Robustness in Generalized Ridge Regression and Related Topics *

by

Herman Rubin
Purdue University

Technical Report #87-27

Department of Statistics
Purdue University

June 1987

* Research supported by the National Science Foundation under Grant DMS-8401996.

⋆

Robustness in Generalized Ridge Regression

and Related Topics

by

Herman Rubin [1]
Purdue University
and
The Ohio State University

ABSTRACT

We start out by considering the formal Bayesian version of generalized ridge regression, namely, we wish to make inferences concerning the mean of a multivariate normal distribution with normal prior, or to decide between normal priors. For estimation, if we restrict ourselves to linear procedures, the assumption of normality becomes unimportant, and in many situations the normality assumption will not be too important in practice. We will also discuss to some extent the importance of the normality assumption, both in the prior and in the distribution of the "errors". Our purpose in this paper is to attack the problem of the sensitivity to assumptions, mainly the assumptions made about the prior covariance matrix. In addition, we present some empirical Bayes type procedures, which are good in some cases. We also present some problems which are not of a regression type for which the general results apply.

Introduction

We consider the problem of estimating the regression parameter from the Bayesian viewpoint. This problem was explicitly discussed by many authors, for example, by the present author in Rubin [1974]. (At the same time, the notion of ridge regression was introduced in Hoerl and Kennard[1970]. The formal similarity between ridge regression and a special case of the problem under discussion was easily seen, for example, in Goldstein and Smith [1974]. Since ridge regression was applied to low dimensional problems where the major problem was multicollinearity, the results are not particularly sensitive to the form of the prior, but are sensitive to its scale, the "ridge prior", a multiple of the identity matrix for the appropriate scaling, may be appropriate. This is, in general, *not quite* a Bayes procedure, and does not work in the problems we are considering.) We will be concerned with the problem of inference, *not* with the problem of prediction.

We wish to study robustness for large samples for problems with a large or even infinite number of dimensions. The asymptotics for infinite dimensional problems can differ greatly from that of the usual fixed dimensional versions. There are two good reasons for taking the infinite dimensional viewpoint; large finite dimensional problems are more likely to look like infinite dimensional problems than low dimensional problems for reasonable sample

[1] Research partially supported by the National Science Foundation under Grant DMS-8401996.

sizes, and there are many problems, including most of those for which the misnamed "nonparametric" procedures are used, which are truly infinite dimensional.

The basic regression problem is inference on the location parameter of a distribution. After the usual reduction, this means that we have a vector θ of parameters, and we observe a vector $y = \theta + x$. We furthere need that x is independent of, or at least uncorrelated with, θ to avoid identification problems. Of course, other assumptions can be made, but would you call inference on the sample size of a binomial distribution a regression problem? In the classical regression problem, the Gauss-Markoff Theorem is the basic robustness theorem. The arguments made in that theorem require modification for Bayesian or even decision-theoretic purposes. If we keep linearity but drop unbiasedness, which we must in high dimensional cases to get anything remotely reasonable, we obtain the usual Bayes procedures described below. Accordingly, let us act in this manner.

Let us therefore make the assumption that θ is normal $(0, \mathrm{T})$, x is normal $(0, \Sigma)$, and θ and x are independent. We wish to investigate the effect of misspecification of the problem on the Bayes risk of the resulting procedures for estimation and testing. This is the general problem of robustness. Other than observing that the central limit theorem can be used to justify normal errors, we consider only the problem of misspecification of the prior. This paper should be considered only as a first step in the investigation; the problem is quite complex.

There are some questions on the relevance of this paper. First, we are operating under the assumption of normality, and it is well known (?) that nothing is normal. If we look only at the estimation problem under quadratic loss and consider only linear estimators, the question is irrelevant; there are other reasons, which we will point out later, which indicate that it might not be too important in many situations. However, if we wish to use an "Occam's razor" prior, or something similar, and there is a complexity cost for the estimate, i.e, estimating a coordinate of θ to be non 0 increases the cost, we suspect that the situation may be much worse. This problem is also discussed in Rubin [1974].

Second, a Bayesian specifies his prior and loss, so where is the problem? Even if we could assume that the "error" x is exactly normal, only a rash Bayesian would be that sure of his prior. The "ordinary" ridge regression priors can find justification only from a non-Bayesian viewpoint, and this does not work in the high dimensional cases, as we will demonstrate. The prior chosen will certainly be wrong; which errors are the most important, and which can essentially be neglected?

