Linear regression with errors in both variables: A proper Bayesian approach

Thomas P. Minka

(Executive summary)

October 8, 1999

Abstract

Linear regression with errors in both variables is a common modeling problem with a 100-year literature, yet we have still not achieved the widespread use of a complete and correct solution. Much of this is due to a confusion between joint and conditional modeling and an unhealthy aversion to priors. This paper expands on the proper Bayesian method of Zellner (1971) and Gull (1989), deriving specific parameter estimators and giving an analysis of their performance. They are shown to perform favorably compared to total least squares.

The main findings are:

- Linear regression, which is a conditional linear model, is different from principal components analysis, which is a joint linear model,
- It is important to use a proper prior on both the variance of the noise and the value of the true inputs,
- Special attention needs to be paid to the case when the scatter of the data is small compared to the noise, which can happen when inputs are correlated,
- Maximum-likelihood estimation can be unreliable on this problem; it is better to use a Bayesian parameter estimate, and
- This estimate provides a desirable shrinkage effect that countless penalized regression methods have been searching for.

1 Introduction

There are two fundamentally different ways to fit a line when there is error in both variables. One way is spatial fitting, where the line represents a symmetric, joint model for x and y, and the density of points along the line is the same no matter what the slope. The model is symmetric because if we interchange the data for x and y (and correct the noise model appropriately), the estimated slope should be the exact inverse. Therefore infinity must be a valid estimate for the slope.

Another way is predictive fitting, or regression, where the line represents an asymmetric, conditional model for y given x, and the density of x is the same no matter what the slope. The model is asymmetric because interchanging the data for x and y need not invert the slope; indeed the slope may remain the same. Infinity is not a valid estimate for the slope, because it makes no sense to predict y as x times infinity.

Virtually all of the literature in applied fields like computer vision, astronomy, biology, and physics concerns joint modeling, even in cases where regression would be more appropriate. For example, a time-series model $p(x_t, x_{t-1}, ..., x_1)$ is more typically parameterized by conditionals $p(x_t|x_{t-1})$ than joints $p(x_t, x_{t-1})$. Statisticians have worked on regression but they often shoot themselves in the foot by using priors that are too weak or too strong.

The main exceptions to this are Zellner (1971) and Gull (1989), who used a proper Bayesian approach to regression with reasonable priors. This paper expands on their method, deriving specific parameter estimators and giving an analysis of their performance. Understanding the estimators requires educating our intuition but in the end they are seen to behave reasonably. Indeed, they automatically exhibit a shrinkage effect which other researchers have deliberately inserted into their algorithms.

As an example of the difference between spatial fitting and predictive fitting, consider the model

$$\begin{aligned} u, v, w &\sim \mathcal{N}(0,1) && (1) \\ x &= u + 2w && (2) \\ y &= v + 2w && (3) \end{aligned}$$

The covariance matrix of $\begin{bmatrix} x \\ y \end{bmatrix}$ is $\begin{bmatrix} 5 & 4 \\ 4 & 5 \end{bmatrix}$. If we sample (x, y) pairs and make a scatter plot (figure 1), then they tend to follow the line $y = x$, convolved with Gaussian noise. The best joint linear model is therefore arguably $y = x$, and this is what principal components analysis (PCA) would say. But if we want to predict y from x, what line should we use? Since x and y are jointly Gaussian, it is straightforward to compute $p(y|x) = \mathcal{N}(4/5x, 9/5)$. So we should predict $y = 4/5x$. What about predicting x from y? As before we compute $p(x|y) = \mathcal{N}(4/5y, 9/5)$, so we should predict $x = 4/5y$. The optimal prediction coefficient is not necessarily the slope of the joint model, and it is not necessarily inverted when we flip variables. As x and y get less correlated, the prediction coefficient goes to zero; it can never be infinity.

We can interpret this as a measurement error model, if we define

$$x_0 = 2w \tag{4}$$

to be the true value of x. In this case we are interested in the relationship between x_0 and y, and the difference between the two approaches is more subtle. With infinite data, both PCA

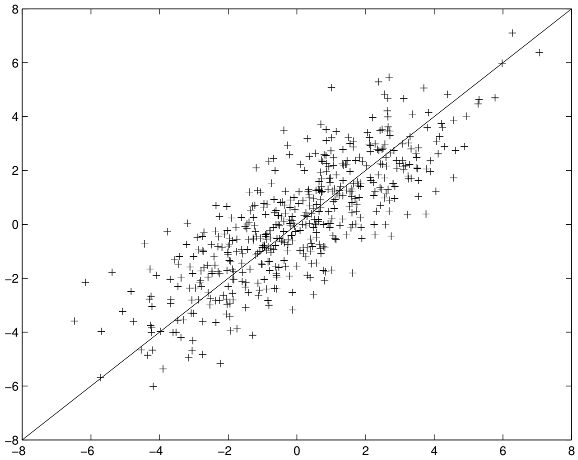


Figure 1: Scatter plot of $\begin{bmatrix} x \\ y \end{bmatrix} \sim \mathcal{N}(\mathbf{0}, \begin{bmatrix} 5 & 4 \\ 4 & 5 \end{bmatrix})$.

and the estimator proposed in this paper will say $y = x_0$. But with finite data, there can be significant differences in accuracy, because PCA is trying to model two things (x and y) while the new estimator only one (y).

