Ridge regression and inverse problems

ANDERS BJÖRKSTRÖM
Stockholm University, Sweden

January 18, 2001

Abstract

Why is ridge regression (RR) often a useful method even in cases where multiple linear regression (MLR) is dubious or inadequate as a model? We suggest that some light can be shed on this question if one notes that RR is an application of Tikhonov regularization (TR), a method that has been explored in the approximation theory literature for about as long as RR has been used in statistics. TR has proven useful for many inverse problems, but it has often been applied without stating a statistical model at all.

In order to indicate how alternatives to MLR might be defined, we give a subjective overview of some inverse problems from the geophysical sciences. We conclude that estimation is often at least as important as prediction.

Key words: inverse problems, Tikhonov regularization, partial least squares, principal components regression, regularized estimators, ridge regression

1 Introduction

1.1 Linear equation systems with random errors

Consider a system of linear equations:

$$A\theta = Y, \tag{1}$$

where A is an $n \times p$ matrix, θ is a p-vector and Y is an n-vector. We assume that data are available for A and Y, but subject to random errors. The purpose is inference about θ.

1.1.1 Example 1: The general linear model

Assume that the uncorrelated random variables Y_i, $i = 1, \ldots, n$ have equal variance σ^2 and expectations that are linear functions of p parameters θ_j, $j = 1, \ldots, p$:

$$\mathrm{E}[Y] = \sum_j A_{ij}\theta_j. \tag{2}$$

In this example, the numbers A_{ij} are assumed known without error. In addition to θ, the parameter σ is also of interest. The standard procedure is to estmate θ with the best approximative solution, in the least-squares sense, to the equation system (1). If the best approximation is not unique, a minimum-length requirement on θ is often added. If A^TA is nonsingular, the estimator is given by the normal equations, $\hat{\theta}_{LS} = (A^TA)^{-1}A^TY$.

The general linear model model includes multiple linear regression and analysis of variance as special cases. In regression, the matrix A is generally denoted X and contains the values of p "explanatory" variables at each of n observations. The purpose is to "explain" the variable Y by ascribing variation in Y to variations in the explanatory variables. This leads up to the equation system

$$\alpha + X\beta = Y, \tag{3}$$

where the intercept α is an n-vector with all elements equal. The p-vector of unknown parameters is here denoted β. It can be used, for example, for predicting the response variable at a new observation. An extension of the regression model is multivariate regression, where several response variables are considered simultaneously. One then needs to solve a matrix equation

$$XB = Y, \tag{4}$$

where B is $p \times q$ and Y is $n \times q$. In the present report we shall mainly confine ourselves to the case $q = 1$.

The general linear model is not adequate for situations where A is a random variable. Problems of this kind arise for example in regression when there are measurement errors in the explanatory variables.

1.1.2 Example 2: A material balance model

Consider a simple hydrological example. Let P denote the amount of rainfall in a region in a certain period of time, let E denote the evaporation, and R the runoff. We want estimates of these three numbers. If the period in question is a year, then to a good approximation $P = E + R$. Since R and E are difficult to measure, a

hydrologist may use data on some chemical compound, for example the chloride ion concentration in precipitation, c_P, and runoff, c_R to determine them. Unlike water, chloride does not evaporate, so if there is no net loss or accumulation of chloride, it must hold that $c_P P = c_R R$. Formally, we can write the balances for water and chloride in matrix-vector notation:

$$\begin{pmatrix} 1 & -1 & -1 \\ c_P & 0 & -c_R \end{pmatrix} \begin{pmatrix} P \\ E \\ R \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{5}$$

Here, not all the matrix elements are free from errors. Those in the first row are, being plus or minus one by construction, but those in the second row are not. Under hypothetical repetitions, research expeditions will observe different concentrations. We may regard the numbers in the second row as outcomes of random variables. Furthermore, the assumption that the right hand side is identically zero may also be questioned, for example if the data were collected during a short period of the year.

More elaborate models, involving many regions and many chemical compounds simultaneously, are important inbased on the same type of material balance considerations, are important in the geosciences. We return to the subject in Section 5.

1.1.3 Example 3: Latent variable models

Suppose one has collected data about p explanatory variables and q response variables at n occasions, yielding two matrices of data, X and Y. Suppose that there exist a unobserved "latent" variables $Z_1, \ldots, Z_a$, that influence the explanatory variables as well as the response variables, according to the linear equations

$$X_j = \sum_{\omega=1}^{a} Z_\omega p_{\omega j} + e_j, \tag{6}$$

$$Y_k = \sum_{\omega=1}^{a} Z_\omega q_{\omega k} + f_k, \tag{7}$$

where $p_{\omega j}$ and $q_{\omega k}$ are unknown parameters, and e_j and f_k are random errors. This is an example of a latent variable model. If equations (6) and (7) describe the actual relation between X and Y, then, unless the random term sare large, any column

vector of Y will be reasonably well approximated by some linear combinations of column vectors of X, so that $Y \approx XB$. Here, however, the usual multivariate regression model (equation 4) can lead to erroneous conclusions. The situation has been analyzed by Burnham et al (1999) who point out that it is not clear how the matrix parameter B should be be related to the parameters of equations (6) and (7) in such case.

1.2 Least squares and other methods

For the general linear model, Gauss-Markovs theorem states that the least-squares solution is the best linear unbiased estimator for θ. However, perhaps for this reason, θ_{LS} is often used quite uncritically as the "solution" to the problem $A\theta = Y$, although the randomness involved may be very different from what is assumed in the model definition. In general, one has no guarantee that the least-squares solution is an unbiased estimator of θ, or that θ_{LS} has minimal variance. A method that is often applied to the problem where A is affected by random errors is Total Least Squares (TLS), see van Huffel (1997). As an alternative to TLS, we focus on estimates of the form

$$\hat{\theta}_\delta = (A^T A + \delta I)^{-1} A^T Y \tag{8}$$

for this class of problems. We have two major reasons for our suggestion:

1) Statisticians have explored a technique called *ridge regression*, RR, which is equivalent to equation (8). The method has thus been explored, theoretically and practically, within the context of at least one well-defined statistical model.

2) Equation (8) defines a method known as *Tikhonov regularization*, which has been the subject of much research in the approximation theory.

We next make some further comments to these two points.

2 Near collinearities and ridge regression

As mentioned, the coefficient matrix A is usually denoted X in regression, and the unknown θ is denoted β. For uniformity we retain the notations A and θ here. In regression, the objective is to "explain" the variation in one or more "response

variables", by associating this variation with proportional variation in one or more "explanatory variables". A frequent obstacle is that several of the explanatory variables will vary in rather similar ways. As result, their collective power of explanation is considerably less than the sum of their individual powers. The phenomenon is known as *near collinearity*. When it occurs, the covariance matrix for $\hat{\theta}_{LS}$, which is $\sigma^2(A^TA)^{-1}$ will be almost singular, making $\hat{\theta}_{LS}$ highly sensitive to random variations in Y, that is, the estimate will depend very much on the particular way the errors ϵ_i happen to come out. Several methods have been developed in order to reduce this sensitivity, and one of them is *ridge regression* (RR), which means that a number δ is added to the elements on the diagonal of the matrix to be inverted, yielding a modified estimator of the form (8).

Ridge regressors are known to have favourable properties. For example, Hoerl & Kennard (1970) showed that $\hat{\theta}_\delta$ has smaller mean square error than the ordinary least-squares estimator, provided δ is small enough, and the standard regression model holds. A number of other properties were also pointed out early (Marquardt, 1970), for example, ridge regressors are "shrinkage regressors", i.e the Euclidean norm $|\hat{\theta}_\delta|$ is a decreasing function of δ. The RR estimators also turn out to play important roles in Bayesian models, provided θ has normal apriori distribution. It has also turned out that several other countermethods often taken against near-collinearity are closely related to ridge regression. Among such methods are continuum regression (Stone & Brooks 1990, Sundberg, 1993), partial least squares (PLS), principal component regression (PCR), and others (Björkström and Sundberg, 1999). These results make it meaningful to continue to explore the properties of ridge regression. In particular, Björkström and Sundberg (1999) have developed an upscaled form of RR, called "least squares ridge regression", (LSRR). The RR regressor is multiplied by a factor, chosen so that the residuals become orthogonal to the fitted values, hence the name "least squares".