Third, let us consider the axiomatic Bayesian (rational decision theorist). He knows that coherence requires a Bayesian procedure, but also requires an infinitely fast computer operating at zero cost. Not having one available, approximate solutions are needed; this leads to the same problem as before.

Finally, take the user of statistical procedures. We are discussing problems where Rubin's first commandment, "Thou shalt know that thou must make assumptions," cannot be ignored. We must inform our clients about which are the important assumptions.

Since the prior assumptions matter, one may ask whether non-linear techniques, such as "empirical Bayes" type procedures, can improve the situation. The answer here is decidedly yes, but we have been able to obtain such procedures only in special cases. One thing to avoid is the reckless use of the hyperparameter approach. A discussion of a special problem of Bayesian testing in the infinite dimensional normal case can be found in Cohen [1972]. The results can be extendend to other situations, such as hyperparameter inference. It is possible, and indeed rather likely, for a few coordinates for which the prior variance is large to dominate a Bayesian hyperparameter inference; however, these coordinates do not appreciably affect the procedure. For the one class of problems for which we believe we have a reasonable empirical Bayes procedure, we do not take that approach.

Robustness of Bayes Procedures

Let us suppose, as in the Introduction, that the random variable y is normally distributed with mean θ and covariance matrix Σ, amd that the parameter θ is normal with mean 0 and covariance matrix T. Then it is well known that if the loss of the estimate t of θ is $(t-\theta)'A(t-\theta)$ the Bayes estimator $\hat{\theta}$ of θ is $\mathrm{T}(\Sigma+\mathrm{T})^{-1}x$, and its Bayes risk is $\mathrm{tr}(A\Sigma(\Sigma+\mathrm{T})^{-1}\mathrm{T})$. It is an easily proved theorem that the matrix multiplying A is unchanged when Σ and T are interchanged, and hence is symmetric. (The problem is essentially unchanged if A is not present, but many problems are more easily formulated with it, and it adds no essential complications.)

Now suppose that the situation is as above, but that we incorrectly use Υ instead of T in computing the estimator. Using the matrix identity $R^{-1}-S^{-1}=R^{-1}(S-R)S^{-1}$, we find that the risk is now increased by

$$\mathrm{tr}A\Sigma(\Sigma+\Upsilon)^{-1}(\mathrm{T}-\Upsilon)(\Sigma+\mathrm{T})^{-1}(\mathrm{T}-\Upsilon)(\Sigma+\Upsilon)^{-1}\Sigma.$$

Note that this expression is *not* symmetric in T and Υ. This is not surprising. This asymmetry is apparent even in the one-dimensional case. In that case, the Bayes procedure corresponds to multiplying the observation y by $\lambda=\tau^2/(\sigma^2+\tau^2)$. If we wish to allow the risk to increase by a factor of c, we may increase or decrease λ by $\sqrt{(c-1)\lambda(1-\lambda)}$. For λ near 0, this interval includes 0, and for λ near 1, it includes 1. The graph in the appendix shows the range by which the multiplier can vary for a given risk factor. In fact, if we set $c=2$, we can use 0 for $\lambda\leq 0.5$ and we can use 1 for $\lambda\geq 0.5$. That is, we at most double the Bayes risk if we ignore the less concentrated of the data and the prior. This fact was used by the author to obtain preliminary results quickly.

The asymmetry is quite pronounced. In the case of large variance of the prior, even for one coordinate using too low a variance can be catastrophic. For example, if $\sigma=1$ and we assume $\tau=10$, the risk is 99.01 for squared error loss if in fact $\tau=1000$, while the Bayes risk is 0.999999. However, if we assume $\tau=\infty$, we obtain a risk of 1, while the Bayes risk for $\tau=10$ is 0.990099. Now this problem can be avoided (see Rubin [1977]) by using a non-linear estimate, and several non-normal priors are considered. However, if the prior is concentrated and does not have a variance, not much can be done about robustness. In the case of a small variance, if we assume $\tau=0$, the error committed is small. Having a moderate error in one or a few coordinates is not too important in the case of low prior variance, but for some priors considered here, the contribution to the total Bayes risk by these coordinates is an appreciable portion of the total, and using too large prior values for τ can be important. Using too small values has little effect on the risk. Certainly little is lost by using 0 if τ is small and using ∞ if it is large.

