\begin{aligned} (\boldsymbol{\Sigma}^{-1} + \boldsymbol{B}(\xi)^{-1})^{-1} &= \boldsymbol{\Sigma} - \boldsymbol{\Sigma}[\boldsymbol{\Sigma} + \boldsymbol{B}(\xi)]^{-1}\boldsymbol{\Sigma} \\ &= \boldsymbol{\Sigma} - \frac{1}{\rho}\xi\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \boldsymbol{A})^{-1}\boldsymbol{\Sigma} \ , \end{aligned} \tag{3.11}$$

and

$$(\boldsymbol{\Sigma}^{-1} + \boldsymbol{B}(\xi)^{-1})\boldsymbol{B}(\xi) = \boldsymbol{\Sigma}^{-1}\boldsymbol{B}(\xi) + \boldsymbol{I}_q = \rho\xi^{-1}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma} + \boldsymbol{A}) \ , \tag{3.12}$$

the numerator is simplified to

$$\int_0^1 [\boldsymbol{I}_q - \frac{\xi}{\rho}\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\boldsymbol{A})^{-1}]\boldsymbol{\delta}_0 \exp\{-\frac{1}{2}(\frac{\xi}{\rho})||\boldsymbol{\delta}_0||^2\} \times[\det(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma}+\boldsymbol{A}))]^{-1/2}\frac{1}{2}(\frac{\xi}{\rho})^{\frac{q-1}{2}}\rho^{1/2}d(\frac{\xi}{\rho}) , \tag{3.13}$$

where $||\boldsymbol{\delta}_0||^2 = \boldsymbol{\delta}_0'(\boldsymbol{\Sigma}+\boldsymbol{A})^{-1}\boldsymbol{\delta}_0$, and similarly the denominator is in the form

$$\int_0^1 \exp\{-\frac{1}{2}(\frac{\xi}{\rho})||\boldsymbol{\delta}_0||^2\}[\det(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma}+\boldsymbol{A}))]^{-1/2}\frac{1}{2}(\frac{\xi}{\rho})^{\frac{q-1}{2}}\rho^{1/2}d(\frac{\xi}{\rho}) . \tag{3.14}$$

Denote $\zeta = \xi/\rho$,

$$\begin{aligned} r_q(v) &= \frac{\int_0^1 \zeta^{\frac{q+1}{2}} \exp\{-\zeta v/2\}d\zeta}{\int_0^1 \zeta^{\frac{q-1}{2}} \exp\{-\zeta v/2\}d\zeta} \\ &= \frac{q+1}{v}\{1-[\frac{q+1}{2}\int_0^1 \zeta^{\frac{q-1}{2}}\exp\{-(\zeta-1)v/2\}d\zeta]^{-1}\} \\ &= \frac{q+1}{v}[1-h_q(v)] , \end{aligned} \tag{3.15}$$

and

$$\begin{aligned} h_q(v) &= [\frac{q+1}{2}\int_0^1 \zeta^{\frac{q-1}{2}}\exp\{-(\zeta-1)v/2\}d\zeta]^{-1} \\ &= \left[\sum_{i=0}^{\infty}\frac{\Gamma(\frac{q+3}{2})(v/2)^i}{\Gamma(\frac{q+3+2i}{2})}\right]^{-1} \\ &\approx \frac{1-\frac{v}{q+1}}{1-(\frac{v}{q+1})^{\sqrt{2(q+7)/\pi}}} , \end{aligned} \tag{3.16}$$

then the abstract version of the new estimator is

$$\hat{\boldsymbol{\theta}} = [\boldsymbol{I}_q - r_q(||\boldsymbol{\delta}_0||^2)\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\boldsymbol{A})^{-1}]\boldsymbol{\delta}_0 , \tag{3.17}$$

which is in the form of the well-known minimax James-Stein estimator [25].

If $||\boldsymbol{\delta}_0||^2 \to 0$ as we guessed in the prior $\mathcal{N}_q(\mathbf{0}, \boldsymbol{A})$, then $r_q(||\boldsymbol{\delta}_0||^2) \to 1$, and the new estimator behaves like Bayesian estimators in Chapter 2. When $||\boldsymbol{\delta}_0||^2 \to \infty$, $r_q(||\boldsymbol{\delta}_0||^2) \to (q+1)/||\boldsymbol{\delta}_0||^2$ so that the new estimator is close to the original ML estimator. Hence it is a rather good approximation that $r_q(||\boldsymbol{\delta}_0||^2) \approx \min\{1, (q+1)/||\boldsymbol{\delta}_0||^2\}$.

It has been shown for the canonical case in [24] that under the generalized quadratic loss $L(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})' \boldsymbol{Q} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$, if $q \geq 5$ and $(q+5)\lambda_{max}\{\boldsymbol{\Sigma Q \Sigma}(\boldsymbol{\Sigma} + \boldsymbol{A})^{-1}\} \leq 2\mathrm{tr}\{\boldsymbol{\Sigma Q \Sigma}(\boldsymbol{\Sigma} + \boldsymbol{A})^{-1}\}$, then $R(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) < R(\boldsymbol{\delta}_0, \boldsymbol{\theta}) = \mathrm{tr}(\boldsymbol{Q\Sigma}), \forall \boldsymbol{\theta}$. The condition is hold if $\boldsymbol{\Sigma Q \Sigma}(\boldsymbol{\Sigma} + \boldsymbol{A})^{-1}$ is a multiple of the identity matrix. Under the first order approximation of the quadratic loss $||\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{\theta})||^2$, $\boldsymbol{Q} \approx \boldsymbol{F}'\boldsymbol{F}$ so that $\boldsymbol{\Sigma Q \Sigma}(\boldsymbol{\Sigma} + \boldsymbol{A})^{-1}$ is close to the identity matrix if $\boldsymbol{A}$ is relatively insignificant to $\boldsymbol{\Sigma}$. For the neural network regression model and the Bayesian methods under our consideration, a diagonal $\boldsymbol{A}$ is usually assumed, and $\boldsymbol{\Sigma} = \sigma_\varepsilon^2(\boldsymbol{F}'\boldsymbol{F})^{-1}$. One can verify that if one canonicalizes $\boldsymbol{F}'\boldsymbol{F}$, $\boldsymbol{\Sigma Q \Sigma}(\boldsymbol{\Sigma} + \boldsymbol{A})^{-1}$ is approximately a multiple of the identity. In general, these properties enable the new estimator to elude the potential unboundedness of the ridge procedure shown in Figure 1.6 and Section 2.2.1, while keeping other desirable aspects of a Bayesian method (see further simulation results in Section 3.2).

3.1.4 The Newton-Raphson iterative version

We utilize the abstract version of the new estimator in (3.17) with the one-step approximation in (2.8). By noticing (3.11), (3.12) and

$$\boldsymbol{\Sigma} - r_{q0}\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \boldsymbol{A})^{-1}\boldsymbol{\Sigma} = \{\boldsymbol{\Sigma}^{-1} + r_{q0}\,[\boldsymbol{A} + (1 - r_{q0})\boldsymbol{\Sigma}]^{-1}\}^{-1} ,$$