2 The model

The model is that y is related to x_0 via a linear scaling plus Gaussian noise:

$$\begin{aligned} y &= ax_0 + e && (5) \\ p(e) &\sim \mathcal{N}(0, v_y) && (6) \end{aligned}$$

which is the usual linear regression model. For the purposes of this summary, there is no offset parameter, the variables are all scalar, and v_y is known. We do not get to observe x_0 directly. We only observe x, a noise-corrupted version of x_0:

$$p(x|x_0) \sim \mathcal{N}(x_0, v_x) \tag{7}$$

where v_x is known. In order to make inferences about a, we need to formalize our prior knowledge about x_0. Here it is assumed that the prior knowledge takes the form of a known Gaussian distribution:

$$p(x_0) \sim \mathcal{N}(0, t) \tag{8}$$

Putting these equations together gives

$$\begin{aligned} p(\begin{bmatrix} x \\ y \end{bmatrix} | a, v_y, v_x, x_0) &\sim \mathcal{N}(\begin{bmatrix} 1 \\ a \end{bmatrix} x_0, \begin{bmatrix} v_x & 0 \\ 0 & v_y \end{bmatrix}) && (9) \\ p(\begin{bmatrix} x \\ y \end{bmatrix} | a, v_y, v_x, t) &\sim \mathcal{N}(\mathbf{0}, \begin{bmatrix} 1 \\ a \end{bmatrix} t \begin{bmatrix} 1 \\ a \end{bmatrix}^{\mathrm{T}} + \begin{bmatrix} v_x & 0 \\ 0 & v_y \end{bmatrix}) && (10) \end{aligned}$$

To simplify the notation in the rest of the paper, let

$$\mathbf{h} = \begin{bmatrix} 1 \\ a \end{bmatrix} \tag{11}$$

$$\mathbf{V} = \begin{bmatrix} v_x & 0 \\ 0 & v_y \end{bmatrix} \tag{12}$$

The problem is to estimate the slope a given a data set $D = \{(x_1, y_1), ..., (x_N, y_N)\}$. Note that the question is "what line explains y given x?" not "what line explains x and y jointly?".

3 Orthogonal Regression and Total Least Squares

One way to eliminate the hidden variable x_0 is to substitute its most likely value conditional on a:

$$\hat{x}_0 = (\mathbf{h}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{h})^{-1}\mathbf{h}^{\mathrm{T}}\mathbf{V}^{-1}\begin{bmatrix} x \\ y \end{bmatrix} \tag{13}$$

$$p(\begin{bmatrix} x \\ y \end{bmatrix} | a, v_y, v_x, x_0 = \hat{x}_0) = |2\pi\mathbf{V}|^{-1/2}\exp(-\frac{1}{2}\begin{bmatrix} x \\ y \end{bmatrix}^{\mathrm{T}}\mathbf{G}\begin{bmatrix} x \\ y \end{bmatrix}) \tag{14}$$

$$\mathbf{G} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{h}(\mathbf{h}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{h})^{-1}\mathbf{h}^{\mathrm{T}}\mathbf{V}^{-1} \tag{15}$$

Under this method, the likelihood for a given D is

$$\prod_i p(\begin{bmatrix} x_i \\ y_i \end{bmatrix} | a, v_y, v_x, x_{0i} = \hat{x}_{0i}) = |2\pi\mathbf{V}|^{-N/2}\exp(-\frac{1}{2}\mathrm{tr}(\mathbf{GS})) \tag{16}$$

$$\mathbf{S} = \sum_{i=1}^{N}\begin{bmatrix} x_i \\ y_i \end{bmatrix}\begin{bmatrix} x_i \\ y_i \end{bmatrix}^{\mathrm{T}} \tag{17}$$

If $\mathbf{q}$ is the principal eigenvector of $\mathbf{SV}^{-1}$, then the maximum is $\hat{a}_{TLS} = q_2/q_1$. Thus, even though we started with a conditional model, we have achieved an answer equivalent to total least squares (TLS) and PCA, which is a spatial fitting (joint modeling) method. This equivalence has probably served to cloud the distinction between spatial fitting and predictive fitting for many people.

Figure 2 plots the log of this likelihood for $\mathbf{V} = \mathbf{I}$ and $\mathbf{S} = \begin{bmatrix} 14 & 8 \\ 8 & 26 \end{bmatrix}$. The maximum at $\hat{a}_{TLS} = 2$ is circled.

An alternative to using the most likely value of x_0 is to include the prior (8) and use the mode of the posterior. This shrinks $\hat{x}_0$ by a constant factor, which increases the slope estimate by the same factor. Otherwise there is little difference.

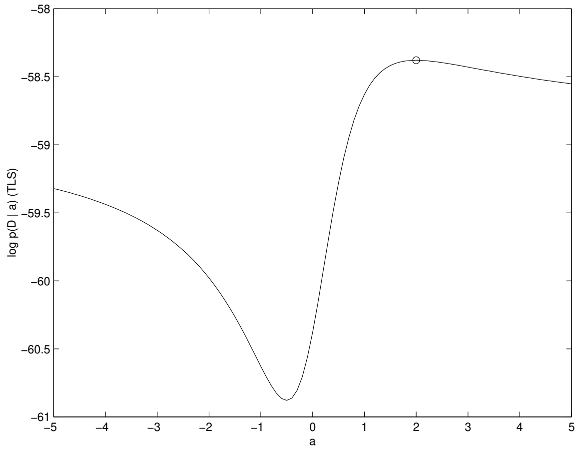


Figure 2: The total least squares log-likelihood

4 Marginal likelihood

The Bayesian approach is to instead integrate out x_0, leaving the marginal likelihood for a:

$$p(D|a, \mathbf{V}, t) = \prod_i p(\begin{bmatrix} x_i \\ y_i \end{bmatrix} | a, v_y, v_x, t) \quad (18)$$