Since using non-normal priors gives good results for large values of τ even if τ is underestimated, can this method be useful for small values of τ? Unfortunately, the answer is no. In the infinite dimensional problem, there may be millions of coordinates with $10^{-6} < \tau < 10^{-3}$. For these coordinates, the observations will not be much different than if τ were 0, and any attempt to guard against tails in the prior founders badly if there are moderate tails, such as t with 20 degrees of freedom or even logistic, in the distribution of the "errors" x. This problem also certainly occurs if the error variance must be estimated.

Specific Results

One of the problems leading to this work is the attempt to obtain an approximate Bayes estimator for densities, suggested by the model in Chen and Rubin [1986]. The problem is somewhat more complicated than this, but it suggests the problem of estimating the mean θ of an infinite dimensional vector in the special case in which the covariance matrix for a single observation is the identity and the sum of the squares of the elements of θ is 1 should be a good starting point for the solution. The method for inference on the "tail" of θ will be close to what should be done in the real problem. Now in this case the prior for θ certainly cannot be normal with mean 0, and the observations certainly are not normal; however, the central limit theorem tells us that the "errors" are approximately normal, and estimates of the early coordinates of θ will not be particularly affected by the normality assumptions. We have seen that the late ones will not be particularly affected either, since we may as well estimate them to be 0. Thus the effect is mainly in the middle. However, unless there is a considerable amount of non-normal dependency in the middle coordinates, either there are few coordinates, so the contribution is not large, or there are many, and again a central limit type effect occurs.

Now what happens if we use the classical ridge approach? The ridge approach does not work at all! If we have an n-dimensional problem, the ridge approach says to estimate the vector θ by maximum likelihood given the length. If we have sample size N, the method can give reasonable results only if N is considerably greater than n, since it uses a prior covariance matrix proportional to the identity. In other words, in a problem with many dimensions, we *must* have non-trivial prior input.

For simplicity in the following, we have assumed that the covariance matrix of x for each observation is the identity and that the covariance matrix of θ is diagonal. Furthermore, let us assume that the loss is the sum of squares of the errors of the estimate. In some cases for the density estimation problem, this corresponds to the Hellinger-Kakutani distance. We will discuss both the good news and the bad news.

First, the bad news. The first problem we have looked at is that for which the k-th diagonal element of T, τ_{kk}, is either k^{-2} or 2^{-k}. Now if k^{-2} is the case, the Bayes risk is on the order of $N^{-1/2}$, but if we use instead 2^{-k} it is only on the order of $1/\log N$;

while if the true prior is 2^{-k} the risk is only on the order of $(\log N)/N$, whereas assuming k^{-2} puts the risk up to the order of $N^{-1/2}$. We include a table of the comparison in the appendix.

Knowing the set of values of the diagonal elements of T is also not enough. For let $k = (2r-1)2^{s-1}, r, s = 1, \ldots$. and suppose we use, as before, k^{-2}. If we interchange r and s, we get the same diagonal elements permuted. The effect on the risk is easily computed to be enormous.

Now the good news. Suppose we know the ordering of the diagonal elements. We know that, for a sample of size N, the expected value of y_k^2 is $\tau_{kk} + \frac{1}{N}$. While we cannot expect to estimate τ_{kk} very well from the value of y_k, we can use the well-known isotonic estimator to obtain better results. There are places where the results will not be too good. These occur for k small and where there is a rapid transition in τ_{kk}. A less critical problem occurs for k large. The small k problem is unlikely to be important, especially if we robustify by using ∞ for large values of the ratio of τ to σ. The rapid transitions can only be a real problem if the risk is greatly affected; since this is important only where the ratios are of the same order of magnitude as 1, not very many terms can be affected. Furthermore, this can only be a major problem if a large coordinate of θ is classified as small. For this to happen, several previous coordinates of y have to be small, which means that it is unlikely that we do not do better overall by shrinkage. The other problem is the long tail problem for k large. Again, if we robustify by using 0 for small values of the ratio, this problem disappears. Simulation studies are being undertaken to ascertain quantitatively the penalty due to ignorance of the actual covariance matrix.