the abstract version can be rearranged as

$$\begin{aligned}
\hat{\boldsymbol{\theta}} &= \{\boldsymbol{\Sigma} - r_{q0}\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \boldsymbol{A})^{-1}\boldsymbol{\Sigma}\}\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}_0 \\
&= \{\boldsymbol{\Sigma}^{-1} + r_{q0}\,[\boldsymbol{A} + (1 - r_{q0})\boldsymbol{\Sigma}]^{-1}\}^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}_0 \\
&= \{\boldsymbol{\Sigma}^{-1} + \left[\frac{1}{r_{q0}}\boldsymbol{A} + \frac{1 - r_{q0}}{r_{q0}}\boldsymbol{\Sigma}\right]^{-1}\}^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}_0 \\
&= \{\boldsymbol{\Sigma}^{-1} + \frac{r_{q0}}{1 - r_{q0}}\boldsymbol{\Sigma}^{-1} - (\frac{r_{q0}}{1 - r_{q0}})^2\boldsymbol{\Sigma}^{-1}\left[\frac{r_{q0}}{1 - r_{q0}}\boldsymbol{\Sigma}^{-1} + r_{q0}\boldsymbol{A}^{-1}\right]^{-1}\boldsymbol{\Sigma}^{-1}\}^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}_0 \\
&= \{\frac{1}{1 - r_{q0}}\boldsymbol{I}_q - (\frac{r_{q0}}{1 - r_{q0}})^2\left[\frac{r_{q0}}{1 - r_{q0}}\boldsymbol{\Sigma}^{-1} + r_{q0}\boldsymbol{A}^{-1}\right]^{-1}\boldsymbol{\Sigma}^{-1}\}^{-1}\boldsymbol{\delta}_0 \\
&= \{\frac{r_{q0}}{(1 - r_{q0})^2}\boldsymbol{\Sigma}^{-1} + \frac{r_{q0}}{1 - r_{q0}}\boldsymbol{A}^{-1} - (\frac{r_{q0}}{1 - r_{q0}})^2\boldsymbol{\Sigma}^{-1}\}^{-1}(\frac{r_{q0}}{1 - r_{q0}}\boldsymbol{\Sigma}^{-1} + r_{q0}\boldsymbol{A}^{-1})\boldsymbol{\delta}_0 \\
&= (\boldsymbol{\Sigma}^{-1} + \boldsymbol{A}^{-1})^{-1}[\boldsymbol{\Sigma}^{-1} + (1 - r_{q0})\boldsymbol{A}^{-1}]\boldsymbol{\delta}_0 ,
\end{aligned}$$

where $r_{q0} = r_q(||\boldsymbol{\delta}_0||^2)$. Since the one-step version of the ML estimator $\boldsymbol{\delta}_0$ is

$$\boldsymbol{\delta}_0 = \boldsymbol{\theta} + (\boldsymbol{F}'\boldsymbol{F})^{-1}\boldsymbol{F}\boldsymbol{\varepsilon} \ , \tag{3.18}$$

and if the variance-covariance matrix $\boldsymbol{A} = \sigma_\varepsilon^2 \boldsymbol{K}^{-1}$ with $\boldsymbol{K} = \mathrm{diag}(k_i)$ and $r_{q0} = r_q(||\boldsymbol{\delta}_0||^2)$ is approximated by $r_q = r_q(||\boldsymbol{\theta}||^2)$ with $||\boldsymbol{\theta}||^2 = \boldsymbol{\theta}'(\boldsymbol{\Sigma} + \boldsymbol{A})^{-1}\boldsymbol{\theta}$, then the one-step approximation for the new estimator $\hat{\boldsymbol{\theta}}$ is

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= (\boldsymbol{F}'\boldsymbol{F} + \boldsymbol{K})^{-1}[\boldsymbol{F}'\boldsymbol{F} + (1 - r_q)\boldsymbol{K}][\boldsymbol{\theta} + (\boldsymbol{F}'\boldsymbol{F})^{-1}\boldsymbol{F}\boldsymbol{\varepsilon}] \\ &= \boldsymbol{\theta} + (\boldsymbol{F}'\boldsymbol{F} + \boldsymbol{K})^{-1}\{[\boldsymbol{I}_q + (1 - r_q)\boldsymbol{K}(\boldsymbol{F}'\boldsymbol{F})^{-1}]\boldsymbol{F}'\boldsymbol{\varepsilon} - r_q\boldsymbol{K}\boldsymbol{\theta}\} \ . \end{aligned} \tag{3.19}$$

It can be easily seen that

$$-(\boldsymbol{F}'\boldsymbol{F} + \boldsymbol{K})$$

is approximately the derivative matrix of the vector

$$\{[\boldsymbol{I}_q + (1 - r_q)\boldsymbol{K}(\boldsymbol{F}'\boldsymbol{F})^{-1}]\boldsymbol{F}'\boldsymbol{\varepsilon} - r_q\boldsymbol{K}\boldsymbol{\theta}\}$$

with respect to $\boldsymbol{\theta}$, so that equation (3.19) is a Newton-Raphson procedure. It is of an immediate interest to figure out the corresponding objective function of this optimization procedure. The new objective function is approximately in the form

$$||\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{\theta})||^2 + (1 - r_q)(\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{\theta}))'\boldsymbol{F}'\boldsymbol{K}(\boldsymbol{F}'\boldsymbol{F})^{-1}\boldsymbol{F}(\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{\theta})) + r_q\boldsymbol{\theta}'\boldsymbol{K}\boldsymbol{\theta} \ . \tag{3.20}$$

The prior hierarchy from (3.3) alters the quadratic loss function in a more peculiar way than a single smoothing (or penalty) term. The counterpart of the second term in optimization theory and statistics needs to be investigated further.

Finally, based on the one-step approximation of the new estimator in (3.19), the practical iterative procedure can be written as

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{\tau+1} &= \hat{\boldsymbol{\theta}}_\tau + (\hat{\boldsymbol{F}}_\tau'\hat{\boldsymbol{F}}_\tau + \hat{\boldsymbol{K}}_\tau)^{-1} \\ &\quad \{[\boldsymbol{I}_q + (1 - \hat{r}_{q\tau})\hat{\boldsymbol{K}}_\tau(\hat{\boldsymbol{F}}_\tau'\hat{\boldsymbol{F}}_\tau)^{-1}]\hat{\boldsymbol{F}}_\tau\hat{\boldsymbol{\varepsilon}}_\tau - \hat{r}_{q\tau}\hat{\boldsymbol{K}}_\tau\hat{\boldsymbol{\theta}}_\tau\} \ . \end{aligned} \tag{3.21}$$

3.1.5 Confidence intervals

Following a similar procedure in Section 3.1.3, the posterior covariance matrix for $\pi(\boldsymbol{\theta}|D_n)$ in the abstract form is

$$\begin{aligned} \boldsymbol{C}(\boldsymbol{\delta}_0) &= \boldsymbol{\Sigma} - r_q(||\boldsymbol{\delta}_0||^2)\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\boldsymbol{A})^{-1}\boldsymbol{\Sigma} + \\ & \quad w_q(||\boldsymbol{\delta}_0||^2)\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\boldsymbol{A})^{-1}\boldsymbol{\delta}_0\boldsymbol{\delta}_0'(\boldsymbol{\Sigma}+\boldsymbol{A})^{-1}\boldsymbol{\Sigma}\ , \end{aligned} \tag{3.22}$$

where

$$w_q(||\boldsymbol{\delta}_0||^2) = \frac{2(q+1)}{||\boldsymbol{\delta}_0||^4}\left[1+\{\frac{||\boldsymbol{\delta}_0||^2}{2\rho}[\rho r_q(||\boldsymbol{\delta}_0||^2)-1]-1\}h_q(||\boldsymbol{\delta}_0||^2)\right] .$$

An approximate $100(1-\alpha)\%$ prediction confidence interval for y_t is

$$\begin{aligned} &\hat{y}_t \pm t_{n-q}^{\alpha/2} s \Big[(1-\hat{r}_q) + \hat{r}_q \hat{\boldsymbol{f}}_t'(\hat{\boldsymbol{F}}_\tau'\hat{\boldsymbol{F}}_\tau + \hat{\boldsymbol{K}}_\tau)^{-1}\hat{\boldsymbol{f}}_t + \\ &\hat{w}_q\{\hat{\boldsymbol{f}}_t'[\boldsymbol{I}_q - (\hat{\boldsymbol{F}}_\tau'\hat{\boldsymbol{F}}_\tau + \hat{\boldsymbol{K}}_\tau)^{-1}\hat{\boldsymbol{F}}_\tau'\hat{\boldsymbol{F}}_\tau]\hat{\boldsymbol{\theta}}\}^2\Big]^{1/2} , \end{aligned} \tag{3.23}$$

where all the $\hat{\cdot}$'s are evaluated at $\hat{\boldsymbol{\theta}}$. Though a similar performance gain as the estimator is expected for the confidence intervals, a full scale investigation would take up the space of another chapter. Therefore, we shall only outline a plan of future study on this topic in the final chapter.