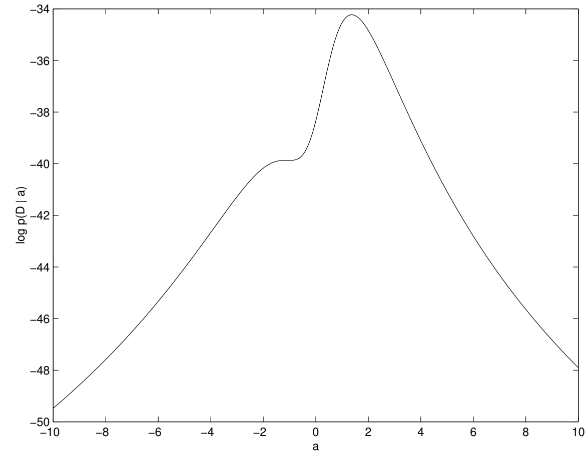
$$= |2\pi\mathbf{V}|^{-N/2} (t\mathbf{h}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{h} + 1)^{-N/2} \exp(-\frac{1}{2}\mathrm{tr}(\mathbf{G}(t)\mathbf{S})) \quad (19)$$

$$\mathbf{G}(t) = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{h}(\mathbf{h}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{h} + t^{-1})^{-1}\mathbf{h}^{\mathrm{T}}\mathbf{V}^{-1} \quad (20)$$

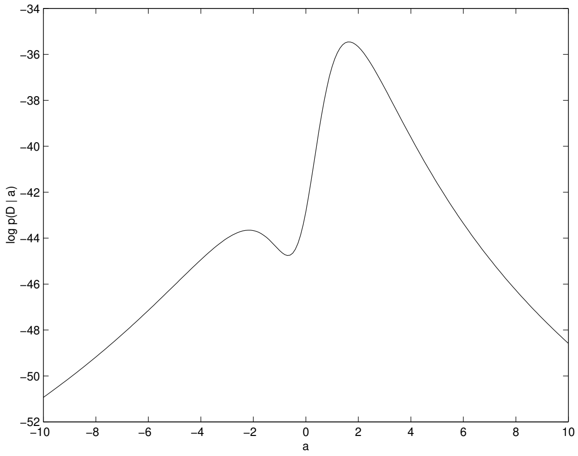
The maximum likelihood estimate $\hat{a}$ is not available in closed form and must be computed by an iterative algorithm such as EM. The likelihood (19) was in the hands of Zellner (1971) and Gull (1989), but they never used it to define an estimator.

Figure 3 plots the marginal likelihood as the eigenvalues of $\mathbf{S}$ change. In each case, the noise variance $\mathbf{V}$ is $\mathbf{I}$, the prior variance t is 1, $N = 10$, and the principal eigenvector $\mathbf{q}$ of $\mathbf{S}$ corresponds to $y = 2x$, i.e. $\hat{a}_{TLS} = 2$, as in figure 2. We see that the eigenvalues, which mean little to the TLS likelihood, are very influential to the marginal likelihood. We also see that the marginal likelihood may have two finite local maxima, which has implications for inference, as discussed below. The dependence of the marginal likelihood on t is not shown here but essentially the marginal likelihood becomes more like the TLS likelihood as t gets large (though they are never identical).

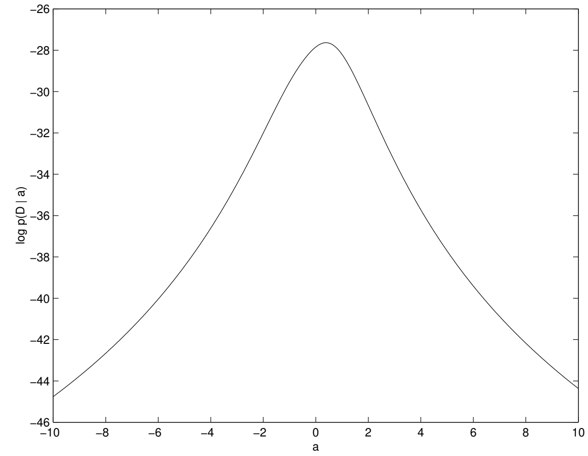
One case of interest is when the largest eigenvalue is small. In this case, the likelihood has one maximum and it is near zero. This makes sense since almost all of the variation in the data can be explained by noise. The likelihood therefore gravitates to the noncommital value of $a = 0$. Why zero? The preference for zero is something specific to predictive fitting, and comes from the normalization term in (19), which arose when we integrated out x_0. The more possible values