Often, RR is used even though the MLR model is known to be wrong. In many cases, such applications are succesful. There is likely to be some reason for the success, some mechanism that can be found and stated more explicitly. Results of that kind should help users to foresee when a method will be fruitful and when it

will not be.

It seems relevant, thus, to explore whether the well-known theoretical advantages of RR, including its limiting cases, remain valid outside the rather strict set of assumptions within which it was originally developed. In statistical terms, we want to find out whether ridge regression is robust to the assumption of an error-free design matrix.

3 Ill-conditioned problems and Tikhonov regularization

We may say that a problem is ill-conditioned if the "known" input is so uncertain that the requested answer becomes unacceptably uncertain. Problems of this kind drew the attention of Hadamard (1902, 1932), who noted that there were problems in which infinitesimally small variations of initial or input data caused large variations in the solution. Hadamard declared that meaningful mathematical problems should have unique solutions that are stable with respect to small variations in input data. A readable survey of ill-conditioned problems is given in Allison (1979). In Allison's words, "the respect for Hadamard was so great that incorrectly posed problems were considered 'taboo' for generations of mathematicians, until comparatively recently it became clear that there are a number of quite meaningful problems, the so-called 'inverse problems', which are nearly always unstable with respect to fluctuations of input data".

More formally, suppose that the "requested" θ and the "given" Y are elements in Hilbert spaces H_1 and H_2 respectively, and let A be a given linear operator with domain $D(A) \subset H_1$ and range $R(A) \subset H_2$. We want to find a θ such that $A\theta = Y$. We say that Y is given, but normally Y is only known to within a margin of error. Suppose we know $Y \in U_2$, where U_2 is some "small" subset of H_2. Then, all we can conclude is $\theta \in U_1 = \{\theta; A\theta \in U_2\}$. If there is a large set of elements $\theta \in H_1$ such that $A\theta \approx 0$, and $|\theta|$ is not negligible, then the set U_1 may be so wide that learning $\theta \in U_1$ adds practically nothing to what we could have said about θ beforehand.

Another complication is that one is often working with an inexact operator A. It may be a matrix of experimentally determined coefficients, with errors to them,

or it may be an approximation of the true physical law connecting θ with Y; for example, we may use finite differences instead of derivatives. Both the "true" A and the approximation we use may lack inverses. If all we actually know about A is that it belongs to some set $\mathcal{A}$ of operators, then all we can say about U_1 is $U_1 = \{\theta; \exists A \in \mathcal{A}; A\theta \in U_2\}$.

It is an important objective in approximation theory to gain as certain knowledge about θ as possible in spite of all these obstacles. Tikhonov (1963) considered the quadratic functional

$$I(\theta, \delta^2) = |A\theta - Y|^2 + \delta^2|\theta|^2.$$

He demonstrated that to every $\delta^2 > 0$ there exists a unique element $\theta_\delta \in H_1$ for which $I(\theta, \delta^2)$ is minimal, (*i.e*, $\exists\ \theta_\delta; \theta \neq \theta_\delta \Rightarrow I(\theta_\delta, \delta^2) < I(\theta, \delta^2)$). It has been shown (Ivanov, 1976; Allison, 1979) that in many practical contexts, these elements perform well as substitutes for the non-existing "solution of $A\theta = Y$". The technique has become known as Tikhonov regularization.

In numerical applications, H_1 and H_2 are finite-dimensional. TR is then equivalent to resolving an ill-conditioned system of linear equations $A\theta = Y$ by using $\theta_\delta = (A^TA + \delta I)^{-1}A^TY$ as an approximate solution.

4 Parameter selection

Any user of TR or RR will need a rule for determination of the method parameter δ. In this section, we describe a method that is well known in regularization theory. The description is intended for statistical readers, and therefore we use the notations X and β for A and θ.

4.1 Ridge traces and L-curves

In ridge regression, the parameter is sometimes determined by inspecting the so-called *ridge trace*. This is a plot of the components of the RR predictor $\hat{\beta}_\delta$ versus δ. The normed residual sum of squares $|y - X\hat{\beta}_\delta|^2/(n-1)$ is also included. Typically, the ridge trace exhibits the features of Figure 1. (This figure is based on a simulated data set, used by Brown (1993, p. 57). As δ runs through a short interval I_0 beginning

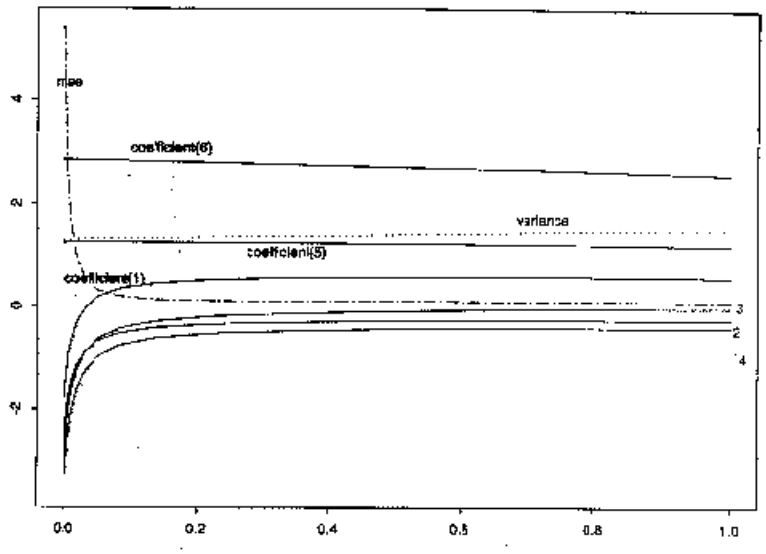

Figure 1: Ridge trace of six regression coefficients versus δ, with variance (from Brown,1993)

at $\delta = 0$ the predictor coefficients $\hat{\beta}_{\delta j}$ approach zero rapidly and may even change sign. To the right of the interval I_0, the predictor $\hat{\beta}_\delta$ stays fairly constant over a range of δ-values, $\delta \in I_1$. The residual sum of squares remains almost constant over the two intervals I_0 and I_1. In this situation, any value of δ in the interval I_1 will correspond to almost the same predictor, call it $\beta(I_1)$, and this predictor yields an error $|y - X\beta(I_1)|$ that is not much larger than the minimum achievable (i.e for OLS). Of course, as δ increases, the residuals will eventually reach an unacceptable size. There exists some more or less well-defined right endpoint of I_1. In Figure 1, I_1 might be $(0.2\,, 1.0)$. To use the predictor $\beta(I_1)$ instead of the OLS solution is a way to exploit the mean square error reduction mentioned in section 1. Hypothetically, if new data were collected, and a new Figure 1 were drawn, one would obtain a ridge trace, probably looking quite differently in the interval I_0, but being nearly the same as before in I_1. Obviously, the method is quite subjective (exactly where does I_1 begin and end?), and Brown (1993, p. 56) observes that the ridge-trace technique has been more or less abandoned in favour of objective estimates of the parameter.