We can handle cases like $k = (2r-1)2^{s-1}, r, s = 1, \ldots$ by using bivariate isotonic estimation if we know the structure. Clearly the results will not be as good. Again, if the structure becomes so complicated that the isotonic estimates are poor, the news is bad. If we have a somewhat incorrect idea of the structure, the results are likely to be at most fair.

The observation that using only the data or the prior, whichever variance is smaller, increases the risk by at most a factor of 2, suggests that in the monotone case the procedure to use the classical estimate for those coordinates of θ for which we estimate the data variance to be smaller than the prior variance, and to estimate the remaining coordinates to be 0, should not be too bad. In cases of smoothly dropping prior variances like k^{-2}, a moderate relative increase in the risk can be expected, since if we know the prior variances, the risk is increased by a factor strictly greater than 1. If the prior variances decrease like 2^{-k}, the increase in the risk will correspond to a fixed number of coordinates. Some of the procedures in the literature, such as some spectral density estimators do this; however, estimators considered better because they use "smoother" weight functions do not always have this robustness property!

Conclusions

The high or infinite dimensional regression problem in the Bayesian setting has some easily established robustness properties, but also can provide problems from the standpoint of robustness which we have not been able to treat. This does not mean that they can not be handled, merely that this author has not been able to find the right approach. Some of them may be very difficult.

We have only considered sums of squares of the errors. Other quadratic forms can be much easier; for example, if Σ is the infinite dimensional identity matrix, the loss is $(t-\theta)'A(t-\theta)$, and tr$A < \infty$, we always have robustness for any reasonable prior, provided reasonable care is used to avoid estimating a coordinate as too small. Here the use of non-normal priors may be good, as the occasional extra-large estimate of a small coordinate of θ will be offset by its small effect on the risk.

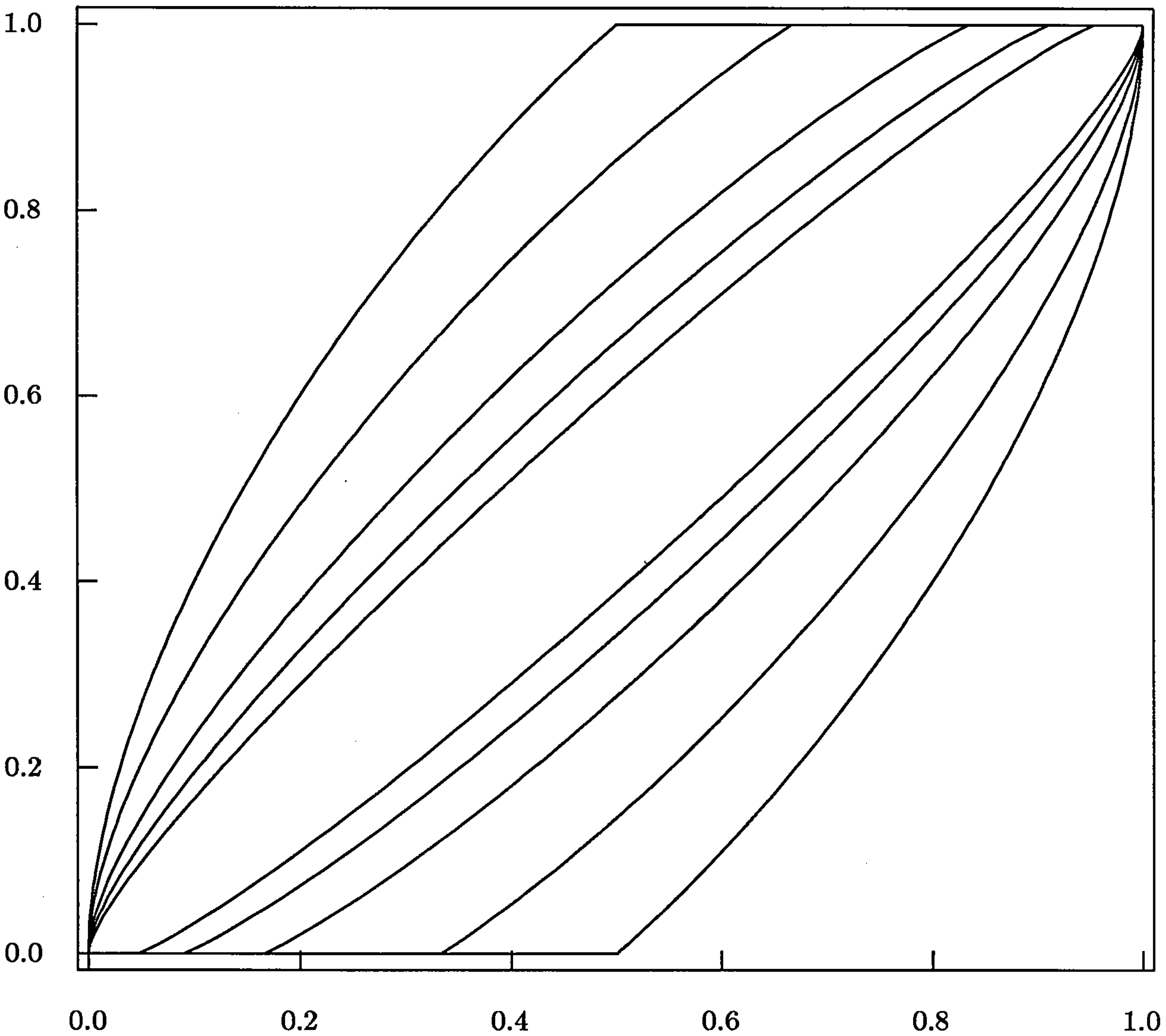
Another thing to keep in mind is that we want to consider what happens for large but not enormous sample sizes. Most of what we have done falls into this category. Some of these procedures are reasonably good even for relatively small samples.