3.2 Numerical Experiments

The robust Bayes (RB) method opens the door to a wide variety of possible implementations that shall carry the desirable characteristics from various estimators under consideration. For the purpose of developing a default (or 'standard') single-run training method for the neural network regression model, we propose the following scenario.

Like all other numerical implementations of nonlinear model in general, the standard quasi-Newton or conjugate gradient methods are recommended as the basic routine to ensure numerical stability. A reasonable way to evaluate the hyperparameter k is to do a combination of random search as in the empirical Bayes (EB) and fixation at a single value as in the single-prior Bayes (SPB). Firstly, a reasonable

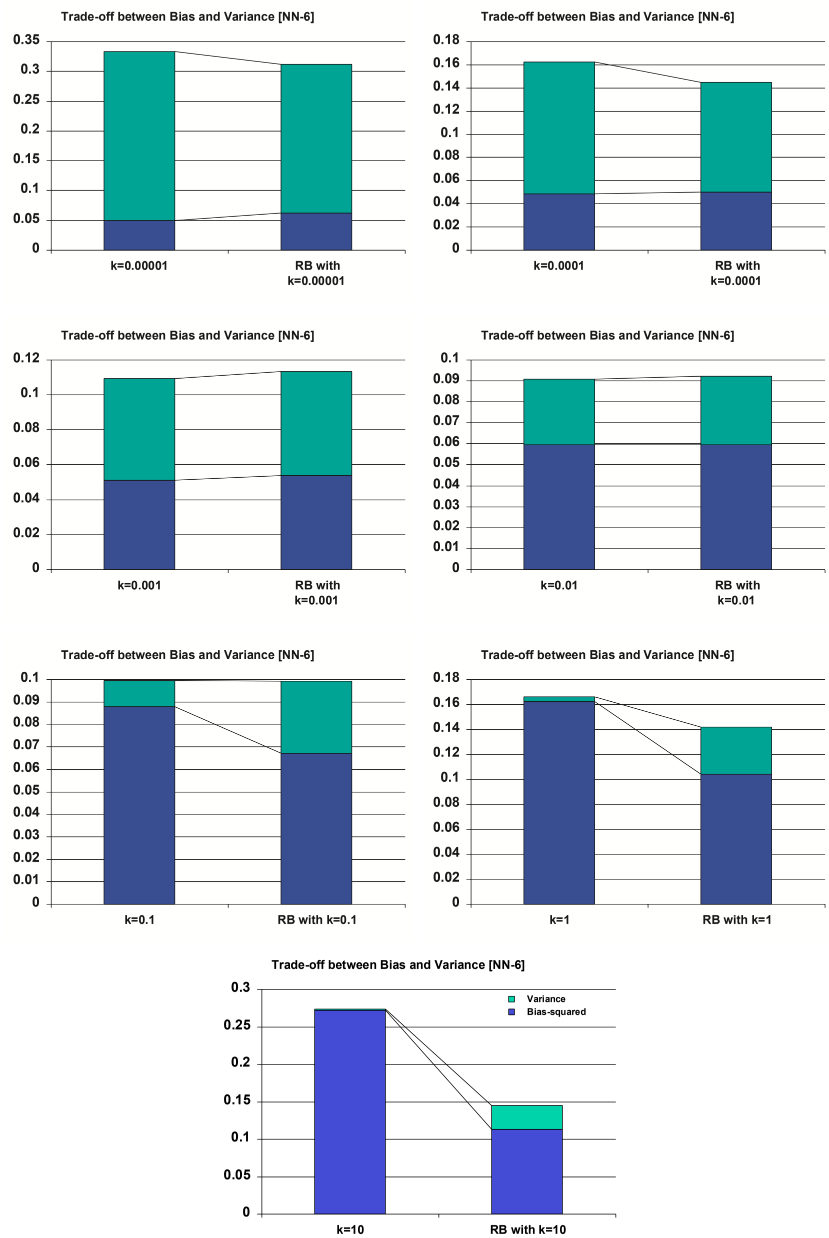


Fig. 3.1. The robust Bayesian approach (RB) is compared with the single-prior Bayes (SPB) method used in neural network regression models with six hidden units on ozone data, in terms of prediction performance (MSPE) and its bias-variance decomposition. The RB method can be seen as being capable to perform multiple compromises with the characteristics from several methods involved. When the guessed k is too small ($k = 0.00001, 0.0001$), RB has a lower MSPE by showing the character of EB with higher bias but lower variance. When the guessed k is about right ($k = 0.001, 0.01, 0.1$), RB has virtually the same MSPE as SPB, with the bias-variance trade-off bearing less character of ML and more character of EB as k increases. When the guessed k is too large ($k = 1, 10$), RB again has a lower MSPE with a much lower bias (the character of ML) and a total MSPE upper bounded by the MSPE of EB, so that the MSPE is not left unbounded as k continues to increase.

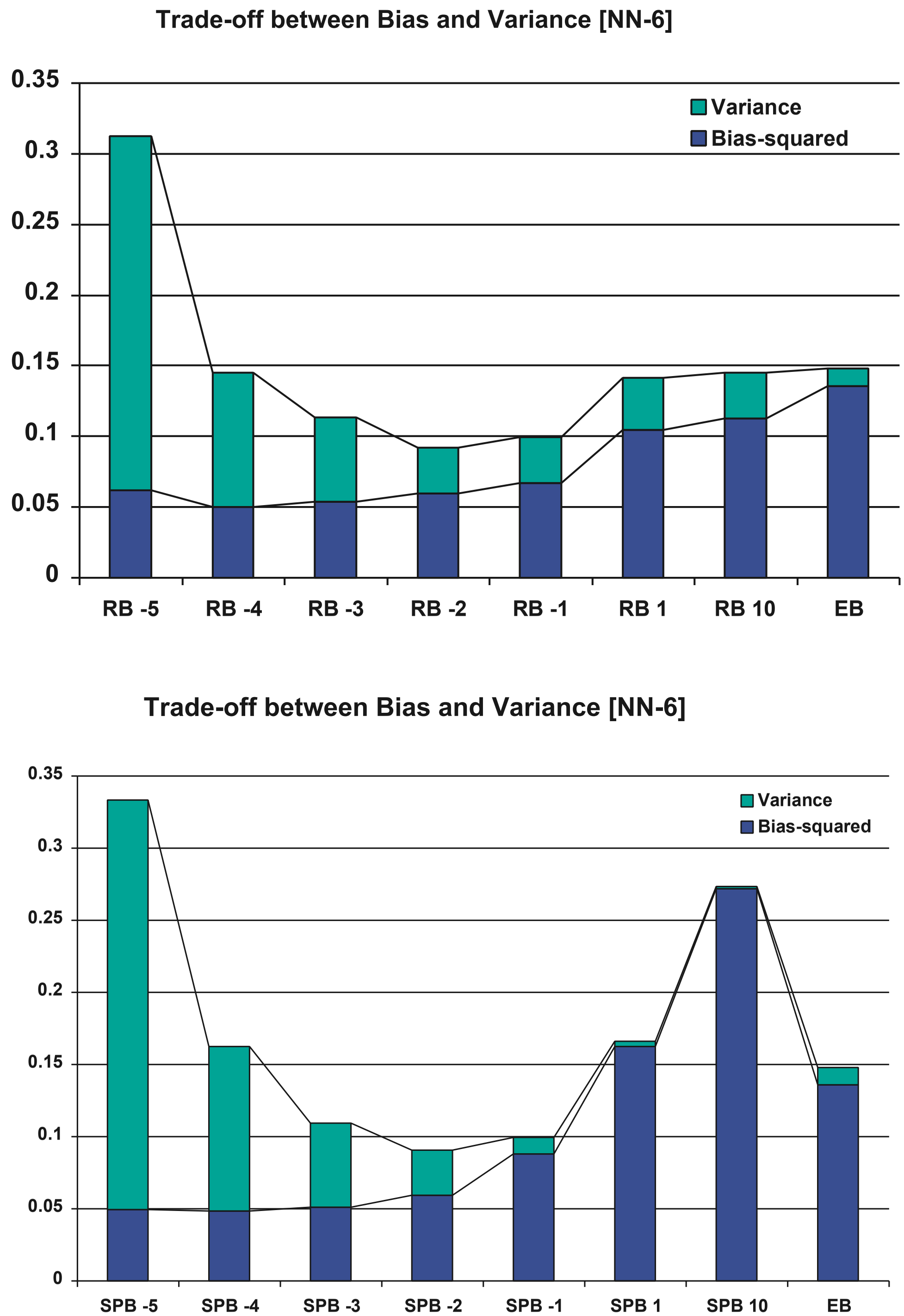


Fig. 3.2. The upper panel compares the robust Bayes (RB) method with the empirical Bayes (EB) method, and the lower panel recites the comparison between EB and single-prior Bayes (SPB) method. When k is too large ($k = 1, 10$ or higher), the MSPE from RB levels off at the level of EB, while the MSPE from SPB continues to increase unboundedly. When k is in a reasonable region ($1e - 4 < k < 1$), RB delivers a better performance than EB by carrying more character of SPB and ML.