The ridge-trace bears some resemblance to the so-called *L-curve*, used for de-

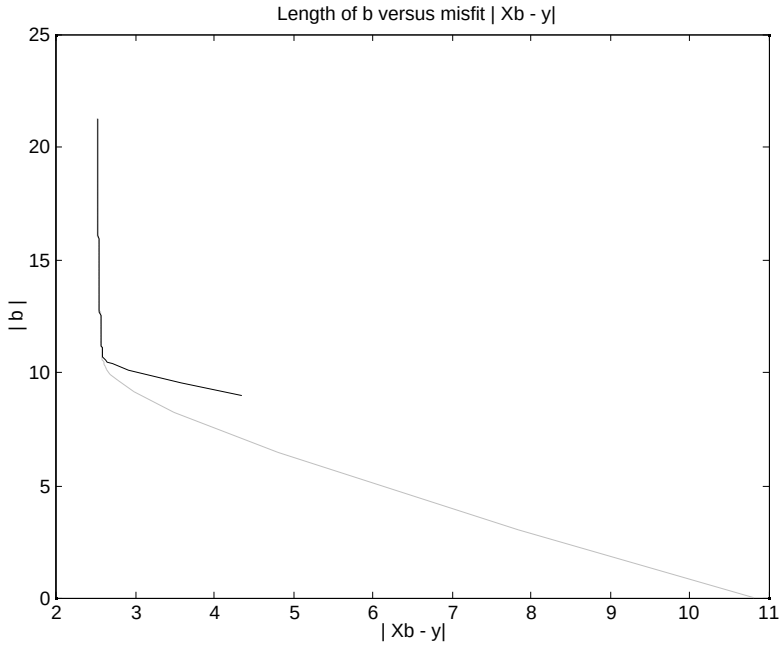

Figure 2: L-curves for Brown's (1993) data, for standard ridge regression (dotted) and with least-squares correction (solid).

termining the parameter in certain other applications of Tikhonov regularization. Instead of plotting all graphs $\beta_j(\delta)$ separately we plot only the Euclidean length of the vector $|\hat{\beta}_\delta|$, and the residual norm $|y - X\hat{\beta}_\delta|$. This information is arranged in a diagram with $|\hat{\beta}_\delta|$ along the ordinata and $|y - X\hat{\beta}_\delta|$ along the abscissa. Theoretically, an interval like I_0, where the coefficients $\beta_j(\delta)$ approach zero and the residuals only increase slightly, corresponds to an almost vertical line in such a plot. The interval I_1 should be compressed to one point, ideally. Values of δ larger than the right endpoint of I_1 would correspond to a segment where $|\hat{\beta}_\delta|$ keeps sinking, but where the most conspicuous effect is the growth of the residuals, which makes the trajectory appear almost horizontal. The combination of an almost vertical segment with an almost horizontal one is the reason for the name L-curves. (The name is ascribed to Lawson and Hanson (1974)).

As an example of an L-curve analysis of a data set from a regression context, we consider Figure 2, which is based on the same data as Figure 1. We see that while the image of I_0 is almost vertical (as could be expected), I_1 clearly is not depicted onto one point only. The coefficient estimates $\hat{\beta}_j$ are slightly dependent on δ, as is

the residual norm $|X\hat{\beta} - y|$, and this dependence stands out more clearly in Figure 2 than in Figure 1. We can also compare the L-curve for (ordinary) RR to the same graph for least-squares RR, (LSRR) (Section 2). For small δ, these two predictors are essentially the same, and their L-curves coincide. For parameters greater than about 0.05, LSRR has visibly better fit and less shrinkage than ordinary RR. The L-curve for LSRR therefore neither extends as far to the right, nor as far down, as the curve for ordinary RR. The "L" has a more distinct "corner".

Hansen (1992) discusses a number of rules for parameter selection in connection with the L-curve. Most of the principles he mentions lead, in practice, to points close to the "corner of the L". Consequently, Hansen and O'Leary (1993) suggest using the parameter that corresponds to the point of maximal curvature. In their experience, most of the other rules tend to yield solutions "to the right of the corner". In statistical terminology, other methods oversmooth at the expense of fit.

4.2 Cross-validation

An often-used method for statistical parameter determination is cross-validation, in practice "leave-one-out": Determine a regression model using all the data except one item, test it by predicting the left-out item; do this as many times as you have items, leaving out a different one each time, and calculate the sum of squares of the prediction errors, PRESS. By applying this procedure, different RR parameter values can be compared to each other, with respect to predictive ability. It often turns out that PRESS decreases as δ increases from zero, and has a minimum for some small positive δ. In order to compare results for different data sets, one sometimes uses a cross-validation index defined as $I_{CV} = 1 - \text{PRESS}/\text{PRESS}_0$, where PRESS_0 is the PRESS-value for the "null" predictor, $\hat{y} = \bar{y}$.

It is of some interest to explore how much the predictor defined by a minimum PRESS criterion differs from the one defined by maximizing the curvature of the L-curve, as described above. We have investigated a few data sets from this point of view. For Brown's data, Figure 3 shows that the best achievable I_{CV} is 0.7816 which occurs for $\delta = 0.17$. The corresponding LSRR predictor is shown in Table 1a. The parameter value corresponding to maximal curvature is smaller, $\delta = 0.05$.

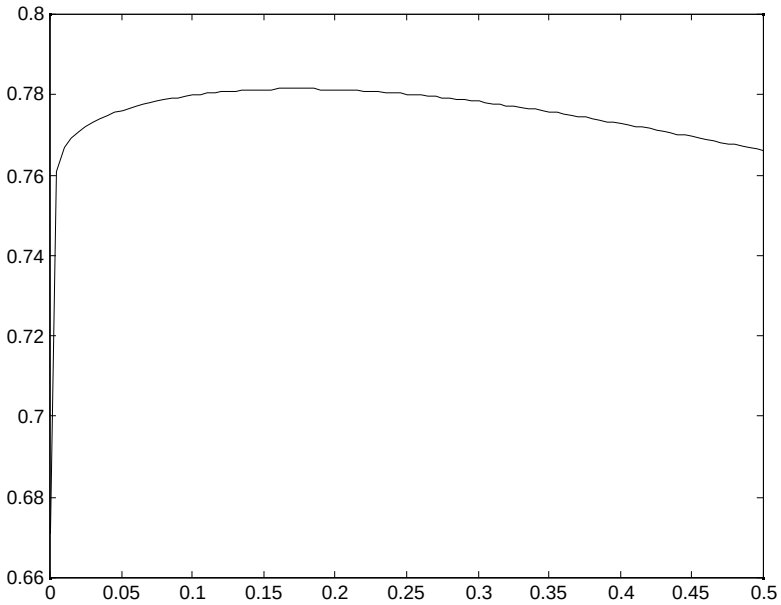

Figure 3: Cross-validation index for Brown's (1993) data, for least-squares ridge regression.

a) Brown's data	Best δ	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$
Cross validation	0.17	0.189	-0.088	0.014	-0.109	1.08	4.90
Maximal curvature	0.05	0.178	-0.092	0.018	-0.107	1.06	5.01
True value		0.603	0.301	0.060	-0.603	0.904	3.01

b) Fearn's data	Best δ	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$
Cross validation	0.067	0.021	-0.080	0.158	-0.231	0.0055	-0.0079
Maximal curvature	0.03	0.049	-0.072	-0.150	-0.250	0.0073	-0.0049

c) Hald's data	Best δ	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
Cross validation	0.025	1.27	0.29	-0.18	-0.36
Maximal curvature	0.006	1.35	0.32	-0.10	-0.33

Table 1: Coefficients in the optimal LSRR predictor, as defined by leave-one-out cross-validation, and by the principle of maximal curvature.

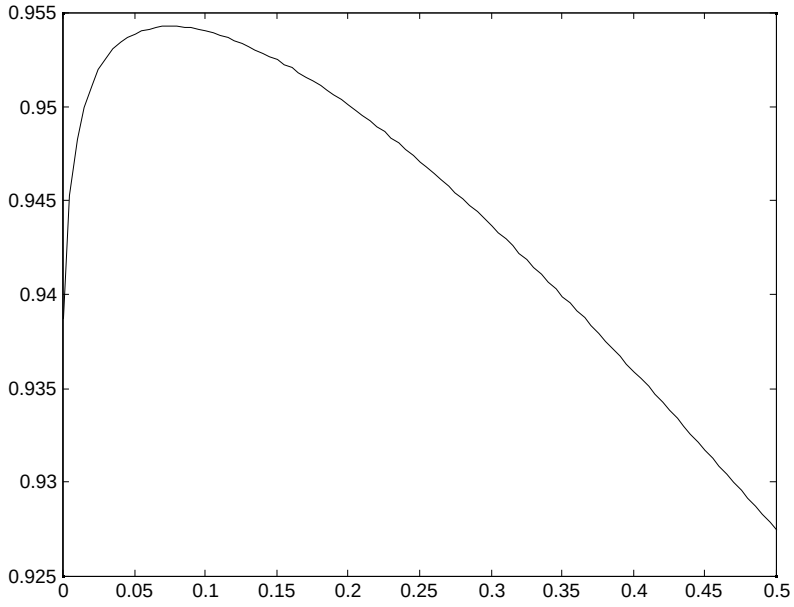

Figure 4: Cross-validation index for subset of Fearn's (1983) data, for least-squares ridge regression.