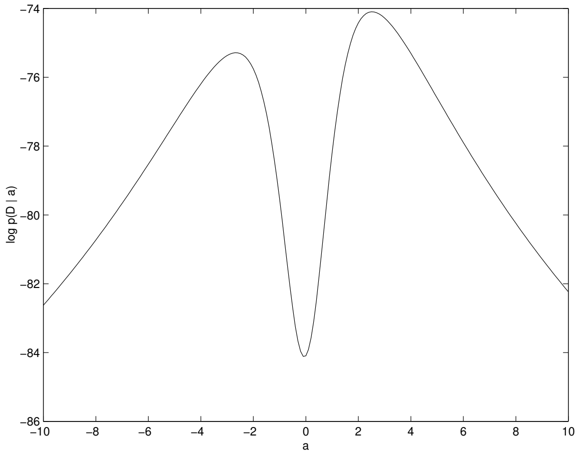
$\lambda = (3, 1)N, \hat{a} = 1.37$



$\lambda = (4, 1)N, \hat{a} = 1.64$



$\lambda = (1, 0.5)N, \hat{a} = 0.39$

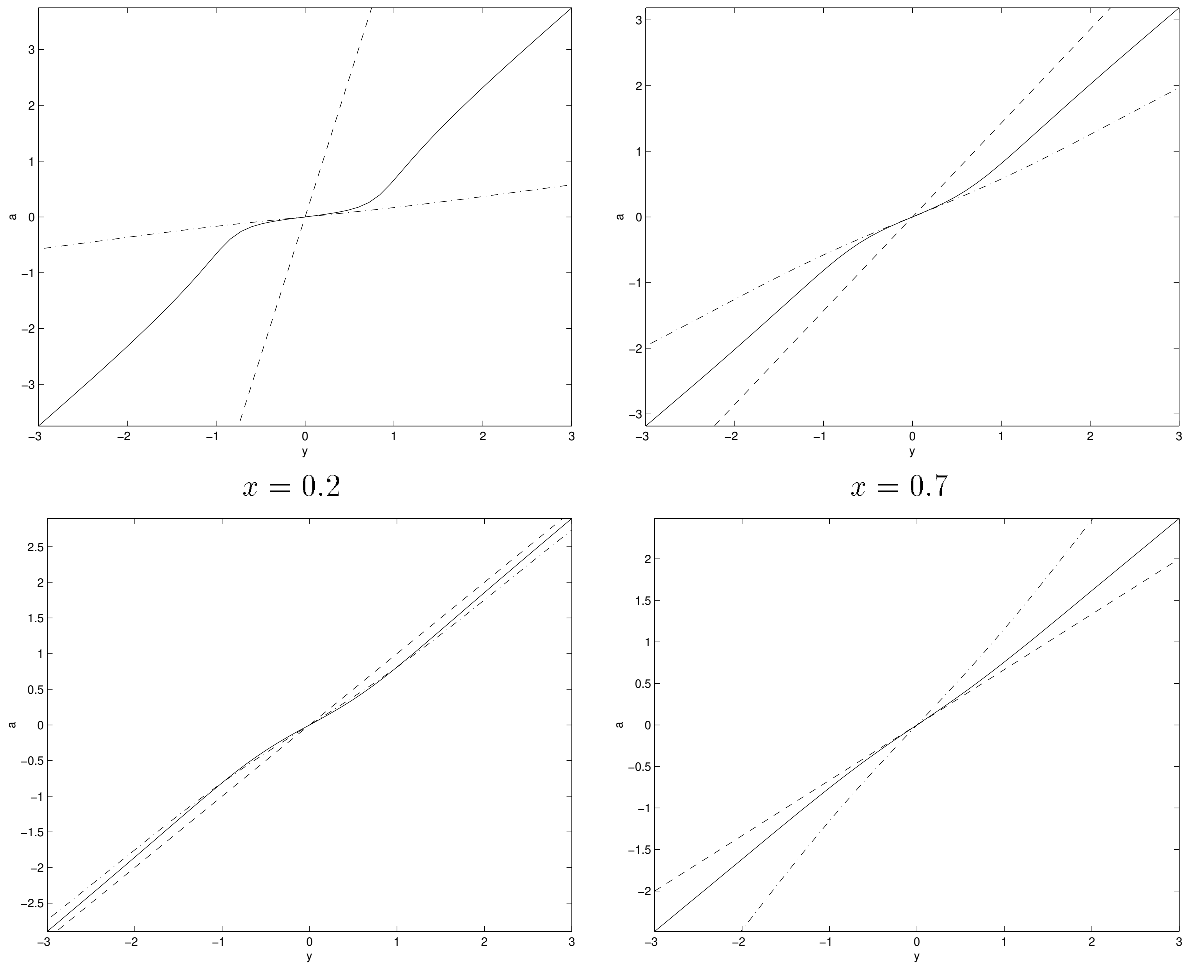


$\lambda = (8.5, 8)N, \hat{a} = 2.52$

Figure 3: The marginal log-likelihood as the eigenvalues of the scatter matrix change. The TLS log-likelihood, and therefore the TLS estimate, is unchanged from figure 2.

x_0 can have, the more probability we accumulate by marginalization. Since $a = 0$ decouples x_0 and y, it allows the most possible values of x_0 and therefore has the highest likelihood, all else being equal. Because it prefers simpler models, this term is called the "Occam factor."

This behavior of the marginal likelihood can be seen by plotting $\hat{a}$ for $D = \{(x, y)\}$ as a function of y. There is only one data point, so both least-squares and total least squares choose the slope which goes through this point: $\hat{a}_{TLS} = y/x$. But the marginal likelihood may peak at a smaller or larger value: see figure 4 (the posterior mean curve is explained in section 5). The marginal likelihood balances two competing effects: an increase in a due to shrinkage of x_0 (mentioned at the end of section 3) and a decrease in a due to the Occam factor. When $\hat{x}_0^2 + y^2 \leq 1$, i.e. the estimated scatter falls under the noise variance, the Occam factor wins and $\hat{a}$ is driven sharply to zero. Otherwise, if t is small compared to x, so that $\hat{x}_0 << x$, then $\hat{a}$ is pushed away from