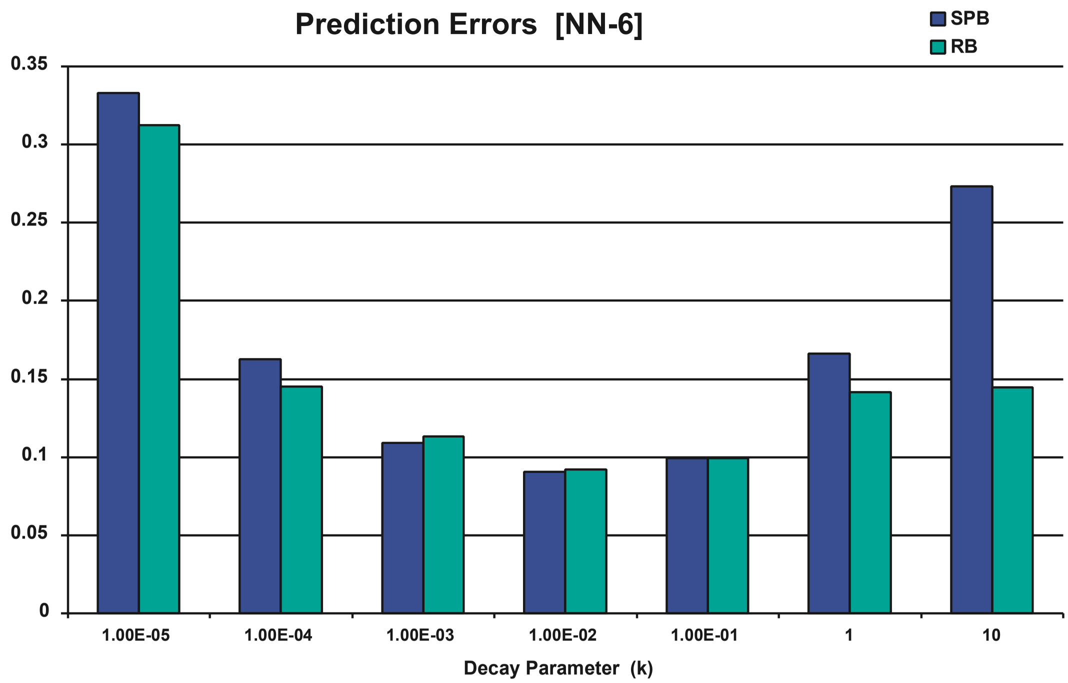


Fig. 3.3. The MSPEs from neural network regression model with six hidden units trained by single-prior Bayes and robust Bayes methods are plotted against each other (see Figure 3.1 for bias-variance decomposition).

single choice or at least a reasonable choice of an interval for k is possible if the data are rescaled to ensure an efficient use of the capacity that the model can provide. Ripley [36] has made a simple argument for this subject. Since the sigmoidal function used in neural network models saturates for domain valued around ± 3 and the inputs $\boldsymbol{x}$ are scaled to the range $[0, 1]$, the standard deviation of the parameters are expected around 5, which suggests a range from 0.001 to 0.1 for k. Secondly, the random walk of $\hat{k}$ allowed in EB has two different effects on the outcomes: it has certain appeal when the guess of k is terribly wrong (be it too large or too small), but $\hat{k}$ always overshrinks the parameters so further constraints are in order. A possible solution is to use the adaptive $\hat{k}$ from EB while adding an guessed upper bound from SPB on $\hat{k}$ so that $\hat{k}$ does not grow too high. With the additional constraint and compromise imposed by r_q form RB, our simulation shows that this scenario is able to improve overall prediction performance of the neural network regression model in the following ways (see Figures 3.1, 3.3, 3.2, 3.4 and 3.5):

1. When the guessed k (the single choice of the upper bound for the adaptive $\hat{k}$ from EB) is wrong (too high or too low), the new estimator, $\hat{\boldsymbol{\theta}}^{RB}$, has a lower MSPE than that of SPBs. If the guessed k is too low, $\hat{\boldsymbol{\theta}}^{RB}$ bears more character of EB with a higher bias but lower variance and a lower MSPE. If the guessed k is too high (the MSPE can go unbounded for SPB), $\hat{\boldsymbol{\theta}}^{RB}$ shows more character of a ML estimator with a lower bias but higher variance and a lower MSPE that levels off at the level of an EB estimator.

2. When the guessed k is about right, $\hat{\boldsymbol{\theta}}^{RB}$ delivers virtually the same MSPE as the SPBs, and avoids a higher MSPE as expected for EB by not overshrinking the parameters.

3. The convex coefficient r_q plays a role in shrinking (which is good) the adaptive $\hat{k}$. A smaller $\hat{k}_\tau$ could lead to a smaller $||\hat{\boldsymbol{\theta}}_\tau||^2$ and so a smaller $\hat{r}_{q\tau}$. A smaller $\hat{r}_{q\tau}$ lowers the penalty imposed by the ridge procedure (the third term in (3.20)) so that the line search procedure in the optimization algorithm one uses has more freedom in choosing the next update of the parameter vector. The trend in parameter magnitude of an RB run is a steady increase after the initial disturbance, instead of steady decrease as in an EB run. This together with a steadily decreasing error (the first term in (3.20)) leads to possibly even lower $\hat{k}_{\tau+1}$ for the next iteration step in optimization. This also results in a lower model bias, since the penalty on the the parameter magnitude is lessened by r_q so that the resulting model bears more character of ML method with low bias.

Besides the experiments on the ozone data, the simulations on the synthetic data with a wide variety of data settings also show overall performance gains by RB (see Figures 3.6 and 3.7).