This LSRR predictor is also shown. The effect of changing δ is in agreement with Figure 1. For these constructed data we know the "truth", which is also given in the table. We see that both methods fail to catch $\beta_1, \ldots, \beta_4$ by at least a factor of three. This, obviously, reflects the near-collinearity arranged by Brown for the first four explanatory variables. As for β_5 and β_6, both methods overestimate them, slightly for β_5, more pronounced for β_6.

Figure 4, which was taken from Björkström and Sundberg (1999) shows cross-validation index as a function of δ when applying LSRR to a subset of Fearn's (1983) data from a NIR analysis of wheat samples (see Stone and Brooks (1990) for details). The figure shows that the best parameter choice is $\delta = 0.067$. Figure 5 shows the L-curves for this data. We locate the point of maximal curvature (by eye) and read off $\delta \approx 0.03$. The two predictors are given in Table 1b. We have no true value to compare with, but the two estimates are in broad agreement.

Our third data set is Hald's (1952) data on evolution of heat in cement. From the point of view of cross-validation, the best parameter is $\delta = 0.025$ (Figure 6). The L-curve in Figure 7 has no easily spotted point where the curvature is largest,

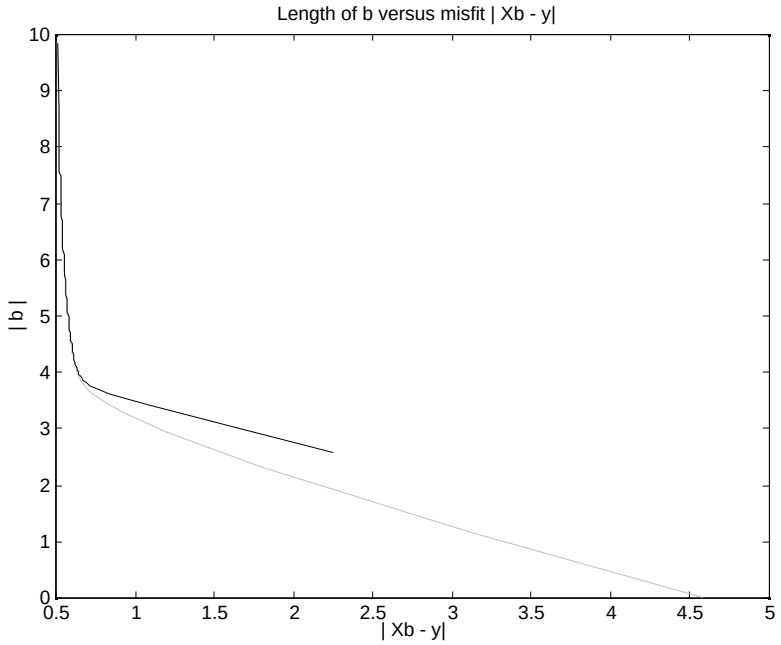

Figure 5: L-curves for subset of Fearn's (1983) data, for standard ridge regression (dotted) and with least-squares correction (solid).

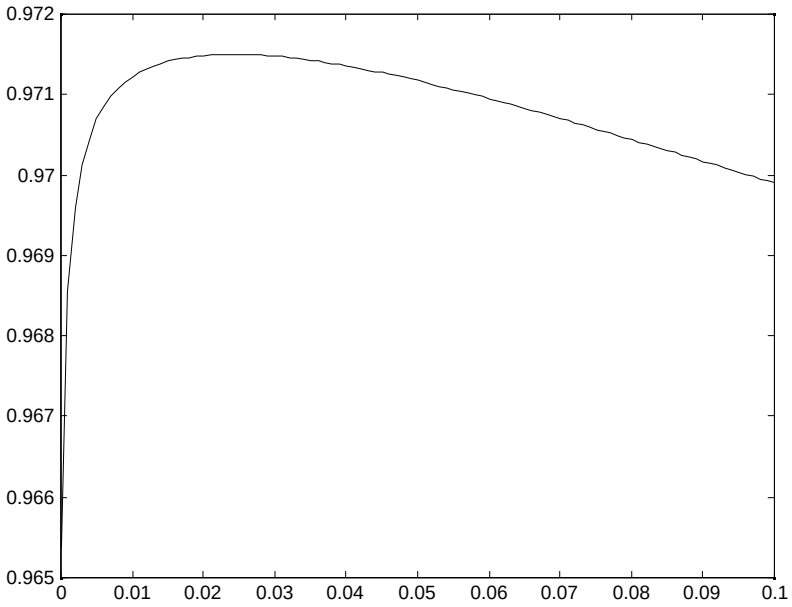

Figure 6: Cross-validation index for Hald's (1952) data, for least-squares ridge regression.

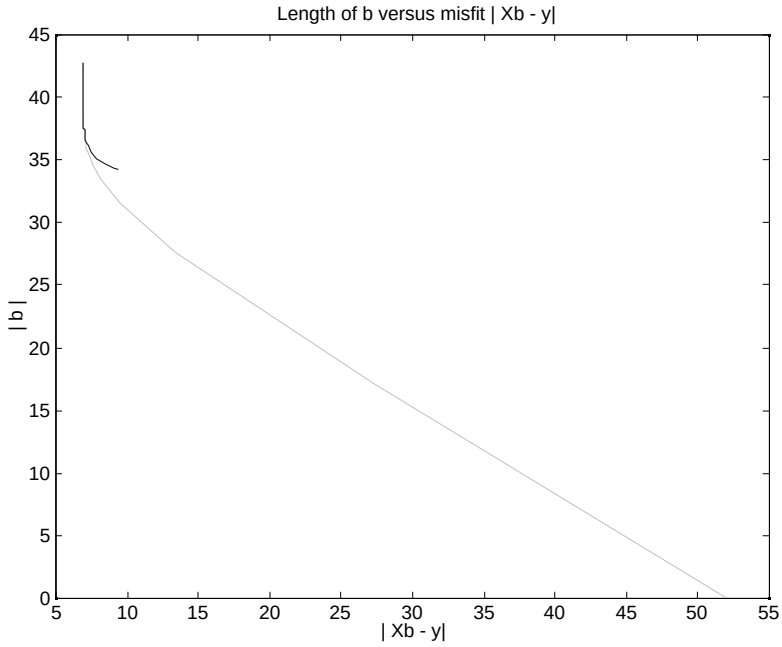

Figure 7: L-curves for Hald's (1952) data, for standard ridge regression (dotted) and with least-squares correction (solid).

but a maximum seems to occur close to $\delta = 0.006$. The two estimators differ only marginally, as Table 1c shows.

To summarize, in two out of our three regression data sets, the L-curve does exhibit if not a unique corner at least a discernible "corner region", where it bends from almost vertical to almost horizontal. The optimal solution, from the cross-validation aspect, occurs somewhat to the right of the corner region. A possible interpretation is that cross-validation overfits. Our observations are thus in agreement with the above-mentioned experince of Hansen and O'Leary (1993). However, we have found an exception to this. Figure 8 shows the L-curve for a data set on nitrate in wastewater (Karlsson et al, 1995), used as example in an overview paper by Sundberg (1999). For these data, Figure 6b in Sundberg's paper shows that the best LSRR parameter value, from the PRESS point of view, is of the order 10^{-3}. This is at least an order of magnitude *smaller* than the values in the corner region of Figure 8. (The corner extends approximately from $\delta = 0.02$ to $\delta = 0.4$.)