Note that, in some situations, relatively simple procedures have good properties, and that procedures which one would think to be better are shown to be worse. Many statistical problems are high or even infinite dimensional; for such problems prior assumptions are very important, and careful mathematical reasoning is needed to establish which assumptions are important and which are not.

Acknowledgments

I would like to thank Prem Goel and Mark Berliner for their comments and discussion while the research was being done; Mark Berliner's questions were especially valuable.

APPENDIX



Contours for risk (incorrect) = c risk (correct) for c = 1.05, 1.1, 1.2, 1.5, 2.

Table of the risk in standard units for (a) using the smaller of the error variance and the prior variance for the correct distribution of the parameter; (b) the correct Bayes risk; (c) the risk if the wrong prior is used; (d) the risk if the procedure using the smaller of the error variance and the wrong prior variance is used. The approach to the asymptotic expression is apparent from these numbers.

Sample size	true state	true 0 or 1	true Bayes	wrong Bayes	wrong 0 or 1
1000	2^{-k}	10.953	9.468	24.701	31.000
1000	k^{-2}	62.746	49.174	99.003	114.169
10000	2^{-k}	14.221	12.788	78.748	100.000
10000	k^{-2}	199.502	156.580	705.093	753.404
100000	2^{-k}	17.526	16.110	250.108	316.000
100000	k^{-2}	631.956	496.229	5633.969	6074.754
1000000	2^{-k}	20.907	19.432	792.001	1000.000
1000000	k^{-2}	1999.500	1570.296	47281.971	51289.823

REFERENCES

CHEN, J. and RUBIN, H. [1986] Drawing a random sample from a density selected at random. *Computational Statistics and Data Analysis* **4**, 219-227.

COHEN, P. L. [1972] Bayes risk for the test of location – the infinite dimensional case. Unpublished dissertation, Purdue University.

GOLDSTEIN, M. and SMITH, A. F. M. [1974] Ridge-type estimators for regression analysis. *J. R. Statist. Soc.*, B,**36**, 284-291.

HOERL, A. E. and KENNARD, R. W. [1970] Ridge regression: biased estimation for non orthogonal problems. *Technometrics* **12**, 55-67.

RUBIN, H. [1974] Decision-theoretic approach to some multivariate problems. In *Multivariate Analysis II*, P. R. Krishnaiah (Ed.). Academic Press, New York.

RUBIN, H. [1977] Robust Bayesian estimation. In *Statistical Decision Theory and Related Topics II*, S. S. Gupta and D. H. Moore (Eds.). Academic Press, New York.