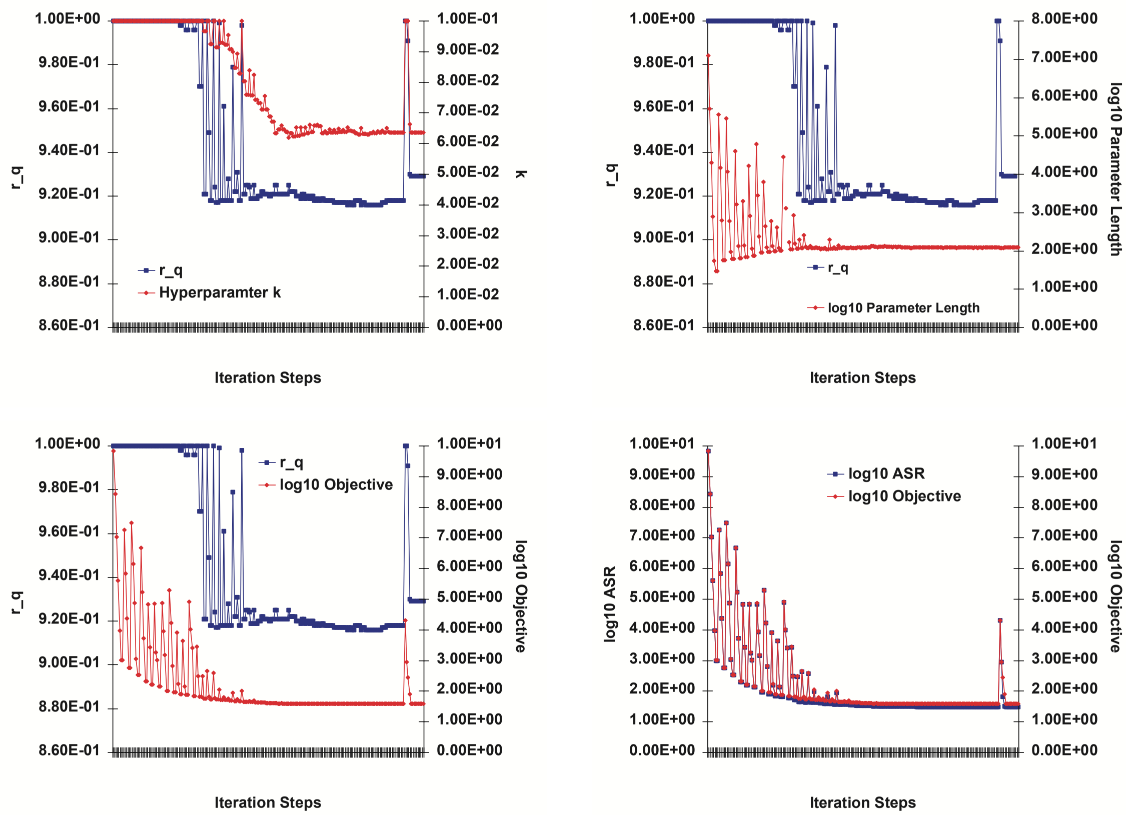


Fig. 3.4. The four panels provide a look at the iterative changes of five quantities (the function r_q, the adaptive hyperparameter k, parameter magnitude $||\boldsymbol{\theta}||^2$, the squared error and the new objective function) in a Newton-Raphson procedure when the robust Bayesian approach is used in a neural network ($h = 6$) on ozone data. The iteratively updated hyperparameter k's are pulled back by a decreasing r_q, so that the parameters are not shrunk as much as by the EB method.

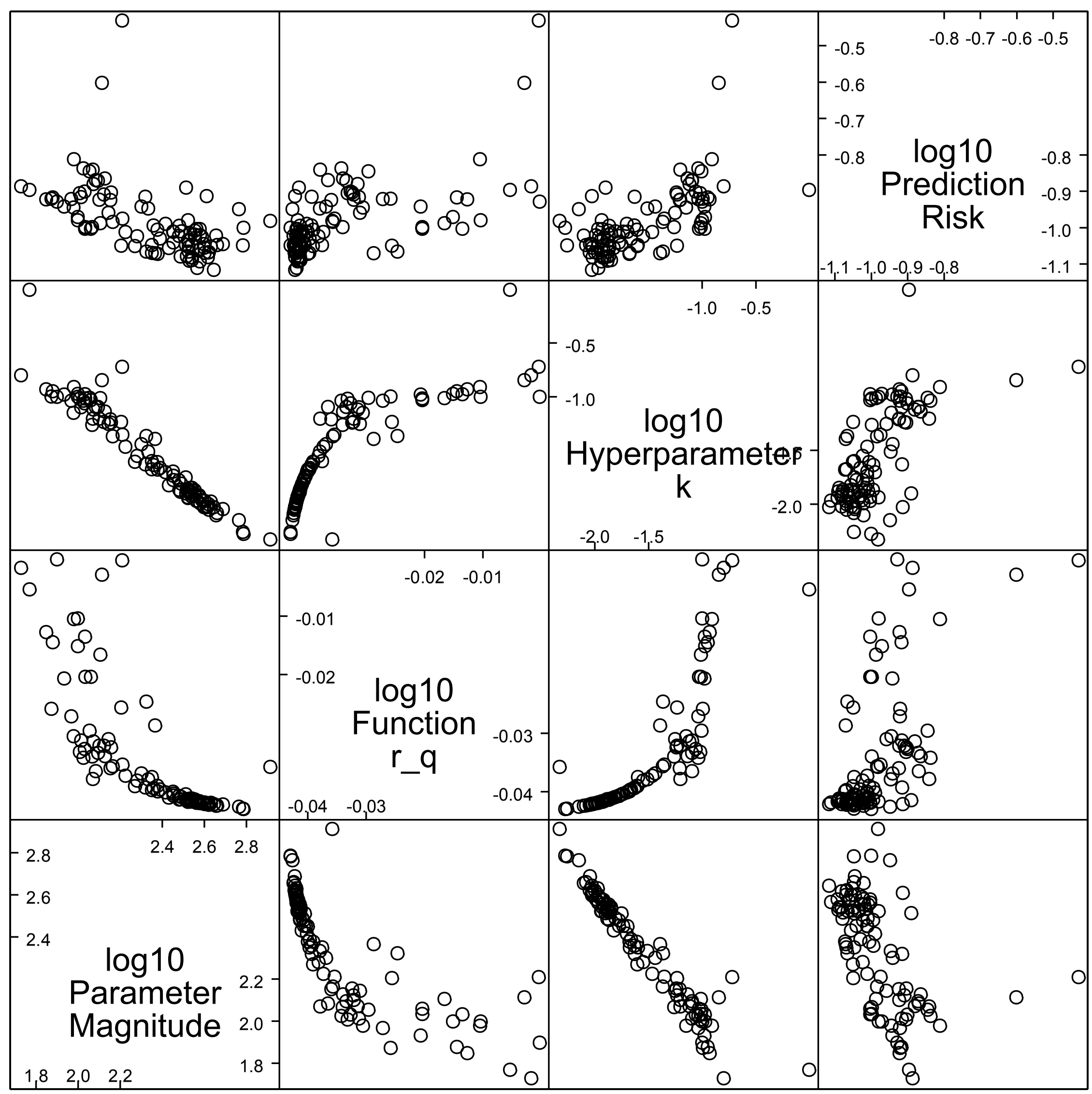


Fig. 3.5. The relations among estimated parameter magnitude, the value of function r_q, hyperparameter k and prediction risk are illustrated by 100 neural network models trained by robust Bayes method with six hidden units on ozone data. All four quantities take their values at the end of Newton-Raphson iteration. Evidently, when r_q decreases, the hyperparameter decreases as well, the parameter length then increases as the result of less shrinkage, and the prediction risk decreases due to the lower model bias.