Above the corner, the L-curve for LSRR coincides with that for RR, for all the data sets we have seen, and to the right of the region, the curve for LSRR is above

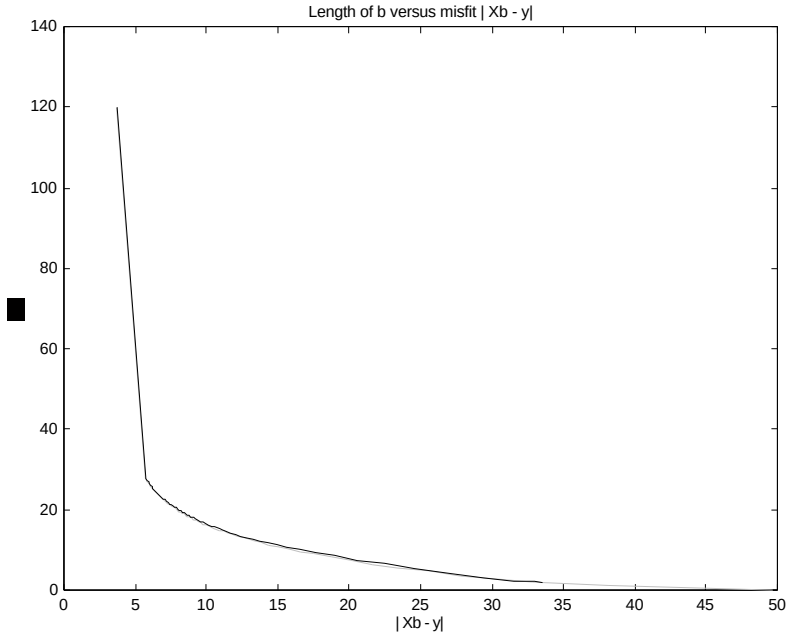

Figure 8: L-curves for the wastewater data by Karlsson et al. (1995) data, for standard ridge regression (dotted) and with least-squares correction (solid).

the one for RR. Thus, in order to achieve a given "fit" (a given size of $|X\hat{\beta} - Y|$), LSRR needs not shrink as strongly as RR. This illustrates that LSRR avoids the part of the shrinkage that has nothing to do with near collinearities.

In the diagrams, the parameter δ runs through the interval $(0\,, \infty)$. Therefore, the right endpoint of the RR curve corresponds to $|\hat{\beta}| = 0$, yielding a residual norm equal to $|Y|$. The curve for LSRR ends in the first-factor PLS estimator for β. We may note that the LSRR curve for Fearn's data (Figure 5) extends further to the right than for the other two data sets (Figures 2 and 7). This means that one-factor PLS does not manage to explain more than a small percentage of the total variation in the response variable for Fearn's data. The same phenomenon is visible for the wastewater data (Figures 8). In contrast, the LSRR curve is very short for Hald's data. This is natural in light of the fact that there are only four explanatory variables in Hald's case, but one hundred in the wastewater example.

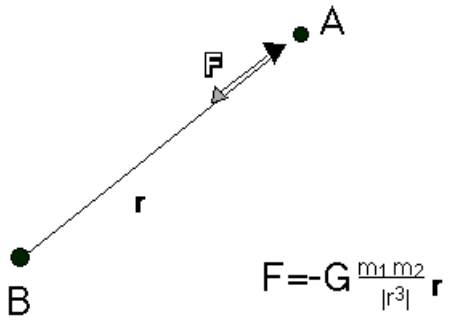

Figure 9: Illustration of the law of gravity.

5 Inverse problems

It is a paradoxical fact that in much research, scientists apply the laws of Nature coversely, logically speaking, to the way the principles can actually be used. Take the law of gravity as example. It states that whenever an object A is at a position $\mathbf{r}$ relative to another object B, then object A will be affected by a force $\mathbf{F}$ that depends on $\mathbf{r}$ and the masses of A and B. The law provides an explicit formula for $\mathbf{F}$ (Figure 9), so that if we know, for, example, the density at each point inside the Earth, we can calculate the force exerted on a kilogram of mass located at the surface. This calculation (a three-dimensional integration) is a "direct" problem, a straightforward exercise that has one and only one correct answer. A more interesting and important question, though, is the inverse problem: What can we say about the distribution of mass within the Earth, given knowledge about the law of gravity and data about the force as observed at a limited number of places? This problem has obvious economic relevance: density anomalies may indicate oil or valuable ores. However, it has no unique answer. A point mass, located at the centre of the Earth, would cause exactly the same force, at sea level, as does an homogeneous three-dimensional distribution of mass.

Another geological example of an inverse problem is to reconstruct the history of the temperature in the Earth's interior. There are data on its present distribution. The relevant Law, in this case, is the heat conduction equation for a sphere. It is

a straightforward task to integrate this equation forward in time, but what can we conclude about past times, given the present? More generally, the same logically "inverse" situation occurs whenever one wants to reconstruct the causes of a phenomenon, given knowledge about its effects. It is just as much present in disciplines where no "hard" mathematical model is available for the "forward" problem, for example when a palaeontologist is deducing the evolution of life on Earth.

We may formalize things this way: A "Law" says that if variable X takes on the value x, then variable Y takes the value $f(x)$. The Law may or may not provide an expression for the function f. We have observed $Y = y$; what can we say about X? (Here, X as well as Y can be many variables simultaneously. We should regard them as vectors in spaces with many dimensions, perhaps infinitely many).

An inverse problem never has a unique answer. The available data can always be explained in more than one way, many different x:s give $f(x)$ equally close to the observed y. A deterministic scientist, not caring much about a statistic model, is likely to accept a least-squares approximation and proceed by putting most belief in the simplest x possible. Simplicity may mean that many of the coordinates x_j are equal to zero, or that $|x|$ is small. It may also mean that there is some pattern to the x_j, so that, for example, x_j falls between x_{j-1} and x_{j+1}. In many cases, the simplicity criterion can be expressed as a wish that $|Lx|$ be small, where L is some suitably chosen matrix. If the determinist accepts to approximate the true $f(x)$ by a linear expression, ($f(x) \approx Ax$, for some linear operator A), one easily sees that he or she faces a the tradeoff between, on the one hand, explaining the available data (*i.e* matching Ax to y) and, on the other hand, satisfying some more or less subjective opinion about simplicity. This leads up to minimizing some weighted average $|Ax - y|^2 + \delta|Lx|^2$, a situation typical for Tikhonov regularization.

Furthermore, if we admit that data are affected by random perturbations, we realize that our observed y is not the true value of Y. Guided by a relevant statistical model, we may compute a confidence region for the true Y. However, to transform this region into a confidence region for the true X may not be rewarding if A is close to singular. It may be better to resort to Tikhonov regularization and consider

confidence regions of the form

$$\{x\,;|Ax-y|^2+\delta|x|^2\leq k\}$$

.

where k depends on the degree of confidence wanted. For positive δ, these sets will be more rounded than the very elliptic shape taken on for $\lambda=0$.

From a statistical perspective, it seems that inverse problems would be well suited for Bayesian analysis. Consider X as a vector of parameters describing the unknown state of nature. The subjective preferences correspond naturally to an a priori distribution for X. The conditional distribution for Y given X can be written down if one knows the physical law that connects X with Y and one has a model for the random effects at work. It is then in principle straightforward to determine the aposteriori distribution for X. Several authors have taken this approach. An example is given is Section 5.3.

In connection with the Bayesian approach, we may note that the ridge regressor has a special interpretation here. If we assume, in model (**??**), that the intercept has a vague a priori distribution on all of the real line, if the conditional distribution of β given σ is N(0, $\sigma^2 I_p/\delta$), then it turns out that the Bayes estimator for β is the ridge estimator $\hat{\beta}_\delta$, which is an unbiased estimate of the expected value of the a posteriori distribution, since β is normally distributed a posteriori. If there is some way to estimate the (common) variance ${\sigma_\beta}^2$ of the components of β, say $\hat{\sigma_\beta}^2$, then a good rule for the parameter selection might be $\delta=\hat{\sigma}^2/\hat{\sigma_\beta}^2$.

A number of inverse problems are described next. The reader is referred to Allison (1979) and Keller (1976) for more extensive reviews.