$x = 0.2$ $x = 0.7$

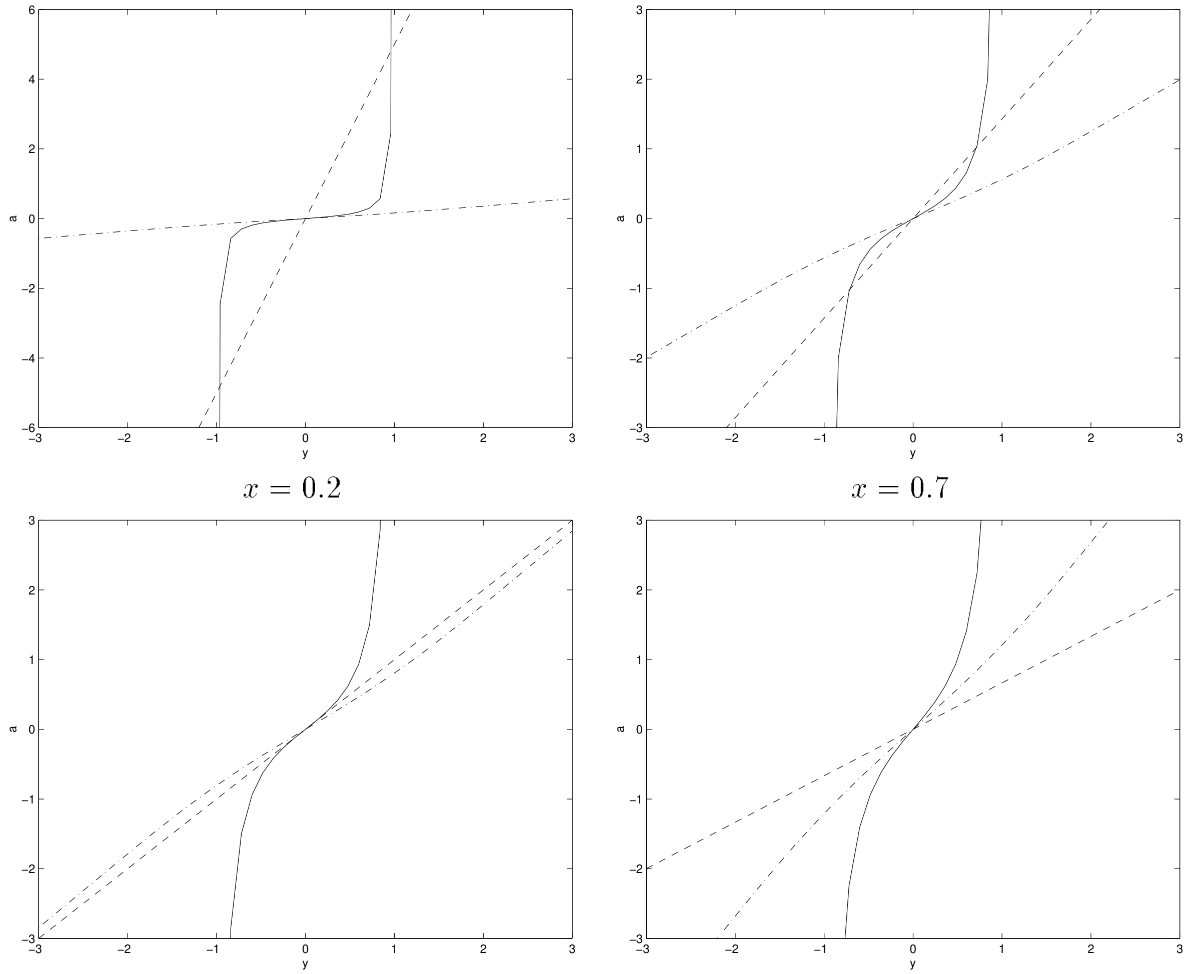
$x = 1$ $x = 1.5$

Figure 4: The estimate of a versus y when $D = \{(x, y)\}$, $t = 1$. Dashed: TLS maximum. Solid: Marginal likelihood maximum. Dash-dot: Posterior mean.

zero, exceeding $\hat{a}_{TLS}$. The behavior of $\hat{a}$ relative to $\hat{a}_{TLS}$ is summarized in the following table:

	y small	y large
$t < x^2$	$\downarrow$	$\uparrow$
$t > x^2$	$\downarrow$	$\downarrow$

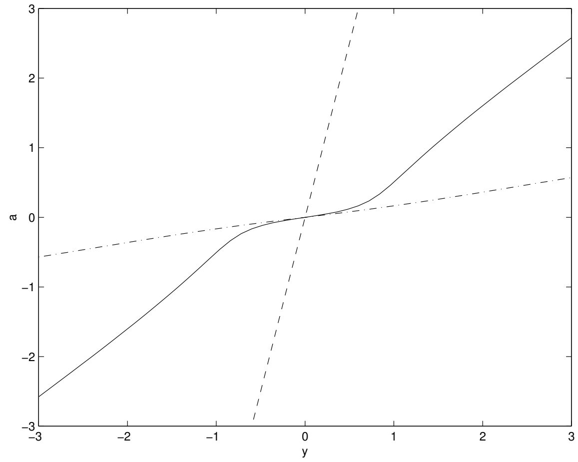
If we set $t = x^2$, then for large y the likelihood maximum approaches TLS. But the Occam factor still reigns for small y. The sensitivity to t is illustrated in figures 5 and 6.