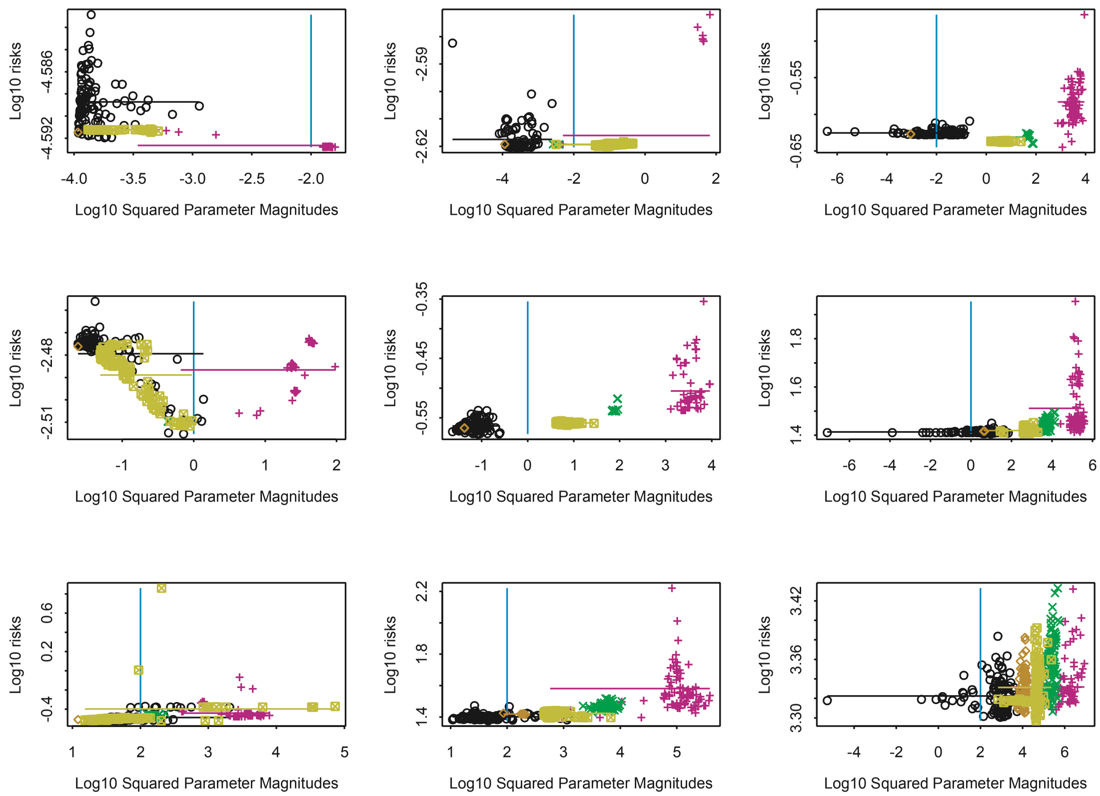


Fig. 3.6. The robust Bayesian approach (RB) ('⊠') is compared with the empirical Bayes (EB) estimator ('○') and the single-prior Bayes estimators with $k = 0.0001$ ('+'), $k = 0.01$ ('×') and $k = 1$ ('⋄') on synthetic data set I in Appendix A with the same setting as in Figure 2.6. It can be observed that the resulting models from RB show characteristics from both SPB and EB, and can be seen as compromises between these two methods. In most cases, RB delivers the best performances or nearly so, while there is no single implementation from either SPB or EB showing such overall improvement with wide variety of data situation. See Figure 3.7 for the corresponding boxplots of prediction risks.

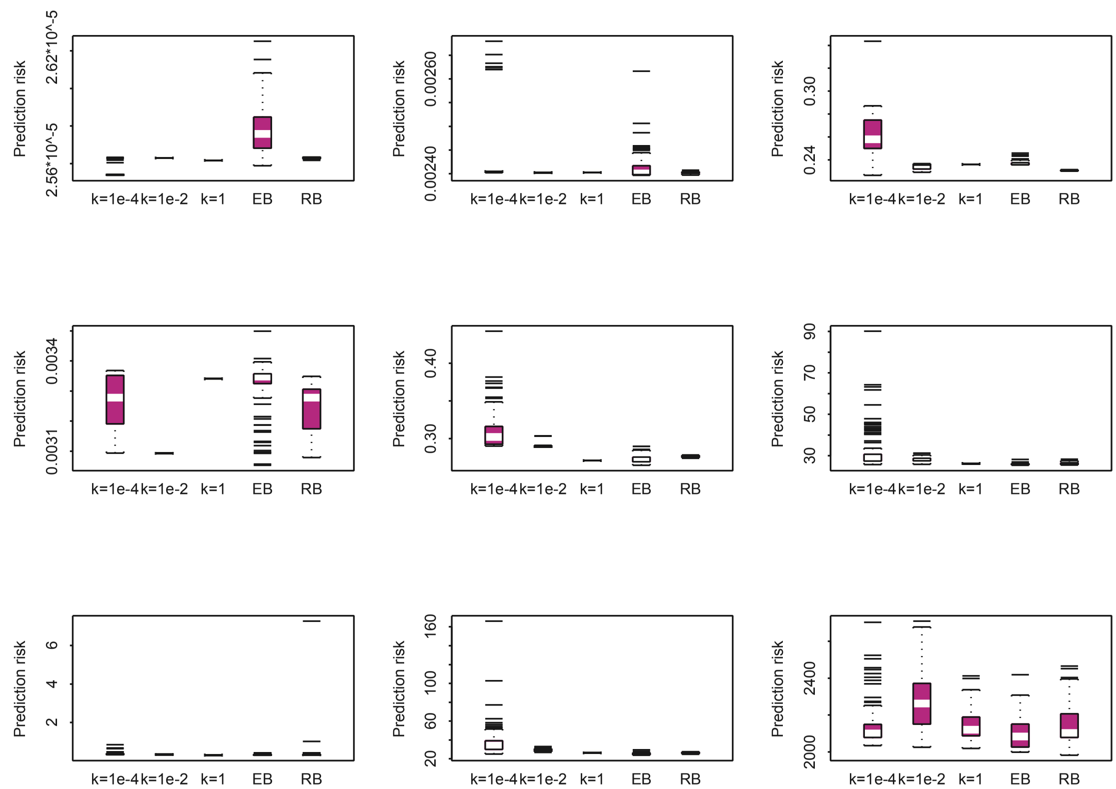


Fig. 3.7. The corresponding boxplots of prediction risks from Figure 3.6.

4. Extensions and Future Work

Implication of the new objective function

Neural network belongs to a class of modern regression model that possesses strong approximation capacity and not-so-slow convergence rate even when the dimensionality of the data is moderately high. Like all other nonparametric models, however, a potentially rather high model variability could undermine its overall performance. We have shown that a carefully designed model bias must be introduced to lower the model variance so that the total prediction risk is reduced. The new robust Bayesian estimator (3.21) developed here implies a new corresponding constricted objective function

$$||\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{\theta})||^2 + (1 - r_q)(\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{\theta}))'\boldsymbol{F}'\boldsymbol{K}(\boldsymbol{F}'\boldsymbol{F})^{-1}\boldsymbol{F}(\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{\theta})) + r_q\boldsymbol{\theta}'\boldsymbol{K}\boldsymbol{\theta} \ . \qquad (4.1)$$

The third term in (3.20) is the familiar penalty term from the single-prior and empirical Bayes methods. The second term is a weighted version of the second term in the linear approximation of the loss function of a prediction action by a ML estimator in (2.12). This term is also called gradient projection in optimization theory, because $\boldsymbol{F}(\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{\theta}))$ is the gradient vector of the quadratic loss function and $(\boldsymbol{F}'\boldsymbol{F})^{-1}$ is the inverted Hessian matrix. It is the projection of the gradient vector on the tangential plane, and can be seen as a measure of nonlinearity around a point $\boldsymbol{\theta}$ where the linear approximation of the loss function is carried out. Hence, an additional penalty is introduced when the underlying nonlinearity is high at the neighborhood of $\boldsymbol{\theta}$ and the linear approximation is inadequate. By combining this weighted projection convexly with the third term, the new objective function takes one more natural measure of smoothness into account. The above discussion is rather informal; we leave more precise technical examination for future work.

Improved confidence regions

There has been little or no attention paid on improving the confidence interval estimation in neural network regression model. We have shown in Chapter 2 that various Bayesian modifications in parameter estimation have a direct impact on their corresponding confidence intervals, because the estimated parameter vector and the estimated confidence intervals come from the first two moment summary of parameter posteriors. It is of great interest to investigate the size and probability of coverage of the ellipsoid defined by the estimated parameter vector (as the posterior mean) and the posterior covariance matrix. And it is also rather interesting to examine the conditions under which the new estimation procedure has not only an improved parameter estimation but also more concentrated corresponding confidence regions.

Small sample asymptotics

So far the outcome of statistical inference on a regression model like the neural network is summarized by the posterior mean and posterior covariance, which is under the asymptotic assumption that the posterior density can be approximated by a normal distribution when the sample size is large. However, this is usually not the case in practice, and the posterior density typically is multimodal and skewed. A small-sample asymptotic inference is in order, when some extra accuracy in prediction is desired. It is unclear at the time being whether this shall benefit at all and how much performance gain one could possibly obtain for a neural network regression model. But a great deal of research in mathematical statistics indicates that this is one of the directions worth consideration. For example, a small-sample asymptotic treatment of logistic regression model (a single-layer feedforward neural network with no hidden layer) by Strawderman, Casella and Wells [55] is shown to be beneficial.

Experimental design and model selection

Other steps in the data analysis process can also benefit from this unified statistical framework based on global-error-property analysis. For example, the preprocessing step is not well formulated and optimized yet in neural network regression. A preposterior and sequential analysis procedure can address the underlying experiment design and sampling issues using a likelihood-based formulation and answer questions such as: what is the sufficient amount of data needed for a prediction or classification of a desired accuracy [24]. An analogous likelihood-based formulation can be utilized for the postprocessing step (e.g., model assessment) as well [56, 57]. Once a likelihood-based framework is formulated, the rest of the analysis can be the same as in the data modeling step investigated in this thesis. Such a global-error-property analysis provides concrete measures and evaluation criteria that allows one to evaluate the relative advantages and disadvantages of different estimation procedures, sampling schemes and post-processing methods.