5.1 An example from the interior of the Earth

In general, inverse problems take the form of integral equations. In the gravitation problem described above, the force of gravitation $\mathbf{F}_0$ at a given point at sea level is

$$\mathbf{F}_0=\int_V \frac{\gamma}{|\mathbf{x}-\mathbf{x_0}|^3}(\mathbf{x}-\mathbf{x_0})\rho(\mathbf{x})dV, \tag{9}$$

where γ is the universal constant of gravitation, and the integral is a volume integral over the Earth. The density at location $\mathbf{x}$ is denoted $\rho(\mathbf{x})$, and is the unknown

function to be estimated. Of course, one has only a finite amount of information about $\mathbf{F}_0$, which limits the degree of detail in which one can resolve the function $\rho(\mathbf{x})$. If we compare equation (9) to our general linear equation system (1), the force $\mathbf{F}_0$ corresponds to Y. The mass elements $\rho(\mathbf{x})$dV make up θ. The expression $\frac{\gamma}{|\mathbf{x}-\mathbf{x_0}|^3}(\mathbf{x}-\mathbf{x_0})$ corresponds to A_{ij}. In numerical applications, the integral in equation (9) is approximated by a sum. This can be done in more than one way, though. One possible discretization is to think of the Earth as made up of a large number of small cubes, each with constant density. Another option is to model it as a large but finite number of concentric shells, each with constant density. The linear equation systems will turn out differently in the two cases. Both models can be equally physically reasonable, but not equally suitable for data analysis. It is important to remember the continuous character of the underlying problem.

In a series of papers around 1970, Backus & Gilbert (1967, 1968, 1970) developed a number of models for the interpretation of geological and seismological data. Some of these models are one-dimensional in the sense that properties only vary with distance from the Earth's centre, not depending on latitude or longitude. Let $\rho(r)$ be the Earth's density at distance r from the entre, and suppose that this is an unknown function that we want to determine. We realize that the total mass of the Earth can be expressed in terms of $\rho(r)$ as

$$\int_0^{R_0} \rho(r)\, 4\pi r^2 dr, \tag{10}$$

where R_0 is the radius of the Earth. If we possess a numerical estimate of the mass of the Earth, say $\hat{M}_e$, we can write down an equation with (10) as left hand side and $\hat{M}_e$ as right hand side, which provides some information about $\rho(r)$. Further, the laws of mechanics enable us to write the Earth's moment of inertia in terms of $\rho(r)$ as

$$\int_0^{R_0} \rho(r)\, 4\pi r^3 dr, \tag{11}$$

so if we have a numerical estimate of its value, $\hat{J}_e$, we obtain more information about $\rho(r)$. Backus and Gilbert (1967) introduce two additional unknown functions, the bulk modulus $\kappa(r)$ and the shear modulus $\mu(r)$. The point is that these two functions (with the density) affect a number of observable properties of the Earth, for example

the travel times for seismic waves between two places. For each of these properties we can derive a "theoretical" expression of the form

$$\int_0^{R_0} (G_1(r)\rho(r) + G_2(r)\kappa(r) + G_3(r)\mu(r))dr, \tag{12}$$

where $G_1(r)$, $G_2(r)$ and $G_3(r)$ are known functions of r (often, but not always powers r^n). For example, when the property we observe is moment of intertia we have $G_1(r) = 4\pi r^3$, $G_2(r) = G_3(r) = 0$.

In their 1968 paper, Backus & Gilbert generalize from 3 to N unknown functions of r. (Consequently, they need N functions $G_i(r)$, $i = 1, \ldots, N$). They attempt to find linear combinations of $G_i(r)$ that are as similar as possible to Dirac spikes, $\delta(r - r_0)$. The reason for this is that if $\sum a_i G_i(r) = \delta(r - r_0)$, then the number $\sum a_i \gamma_i$ would be a good estimator of $m(r_0)$. They assume the covariances $E_{ij} = \mathrm{Cov}(\epsilon_i, \epsilon_j)$ to be known. They face a conflicting minimization problem: Find numbers $a_1, \ldots, a_N$ such that, on one hand, the function $\sum a_i G_i(r)$ is as Dirac-like as possible, while, on the other hand, the quadratic form $\sum a_i a_j E_{ij}$ is kept small, this number being the variance of the estimators. The conflicting minimization that they encounter is similar to the tradeoff between bias and variance in ill-conditioned regression. Backus & Gilbert do not explicitly use the term Tikhonov regularization, but their geometrical construction and their use of Lagrangian multipliers is equivalent, in practice.

5.2 Inverse problems in oceanography

A fundamental objective in oceanography is to determine the motion patterns in the sea. A particularly difficult problem here is to resolve the vertical part of the circulation. The motion of water up and down is generally slow, in most places not more than one or a few metres per year, and there are no instruments that can provide direct measurements of it. Instead, estimates are derived from indirect sources. For example, if one observes that currents at the surface bring more water into a certain region than out of it, one can conclude that downward motion must be going in that region. Similarly, one can formulate mass balance equations for salt and other compounds dissolved in the water in a specified region, and derive

additional information about the water motions. Aside from its intrinsic interest, knowledge about the vertical circulation is crucial for a correct assessment of the risks with increasing carbon dioxide emissions.

An important source of information in this context is the "apparent age" of seawater sampled at various locations. This is defined as the age that an archaeologist would assign to the water, based on the abundance of radioactive carbon. (Seawater contains carbon, in round numbers 2 moles per m^3, mainly as bicarbonate ion). However, the interpretation of the apparent age is not straightforward, since any sample of water is in fact a mixture of water of different ages, having travelled along different paths before the sampling. This problem is best resolved by noting that radioactive carbon is not the only substance that carries information about oceanic motion. Concentration profiles for compounds necessary for marine life also contain information that can be interpreted in terms of water circulation. Substances of this type are, typically, less abundant near the surface, where they are continuously "eaten" by living organisms, than in the deep sea, where the debris of dead organisms dissolves. Oceanographers have tried to deduce information about the biological production and the motion patterns simultaneously, by treating both types of material fluxes as unknowns and formulating systems of mass balance equations. The forward problem, in this example, consists of calculating how a substance will spread, knowing the circulation and diffusion. If we put one unit of ink at a given position in the sea, it will gradually become a more and more diffuse "cloud", because of random mixing. This proceeds much like heat conduction and can be described by a similar equation. Simultaneously, the cloud as a whole will be transported and deformed because of the large-scale circulation systems. If we know the velocities, it is straightforward to calculate this effect also. (See Appendix A). The forward problem is thus in principle solved, but our problem is the inverse one: We have data on concentrations of various compounds, and wish to determine the velocities and diffusion coefficients. Several authors have treated this inverse problem. Their work has dealt with all of the ocean (Bolin et al, 1983) as well as regions like the Indian Ocean (Metzl et al, 1990), the North Atlantic (Bolin et al, 1987), or the Weddel Sea (Lindegren and Josefson, 1998).

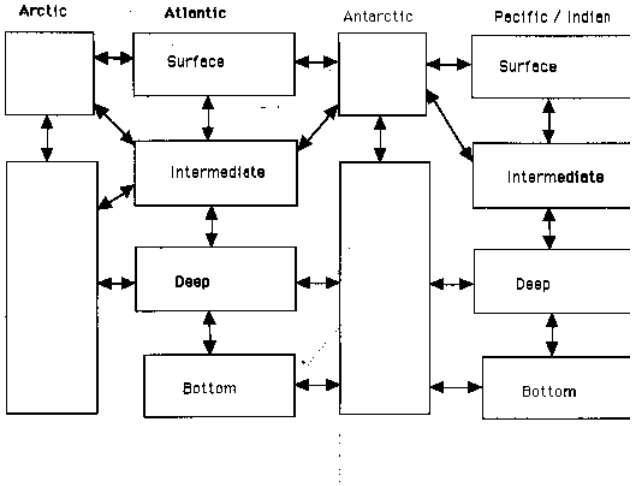

Figure 10: A 12-box model of the ocean, from Bolin et al (1983). The four columns represent, from left to right: The Arctic Ocean, the Atlantic Ocean, The Antarcic Ocean and the Pacific/Indian Oceans.