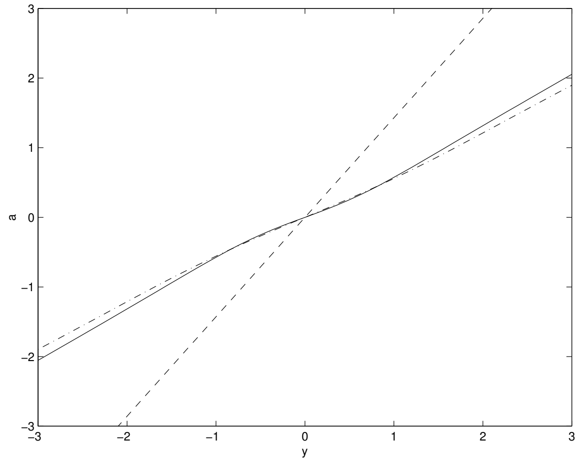
$x = 0.2$ $x = 0.7$

$x = 1$ $x = 1.5$

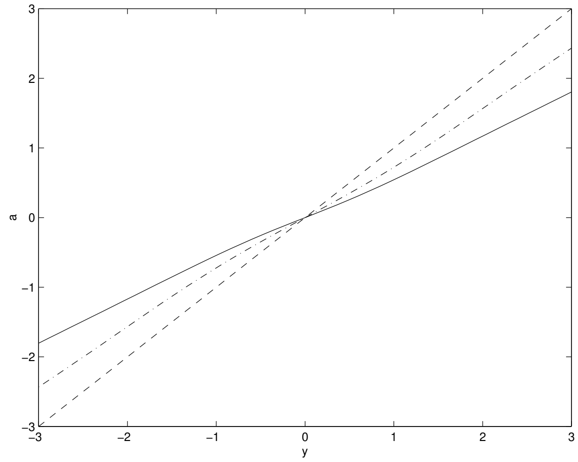
Figure 5: The estimate of a versus y when $D = \{(x, y)\}$, $t = 10^{-5}$. Dashed: TLS maximum. Solid: Marginal likelihood maximum. Dash-dot: Posterior mean.



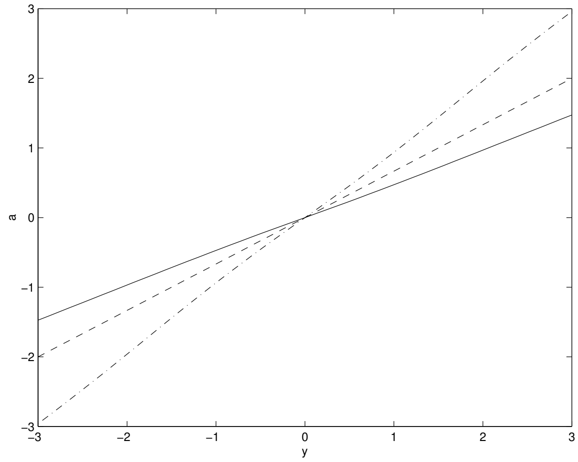
$x = 0.2$



$x = 0.7$



$x = 1$



$x = 1.5$

Figure 6: The estimate of a versus y when $D = \{(x, y)\}$, $t = 10^5$. That is, the prior for x_0 is essentially uniform. Dashed: TLS maximum. Solid: Marginal likelihood maximum. Dash-dot: Posterior mean.

5 Posterior mean

The second case of interest is when the eigenvalues are comparable to eachother (relative to N). The behavior in this case is more disturbing. Since the eigenvalues are comparable, there are two maxima of almost equal quality. Furthermore, these maxima are *pushed away* from $a = 0$. So a maximum-likelihood estimate of slope could suddenly flip from a large postive solution to a large negative solution, if the data were to change slightly (by shifting left or right). The problem here is not with the marginal likelihood but with the idea of *maximizing* the marginal likelihood.

With a prior on a, we can turn the likelihood into a posterior, and then apply decision theory to get the most appropriate estimate. For example, if we want to minimize the expected squared error between the estimate and the true value, then we should use the posterior mean. The posterior mean is also appropriate from the standpoint of prediction, since

$$E[y|x, D] = E[a|D]x \tag{21}$$

Figures 4, 5, and 6 compare $\hat{a}$ with $E[a|D]$ for $N = 1$. The prior $p(a)$ is uniform and the posterior mean was computed via numerical integration. The posterior mean does not have a sharp change between two regimes; it picks one regime or the other and stays that way. When x is small, it stays in the Occam regime, regardless of the value of t. In this regime, $E[a|D] \approx xy$. When x is large, it stays in the inflation regime, with the amount of inflation determined by t but never exceeding x^2. So overall the posterior mean is less sensitive to t than $\hat{a}$. Of course, this is not necessarily true for large N: if we replicate the data set, $\hat{a}$ is unchanged but the posterior mean converges to $\hat{a}$.

6 Results

For a theoretical comparison of these three estimators (TLS, marginal ML, and posterior mean), we can synthesize data from the model and compare the estimation errors. In the first experiment, the model is $a = 2$ and $t = 1$ where t is known. Figure 7 plots the cumulative error histograms for these three estimators. The Bayesian estimators are better on average and in worst-case, especially for small N.

Of course, the Bayesian estimators are being given additional information in the form of t. Figure 8 illustrates what happens if the value of t used in the estimators is wrong. However, if we are uncertain about t, then we shouldn't use an estimator that assumes t is known. It is better to put a prior on t and marginalize it out as well. This is described in the full paper.