A. Data Sets

A.1 Ozone Data

The ozone data set analyzed in [58, 59, 60, 3] is composed of one response variable `ozone` and nine predictors with 330 records in 1976. The name list of the variables is as follows:

`ozone`: The daily maximum of the hourly-average atmospheric ozone concentrations in Upland, California.

`vh`: 500 millibar pressure height at the Vanderberg air force base.

`wind`: Wind speed (mph) at Los Angeles airport (LAX).

`humidity`: humidity (%) at LAX.

`temp`: Temperature (^{o}F) at the Sandberg air force base.

`ibh`: Temperature inversion base height (feet).

`dpg`: Pressure gradient (mm Hg) from LAX to Daggert, California.

`ibt`: Inversion base temperature (^{o}F) at LAX.

`vis`: Visibility (miles) at LAX.

`doy`: Day of the year.

A.2 Synthetic Data

Following a typical setting of simulation proposed for examining ridge procedures in linear regression [48, 31, 32, 30], two synthetic data sets are created. The data sets are composed of one response variable Y and three (or nine) uniformly distributed predictors $\boldsymbol{X} = (X_1, ..., X_d)' \in [0,1]^d$, where data set I with $d = 3$ is designated to represent relatively small neural network model and data set II with $d = 9$ resembles the situation of ozone data when a large network model is needed. For data set I (II), there are $q = 16$ (100) parameters $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$ in a feedforward neural network model with $h = 3$ (9) hidden units (without skip layer)

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{k=1}^{h} \beta_k g(\sum_{i=1}^{d} x_i \alpha_{ki} + \alpha_0) + \beta_0 \ .$$

Three signal-to-noise ratio (SNR), $f^2(\cdot)/\sigma_\varepsilon^2$, are used at the values of 100, 1 and 0.01, so that the significance of the parameters varies from very high, about even, to very low. Three different 'true' parameter magnitudes, $\boldsymbol{\theta}'\boldsymbol{\theta}$, are also used at the values of 0.01, 1 and 100 for data set I and 0.1, 10, 1000 for data set II, so that there are 9 cases combined for each data set. $\boldsymbol{\theta}$ is created as a uniform random vector in $[-1/2, 1/2]^q$, and then rescaled so that $\boldsymbol{\theta}'\boldsymbol{\theta} = r^2$ with $r^2 = 0.01, 1, 100$ for data set I and $r^2 = 0.1, 10, 1000$ for data set II. For each $\boldsymbol{\theta}'\boldsymbol{\theta}$, 400 (2000) $\boldsymbol{X}$'s are created for data set I (II) and 400 (2000) Y's are then calculated for each of the 3 levels of SNR with a normal distributed noise added

$$y = f(\boldsymbol{x};\boldsymbol{\theta}) + \varepsilon \ ,$$

where $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ and $\sigma_\varepsilon^2 = f^2(\cdot)/SNR$. For each of 9 cases of the combination of $\boldsymbol{\theta}'\boldsymbol{\theta}$ and SNR, $n = 200$ (1000) pairs of $(\boldsymbol{X}, Y)$ are used as training set, and the rest $n = 200$ (1000) as test set.

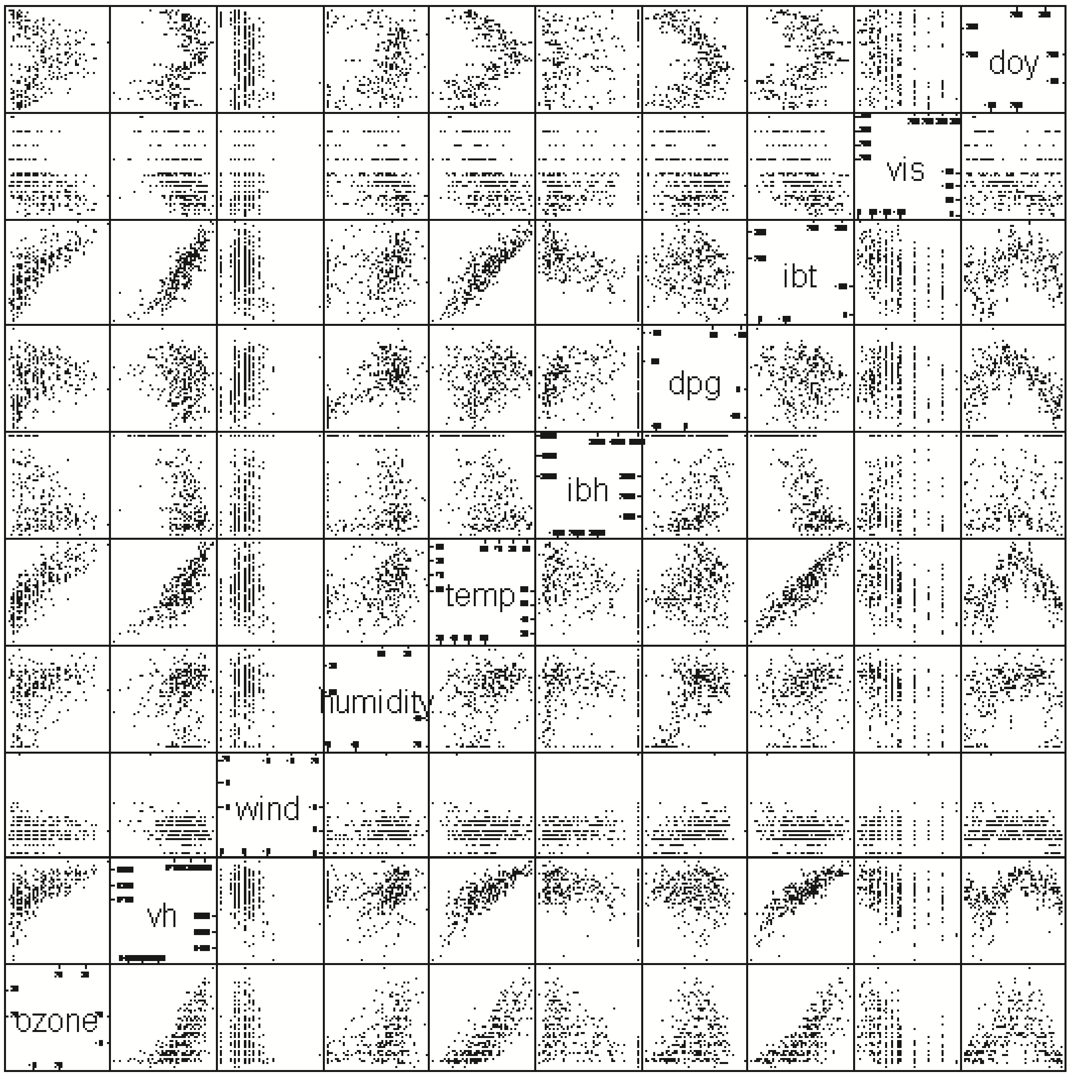


Fig. A.1. The scatterplot of the ozone data.

LIST OF REFERENCES

[1] E. L. Lehmann and George Casella. *Theory of Point Estimation (Second Edition)*. Springer-Verlag, New York, 1998.

[2] Charles J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053, 1982.

[3] Trevor J. Hastie and Robert Tibshirani. *Generalized Additive Models*. Chapman and Hall, London, 1990.

[4] Andrew R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39:930–945, 1993.

[5] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 2:303–314, 1989.

[6] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2:359–366, 1989.

[7] Charles J. Stone. Additive regression and other nonparametric models. *The Annals of Statistics*, 13(2):689–705, 1985.

[8] Charles J. Stone. The dimensionality reduction principle for generalized additive models. *The Annals of Statistics*, 14(2):590–606, 1986.

[9] E. Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33:1065–1076, 1962.