5.2.1 A case study

We consider one of the studies in more detail. Figure 10 illustrates a model to extract information from data on radioactive carbon and other dissolved compounds in the water. The complexity of the figure represents the state of the art some fifteen years ago (Bolin et al., 1983), but it can serve as example here. The ocean is modelled as twelve compartments, bordering each other at twenty boundaries. The circulation is described by specifying two numbers at each boundary: the net flow of water and a coefficient for "diffusion". Water that leaves a compartment is assumed to carry dissolved substances in concentrations characteristic for the compartment it leaves. A balance equation can be formulated for each substance and each compartment considered, and , assuming that the state of the ocean as we see it today represents an equilibrium, we can set all net balances equal to zero and obtain a linear equation system. Formulating balance equations for water, radioactive carbon, and three more substances, Bolin et al arrived at a system of 80 equations in 56 unknowns. Estimates obtained were in general agreement with conventional knowledge. Although the main purpose with the work was to

study the CO_2 uptake, the estimates of the circulation rates of course attracted interest. The authors were able to reproduce the already-known gross features of the oceanic circulation correctly and also came up with estimates of the circulation in the so-called intermediate ocean (down to about 1000 m below sea level). They also concluded that their model was too simple to be adequate for CO_2 uptake studies. An obvious shortcoming with the model was the fact that some of the parameters were clearly wrong, for example, turbulent diffusion coefficients became negative at some places (which means that heat would flow from a colder object to a warmer one). The only attempt the authors made to analyze the uncertainty of their reults was to add random perturbations to their input data and compute ten alternative solutions. For each of the unknowns, its standard deviation (over this sample of ten) was computed and used as a measure of the uncertainty.

A better investigation of the uncertainty of this particular model was carried out by Mansbridge & Enting (1986). They applied a number of regularization techniques, among them Tikhonov regularization with various principles for selecting the parameter. In particular, the generalized method suggested by Goldstein & Smith (1974) was employed, and also methods similar to PCR. Having observed that $\hat{\beta}_{\delta j}$ often changes sign for small values of δ (corresponding to the interval I_0 in Section 4.1), they had hoped to correct the sign errors for the diffusion coefficients. Unfortunately, they did not manage to come to grips with this particular fault.
Neither group of authors attempted to formulate a specific model for the errors. Obviously, the right-hand side contains errors, but a regression model is not adequate, since the most important data are measurements of concentrations of substances, and these enter in the coefficients A_{ij}. Actually, the structure is that we have a number of measurements of various compounds at various locations in the sea, denote them c_1, c_2 ..., and the structure is $A_{ij} = \sum_k \tau_{ijk} c_k$. where the "design tensor" τ_{ijk} is a sparse arrangement of $+1$:s and -1:s. It may not be impossible to follow up the consequences of assuming all c_k to be independent, unbiased and normally distributed. We do not plan to analyze this model further, however, since more detailed models have superseded it nowadays. We note that although the various regularization schemes gave quite differing results, none of them was really satisfac-

tory from an oceanographic point of view. It is probably correct to ascribe this to something fundamentally wrong with the assumptions. For example, turbulent mixing on scales as these may not actually lend itself to modelling by a heat conduction equation.

In later years, tracer-based inverse models have been replaced by models based on dynamical considerations, analyzing the motions from first principles, such as the laws of motion. However, these models also lead to a large number of inverse problems, as described in the monograph by Bennett (1992).

5.3 A Bayesian example from the atmosphere

The material balance approach described in the previous example is also being applied in atmospheric chemistry. Then, it is not primarily the motion patterns that are unknown. (Weather observers are measuring the speed and direction of the wind continuously). Instead, the aim is to determine the sources and sinks (deposition patterns) for various substances present in the air. Both the location and the magnitude are poorly known (where does this compound enter the atmosphere, how much mass comes in per unit time) and the objective is to gain knowledge about these numbers from measurements of concentrations in the air at various times and places. The direct problem is again to solve a transport equation, *i.e* a system of partial differential equations. In this case, the desired solution has a known gradient at the boundary of the domain (the sources and sinks are at ground level). An interesting paper by Kandlikar (1997) discusses the methane cycle from a Bayesian perspective. Methane (CH_4) is emitted into the atmosphere via a number of processes, natural as well as man-made ones. The most important contributions are from biological decay in wetland ecosystems, from certain mammals (ruminants), and underwater rice fields. In recent years, leakage from handling of natural gas has become a considerable source. Methane is removed from the air primarily by oxidation to carbon dioxide.

If $C(t)$ denotes the total amount of methane in the atmosphere at time t, we can

write a simple balance equation as

$$\frac{dC}{dt} = \sum_i Q_i - \sum_j S_j \tag{13}$$

where Q_i are the various sources (inflows) and S_j are the various sinks (outflows). Scientists representing different subjects have estimated the different source and sink terms, with their individual limits of uncertainty. There are also observations of the local rate of change of methane concentrations at a number of places around the Earth. These numbers can be interpreted in terms of dC/dt, but one is of course introducing some error in that extrapolation. Kandlikar (1997) gathers estimates from diverse references, and translates this information into a priori distributions for the terms on the right hand side of (13). He simulates from these distributions and uses equation (13) to obtain a prior distribution for dC/dt. Comparing this result to the observations of the rate of change, and using Bayes' formula, he is able to deduce updated a posteriori distributions for the source and sink terms. The method permits him to reduce the uncertainties about the fluxes with up to 30 %, in some cases.

Kandlikar's method is extremely simple in that he treats all of the atmosphere as one single reservoir. Thus he is unable, for example, to say where on Earth the different source and sink processes go on. More elaborate models have been developed by for example Hartley and Prinn (1993) who use a Bayesian approach to study the geographical distribution of the emissions.

6 Conclusions

We summarize the previous sections in a few points.

1. The standard multiple regression (MLR) model can be seen as a special case of a more general situation where one wants to solve an equation system $Ax = b$, where A and b are subject to random errors. Methods designed for MLR sometimes perform well even when the correct model is another one. We do not know why this is the case, and therefore we are not able to foresee when a method will be adequate for a given data set.

2. In many applications of regression methods, the solution β is not the result of primary interest. Instead, the important purpose has been to be able develop predictors of new values for the response variabel, given new cases, where only the explanatory variables are observed. Criteria for method evaluation often reflect the prediction aspects. However, as we have seen, there are many situations where the components of x have important scientific interpretation. It seems meaningful to formulate and explore evaluation criteria based on the difference between the estimated solution and the "true" x.

3. In ridge regression, several principles are known for selecting the best parameter value. To the extent the consequences of these principles have been explored, it has been within the framework of the standard regression model. Little seems to be known about their performance under other models. Statisticians can probably learn interesting methods for ridge parameter selection by getting acquainted with of the literature about Tikhonov regularization.

4. Inverse problems is an area that seems potentially suitable for Bayesian analysis. Several scientists have taken that approach. However, a) some people are reluctant to translate their uncertainties into terms of probability distributions and b) it is important to be clear about what is known beforehand and what is based on data.

5. In the above examples, as in most applications in environmetrics, data have not arisen as the result of controlled experiments. Rather, all the variables involved should be considered stochastic, whether they contribute to the right or left hand side of the linear equation system. Drawing a parallel to regression, we may say that the situation resembles the "natural calibration" case, and it is natural to assume that explanatory and response variables have a simultaneous distribution. This is a hint that appropriate models might be resemblant of for example canonical coordinate regression, or latent variable models. However, we also see, particularly in the oceanic and atmopheric examples, that the division into explanatory and response variables seems unnatural. The situation is not similar to regression models at all.

REFERENCES

Allison, H. (1979). Inverse unstable problems and some of their applications. *Math. Scientist* **4**, 9-30.

Backus, G. & Gilbert, F. (1967). Numerical applications of a formalism for geophysical inverse problems. *Geophys. J. R. Astr. Soc* **13**, 247-276.