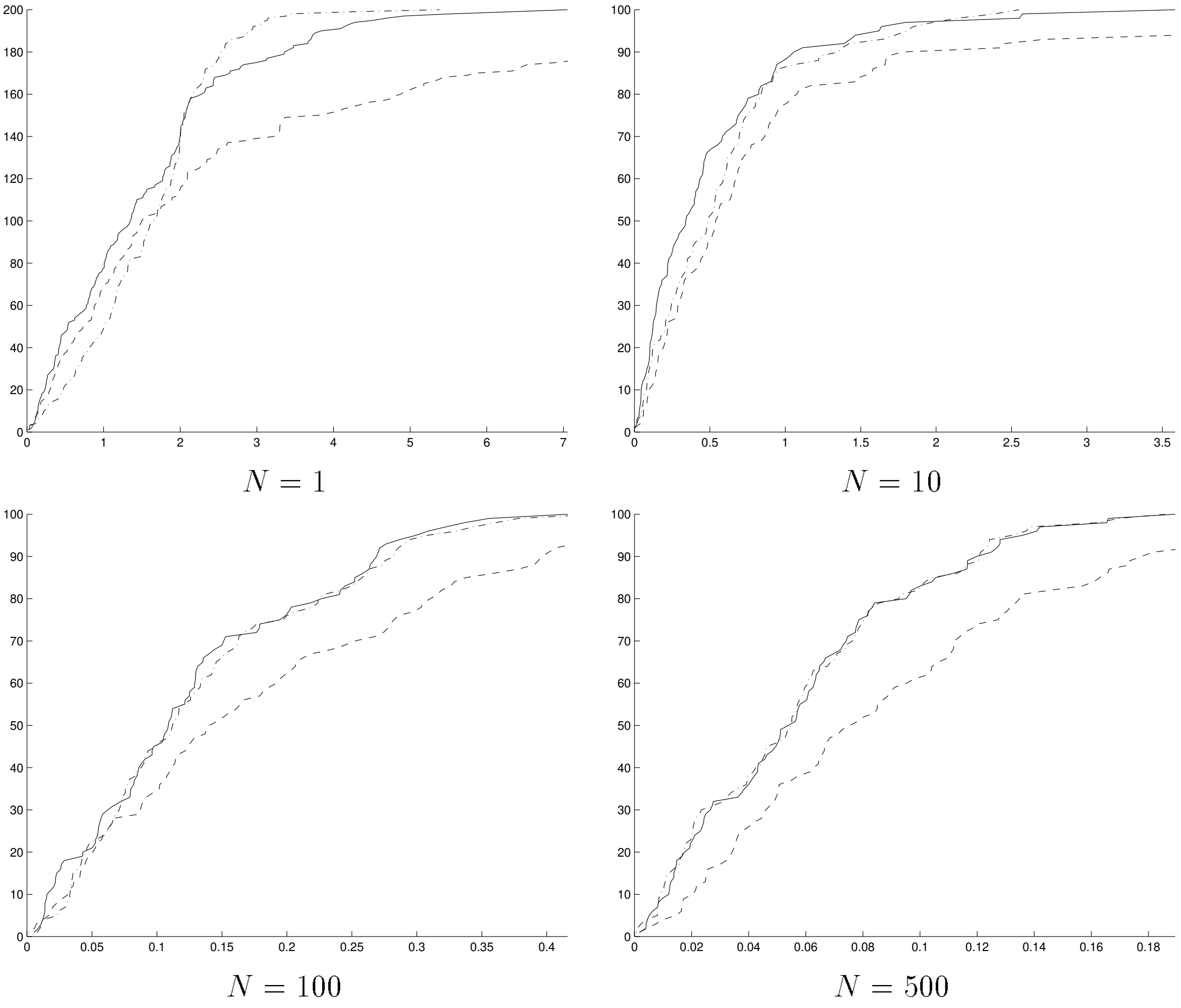
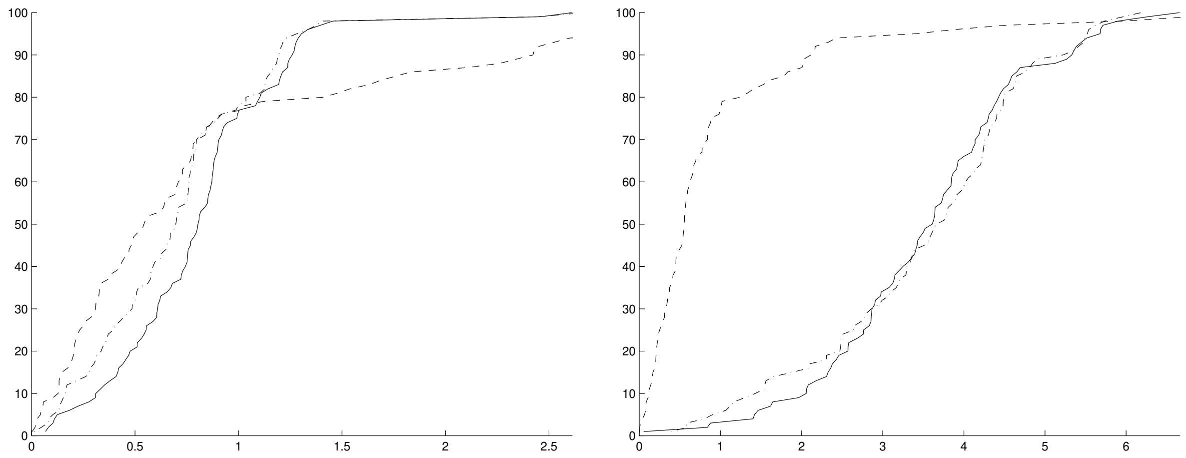