[10] M. Rosenblatt. Remarks on some non-parametric estimates of a density function. *The Annals of Mathematical Statistics*, 27:642–669, 1956.

[11] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London, 1986.

[12] E. A. Nadayara. On estimating regression. *Theory of Probability and its Applications*, 10:186–190, 1964.

[13] G. S. Watson. Smooth regression analysis. *Sankhyā A*, 26:359–372, 1964.

[14] Grace Wahba. Smoothing noisy data with spline functions. *Numer. Math.*, 24:383–393, 1975.

[15] B. W. Silverman. Spline smoothing: The equivalent variable kernel method. *The Annals of Statistics*, 12(3):898–916, 1984.

[16] Charles J. Stone. Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, 8(6):1348–1360, 1980.

[17] L. K. Jones. On the conjecture of huber concerning the convergence of prejection pursuit regression. *The Annals of Statistics*, 15(2):880–882, 1987.

[18] L. K. Jones. A simple lemma on greedy approximation in hilbert space and convergence rates for projection pursuit regression and neural network training. *The Annals of Statistics*, 20(1):608–613, 1992.

[19] Andrew R. Barron. Approximation and estimation bounds for artificial neural networks. *Machine Learning*, 14:114–133, 1994.

[20] Federico Girosi and Gabriele Anzellotti. Rates of convergence for radial basis functions and neural networks. In Richard J. Mammone, editor, *Artificial Neural Networks for Speech and Vision*, pages 97–114. Chapman & Hall, London, 1993.

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

[22] Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Applications to nonorthogonal problems. *Technometrics*, 12(1):69–82, 1970.

[23] Arthur E. Hoerl and Robert W. Kennard. Erratum to ridge regression: Applications to nonorthogonal problems. *Technometrics*, 12:723, 1970.

[24] James O. Berger. *Statistical Decision Theory and Bayesian Analysis (Second Edition)*. Springer-Verlag, New York, 1985.

[25] W. James and C. Stein. Estimation with quadratic loss. In Jerry Neyman, editor, *Proceedings of the Fourth Berkeley Symposium, Vol. I*, pages 361–379. University of California Press, Berkeley and Los Angeles, 1961.

[26] Charles M. Stein. A necessary and sufficient condition for admissibility. *The Annals of Mathematical Statistics*, 26:518–522, 1955.

[27] Charles M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151, 1981.

[28] James O. Berger. Minimax estimation of location vectors for a wide class of densities. *The Annals of Statistics*, 3(6):1318–1328, 1975.

[29] Harold M. Hudson. *Empirical Bayes Estimation*. PhD thesis, Stanford University, 1974.

[30] Ronald A. Thisted. *Ridge regression, minimax estimation, and empirical Bayes methods*. PhD thesis, Stanford University, 1976.

[31] A. E. Hoerl, R. W. Kennard, and Kent F. Baldwin. Ridge regression: Some simulations. *Communications in Statistics*, 4:105–123, 1975.

[32] A. E. Hoerl and R. W. Kennard. Ridge regression: Iterative estimation of the biasing parameter. *Communications in Statistics*, A5:77–88, 1976.

[33] Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58, 1992.

[34] M. Stone. Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of Royal Statistical Society*, 36(2):111–147, 1974.

[35] Wray L. Buntine and Andreas S. Weigend. Bayesian back-propagation. *Complex Systems*, 5:603–643, 1991.

[36] B. D. Ripley. *Pattern Recognition and Neural Networks.* Cambridge University Press, Cambridge, 1996.

[37] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S-PLUS (Second Edition).* Springer, New York, 1997.

[38] David J. C. MacKay. A practical bayesian framework for backprop networks. *Neural Computation*, 4:448–472, 1992.

[39] Radford M. Neal. *Bayesian Learning for Neural Networks.* PhD thesis, University of Toronto, 1995.

[40] Peter Müller and David Rios Insua. Issues in bayesian analysis of neural network models. *Neural Computation*, 10:749–770, 1998.

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

[42] Charles J. Stone. Large-sample inference for log-spline models. *The Annals of Statistics*, 18(2):717–741, 1990.

[43] B. Efron and R. Tibshirani. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1(1):54–77, 1986.

[44] Xiaohong Chen and Halbert White. Improved rates and asymptotic normality for nonparametric neural network estimators. *IEEE Transactions on Information Theory*, 45(2):682–691, 1999.

[45] Xiaotong Shen. On the methods of sieves and penalization. *The Annals of Statistics*, 25(6):2555–2591, 1997.

[46] D. M. Bates and D. G. Watts. Relative curvature measures of nonlinearity (with discussion). *Journal of Royal Statistical Society*, 42(1):1–25, 1980.

[47] P. J. Brown and J. V. Zidek. Adaptive multivariate ridge regression. *The Annals of Statistics*, 8(1), 1980.

[48] N. R. Drapper and R. Craig Van Nostrand. Ridge regression and james-stein estimation: Review and comments. *Technometrics*, 21(4):451–466, 1979.

[49] William E. Strawderman. Minimax adaptive generalized ridge regression estimators. *Journal of American Statistical Association*, 73:623–627, 1978.

[50] William J. Hemmerle. An explicit solution for generalized ridge regression. *Technometrics*, 17(3):309–314, 1975.

[51] Lawrence D. Brown. Minimaxity, more or less. In S. S. Gupta and J. O. Berger, editors, *Statistical Decision Theory and Related Topics V*, pages 1–18. Springer-Verlag, New York, 1994.

[52] William E. Strawderman. Proper bayes minimax estimators of the multivariate normal mean. *The Annals of Mathematical Statistics*, 42(1), 1971.

[53] James O. Berger. Minimax estimation of a multivariate normal mean under arbitrary quadratic loss. *Journal of Multivariate Analysis*, 6:256–264, 1976.

[54] James O. Berger. A robust generalized bayes estimator and confidence region for a multivariate normal mean. *The Annals of Statistics*, 8(4):716–761, 1980.

[55] Robert L. Strawderman, George Casella, and Martin T. Wells. Practical small-sample asymptotics for regression problems. *Journal of American Statistical Association*, 91:643–654, 1657, 1996.

[56] Charles J. Stone. Local asymptotic admissibility of a generalization of akaike's model selection rule. *Ann. Inst. Statist. Math.*, 34A:123–133, 1982.

[57] Charles J. Stone. Admissibility and local asymptotic admissibility of procedures which combine estimation and model selection. In S. S. Gupta and J. O. Berger, editors, *Statistical Decision Theory and Related Topics III (2)*, pages 317–333. Academic Press, New York, 1982.

[58] L. Breiman and Jerome H. Friedman. Estimating optimal transformations for multiple regression and correlation (with discussion). *The Journal of American Statistical Association*, 80:580–619, 1985.

[59] Andreas Buja, Trevor Hastie, and Robert Tibshirani. Linear smoothers and additive models (with discussions). *The Annals of Statistics*, 17(2):453–555, 1989.

[60] Jerome H. Friedman and Bernard W. Silverman. Flexible parsimonious smoothing and additive modeling (with discussions). *Technometrics*, 31(1):3–39, 1989.

VITA

1964: Born January 29 in Shanghai, the People's Republic of China.

1980-1985: Attended Shanghai Jiao Tong University, Shanghai, P. R. China; majored in Biomedical Engineering.

1985: B.S., Shanghai Jiao Tong University.

1985-1988: Graduate work in Engineering; Shanghai Jiao Tong University.

1985-1988: Research Assistant, Department of Precision Instruments, Shanghai Jiao Tong University.

1988: M.S., Shanghai Jiao Tong University.

1988-1992: Research Associate, Laboratory of Visual Information Processing, Institute of Biophysics, Chinese Academy of Sciences, Beijing, P. R. China.

1993-1999: Graduate work in Electrical and Computer Engineering; School of Electrical and Computer Engineering, Purdue University, West Lafayette, Indiana.

1993-1997: Research Assistant, School of Electrical and Computer Engineering, Purdue University.

1996: Summer Research Assistant, NEC Research Institute, Princeton, New Jersey.

1997-1999: Principal Engineering Aid, Electrical Engineering Department, University of California, Los Angeles, California.

1999: Ph.D. in Electrical and Computer Engineering, Purdue University.