Backus, G. & Gilbert, F. (1968). The resolving power of gross Earth data. *Geophys. J. R. Astr. Soc* **16**, 169-205.

Backus, G. & Gilbert, F. (1970). Uniqueness in the inversion of inaccurate gross Earth data. *Royal Soc. London, Philosophical Transactions, A***266** , 123-192.

Bennett, A. (1992). Inverse methods in physical oceanography. Cambridge Univ. Press.

Berkson, J. (1950). Are there two regressions? *J. Am. Statist. Ass.* **45**, 164-180.

Björkström, A. & Sundberg, R. (1996). Continuum regression is not always continuous. *J. Roy. Statist. Soc. Ser. B* **58**, 703-710.

Björkström, A. & Sundberg, R. (1998). A generalized view on continuum regression. *Scand. J. Statist* **25**, 17-30.

Bolin, B., Björkström, A., Holmén, K. & Moore, B. (1983). The simultaneous use of tracers for ocean circulation studies. *Tellus* **35 B**, 206-236.

Bolin, B., Björkström, A., Holmén, K. & Moore, B. (1987). On inverse methods for combining chemical and physical oceanographic data: A steady-state analysis of the Atlantic Ocean. Report CM-71, Dept. of Meteorology, Stockholm University.

Brown, P. J. (1993). *Measurement, Regression, and Calibration.* Oxford Univ. Press, Oxford.

Burnham, A. J., MacGregor, J. F. & Viveros, R. (1999). Interpretation of regression coefficients under a latent variable regression model. (Manuscript)

Fearn, T. (1983). A misuse of ridge regression in the calibration of a near infrared reflectance instrument. *J. Appl. Statist.* **32**, 73-79.

Goldstein, M. & Smith, A. F. M. (1974). Ridge-type estimators for regression anal-

ysis. *J. Roy. Statist. Soc. Ser. B* **36**, 284-291.

Hadamard, J. (1902). Sur les problèmes aux dérivées partielles et leur signification physique. *Bull. Univ. Princeton*, **13**, 49-52

Hadamard, J. (1932). Le problème de Cauchy et les équations aux dérivées partielles linéaires hyperboliques. Hermann, Paris.

Hald, A. (1952) Statistical theory with engineering applications. Wiley, New York.

Hansen, P.C. (1992). Analysis of ill-posed problems by means of the L-curve. *SIAM Review* **34**, 561-580.

Hansen, P.C. and D. P. O'Leary (1993). The use of the L-curve in the regularization of discrete ill-posed problems. *SIAM J. Sci. Comput.* **14**, 1487-1503.

Hartley, D. & Prinn, R. (1993). Feasibility of determining surface emissions of trace gases using an inverse method in a three-dimensional chemical transport model. *J. Geophys. Res.* **98**, 5183-5197.

Hoerl, A.E. & Kennard, R.W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**, 55-67.

Ivanov, V.V. (1976) The theory of approximate methods. Noordhoff International Publishing, Leyden, The Netherlands.

Karlsson, M., Karlberg, B. & Olsson, R. J. O. (1995) Determination of nitrate in municipal waste water by UV spectroscopy. Anal. Chim. Acta **312**, 107-113.

Kandlikar, M. (1997). Bayesian inversion for reconciling uncertainties in global mass balances. *Tellus* **49 B**, 123-135.

Keller, J. (1976) Inverse problems. *Amer. Math. Mon.* **83**, 107-118.

Lawson, C. L. & R. J. Hanson (1974) Solving least squares problems. Englewood Cliffs, N.J.: Prentice-Hall Inc.

Lindegren, R. & M. Josefson (1998) Bottom water formation in the Weddell Sea resolved by principal component analysis and target estimation. *Chemolab.* **44**, 403-409.

Mansbridge, J. V. & Enting, I. G. (1986). A study of linear inversion schemes for

an ocean tracer model. *Tellus* **38 B**, 11-26.

Marquardt (1970). Generalised inverses, ridge regression, biased linear estimation and nonlinear estimation. *Technometrics* **12**, 591-612.

Metzl, N., B. Moore & A. Poisson (1990). Resolving the intermediate and deep advective flows in the Indian Ocean by using temperature, salinity, oxygen and phosphate data: the interplay of biogeochemical and geophysical tracers. *J. Geophys. Res* **89**, 81-111.

Pasquill, F. & F. B. Smith (1983) Atmospheric Diffusion: The dispersion of windborne material from industrial and other sources. Ellis Horwood Ltd, Chichester. Third edition.

Stone, M. & Brooks, R. J. (1990). Continuum regression: Cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. (With discussion) *J. R. Statist. Soc.* B **52**, 237-269; Corrigendum (1992). **54**, 906-907.

Sundberg, R. (1993). Continuum regression and ridge regression. *J. R. Statist. Soc.* B **55**, 653-659.

Sundberg, R. (1999). Multivariate calibration – Direct and indirect regression methodology (with discussion) *Scand. J. Statist* **26**, 161-207.

Tikhonov, A.N. (1963) Dokl. Akad. Nauk SSSR, 153(1963) 49-52, MR 28# 5577.

van Huffel, S. (Ed.) (1997). Recent advances in total least squares techniques and errors-in-variables modeling. *Proc. of the Second International Workshop on Total Least Squares and Errors-in-Variable Modeling*, SIAM, Philadelphia.

Appendices

A The general transport equation

Consider a substance that is dissolved in the atmosphere or in the ocean. Suppose that the substance is being transported with the motions of the air or the water, without itself influencing these motions. The concentration c (mass units per unit volume) will be a function time and three spatial coordinates, $c = c(x, y, z, t)$. We now derive a partial differential equation for c. The flux of the compound, i.e, the amount transported through a unit area per unit time, in made up by two processes, $\mathbf{F} = \mathbf{F_a} + \mathbf{F_d}$. The "advective" flux $\mathbf{F_a}$ is that brought about by the motions of the fluid, $\mathbf{F_a} = c\mathbf{v}$; the "diffusive" flux $\mathbf{F_d}$ is accomplished by motions on smaller scale (gusts and eddies in the wind, or the equivalent in the sea). Modellers often assume that this flux is proportional to the concentration gradient, $\mathbf{F_d} = K\nabla c$, like molecular diffusion, but with a much larger K. The local rate of change of concentration, $\partial c/\partial t$ involves the divergence of these two fluxes, $\partial c/\partial t = -\nabla(c\mathbf{v}) + K\nabla^2 c +$ other terms.

Most important among "other terms" are processes such as biological consumption, chemical decomposition and in some cases radioactive decay. To a first approximation, we may assume that all these processes go on at a rate proportional to c, so they contribute a term $-\lambda c$ to the equation. Also, there may be processes supplying the compound that are independent of c. (For example, dissolution of sediments, fires,...). We denote the sum of all these contributions by Q, and get the following general transport equation (sometimes called the *diffusion equation*):

$$\partial c/\partial t = -\nabla(c\mathbf{v}) + K\nabla^2 c - \lambda c + Q \tag{14}$$

For a more detailed discussion of equation (14), see for example Pasquill and Smith (1983).

In "forward" uasge of equation (14), one assumes everything to be known except the function $c(x, y, z, t)$, which is required. This problem amounts to solving (numerically) the differential equation (14), usually on a multidimensional grid. A typical application is prediction of the spread of a pollutant.

In inverse problems, data on the concentration c are available, and we want to gain information from this data. In some problems, the unknowns are the three-dimensional velocity field $bfv(x, y, z, t)$ and the diffusion coefficient $K(x, y, z, t)$. A possible approach is to assume that $\mathbf{v}$ and K have been constant in time for very long, so that the concentration field c as we observe it represents a stationary state. We can then assume $\partial c/\partial t = 0$. It is clear that equation (14) readily gives rise to a system of linear equations, if one approximates the gradients $\nabla(c\mathbf{v})$ and the Laplacian $\nabla^2 c$ with finite differences and assumes that c is known at all grid points. The model by Bolin et al (1983) is an example of this approach.

In other problems, particularly in atmospheric chemistry, data are available for $\mathbf{v}$ as well as c. The interesting unknowns are primarily the sources, Q.