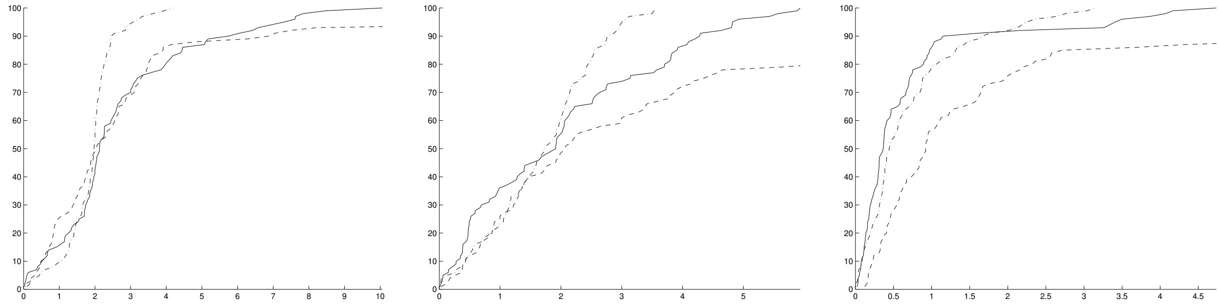
Figure 7: Cumulative error histograms for 100 replications. The horizontal axis is the error and the vertical axis is the percentage of replications which fell within that error. Higher curves are better. The graph stops when the ML estimate reaches 100%. Dashed: TLS maximum. Solid: Marginal likelihood maximum. Dash-dot: Posterior mean.



$t = 10$ $t = 0.1$

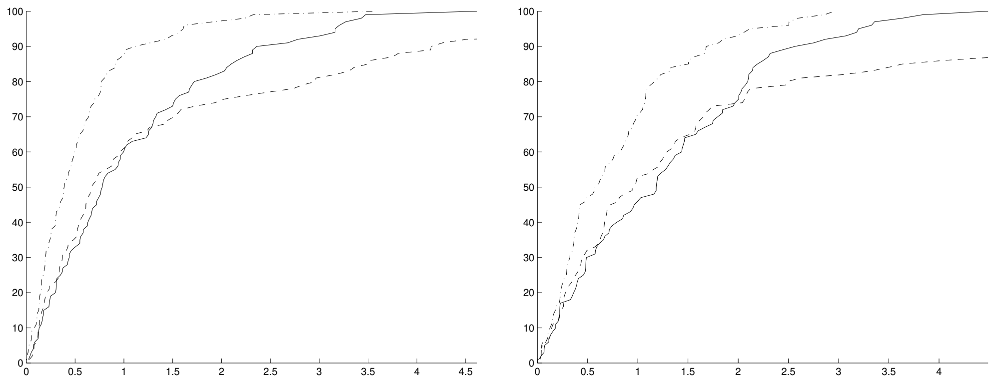
Figure 8: Same as figure 7 but the Bayesian estimators are given the wrong value of t (the true value is $t = 1$). Here $N = 10$.

The next set of experiments make the difference between $\hat{a}$ and $E[a|D]$ more apparent. Here $t = 0.1$, making the noise a major part of x. Figure 9 has the results. This situation arises especially in the multivariate case when inputs are correlated, since that is equivalent by rotation to having an input with small variance. If additionally a is smaller than 2, so that noise is a major part of y, the difference is more dramatic, as shown in figure 10.

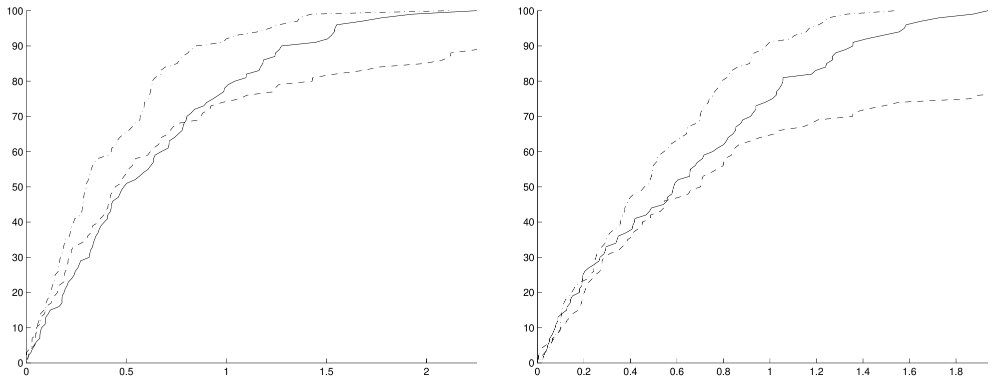


$N = 1$ $\qquad$ $N = 10$ $\qquad$ $N = 100$

Figure 9: Same as figure 7 but the true value of t is 0.1.



$a = 0.1,\ N = 10$ $\qquad$ $a = 0.5,\ N = 10$



$a = 0.1,\ N = 100$ $\qquad$ $a = 0.5,\ N = 100$

Figure 10: Same as figure 9 but smaller values of a. Here $N = 10$.

References

[1] S. F. Gull. Bayesian data analysis: Straight-line fitting. In J. Skilling, editor, *Maximum Entropy and Bayesian Methods*, pages 511–518. Kluwer Academic Publishers, 1989. `http://bayes.wustl.edu/sfg/gull.html`.

[2] Arnold Zellner. *An Introduction to Bayesian Inference in Econometrics*. John Wiley & Sons, 1971.