Better predictions when models are wrong or underspecified

Ommen, M. van

Chapter 5

Bayesian Inconsistency: More Experiments

This chapter provides additional linear regression experiments to test whether the results of Chapter 3 also hold with different priors, models, methods, and data-generating distributions. We find that three versions of SafeBayes consistently perform well, while other methods, including Bayes and AIC, perform badly.

5.1 Experiments on variations of the prior and the model

Apart from the priors on parameters given the models we used in our main experiments, we tried several alternative prior distributions, described in the subsections below. The first subsection describes experiments with fixed (i.e., a degenerate prior on) σ^2.

5.1.1 Experiments with fixed σ^2

When models with fixed σ^2 are used, our two SafeBayes methods become *R*-square- and *I*-square-SafeBayes, as defined in Section 3.4.2. These also have a direct interpretation as trying to find the best η for predicting with a square-loss function, as was explained in that section. In this context, the value $\eta = 1$ has no special status, so we now also tried values $\eta > 1$ (we did experiment with varying η in the previous varying σ^2 experiments as well, but there it did not make any substantial difference in the results). Specifically, we set $\mathcal{S}_n$ in the SafeBayesian algorithm to $\{2^{\kappa_{\max}}, 2^{\kappa_{\max}-\kappa_{\text{STEP}}}, 2^{\kappa_{\max}-2\kappa_{\text{STEP}}}, \ldots, 2^{-\kappa_{\max}}\}$, with $\kappa_{\text{STEP}} = 1/2$ and $\kappa_{\max} = 6$. All priors on the regression coefficients β remain as described in Section 3.5.1.

5.1.1.1 Model averaging experiment, fixed σ^2

The model-correct experiment showed no surprises (all methods performed well), so we only show results for the model-wrong experiment, as described in Section 3.5.1, testing each of Bayes, R-square- and I-square-SafeBayes twice: once based on a model with variance σ^2 overly large (3 times $\tilde{\sigma}^2$), and once with σ^2 overly small (1/3 times $\tilde{\sigma}^2$) variance. To allow precise comparison with the results in the main text, we also show behaviour of R-log-SafeBayes with varying variance (defined precisely as in Figure 3.3) in Figure 5.1.

5.1.1.2 Ridge regression experiments, fixed σ^2

Again we only show results for the model-wrong experiment.

Note that here standard Bayes — as can be seen from plugging $\eta = 1$ into (3.12) — does not depend on σ^2 and thus coincides in terms of square-risk behaviour with standard Bayes in the variable σ^2 case as in Figure 3.7. Also (see below (3.12)) I-square-SafeBayes for fixed σ^2 does not itself depend on σ^2 and simply minimizes the cumulative sum of squared errors.

Just as for ridge regression with variable σ^2, one may equivalently interpret the η-generalized-posterior means $\bar{\beta}_{i,\eta}$ as the standard, nongeneralized Bayesian posterior means that one would get with a modified prior on β, proportional to the original prior raised to the power η^{-1} (see above (3.31), Section 3.5.4). It may then once again seem reasonable to learn η itself in a Bayesian or likelihood-based way such as empirical Bayes.[1] Indeed, this was suggested implicitly as early as 1999 by one of us (Grünwald, 1999). The procedure described in Section 3.4.3 ('hierarchical loss') of Bissiri et al. (2013) also arrives, via a different derivation, at a similar prescription for finding η (we immediately add that the authors describe many ways for determining η, of which this is just one). Unfortunately, just as for the empirical Bayes learning of η with varying σ^2, the figures below indicate that it does not perform well at all.

Conclusion Standard Bayes again performs comparably badly in both experiments (note the difference in scale in the first graphs of Figures 5.1 and 5.2). I-square-SafeBayes behaves excellently in both experiments. But now in the ridge experiment R-square-SafeBayes becomes a highly problematic method for small samples, worse even than standard Bayes. The reason is its dependence on the specified σ^2 as can be clearly seen from (3.23). If σ^2 was set to be much larger than the actual average prediction error on the sample, then the third term in (3.23) dominates. This term decreases with η and thus automatically pushes $\hat{\eta}$ 'upward' by an arbitrary amount. The term also decreases with n, so that the problem disappears at a large enough sample size. The problem did not occur in the model averaging experiment; we suspect that this is because in this experiment, there is substantial prior mass on a small model

[1] In the present setting, learning η by empirical Bayes has a second interpretation: if one fixes the variance σ^2 appearing in the prior on β, uses the linear model with a different variance σ'^2, and then learns σ'^2 by empirical Bayes, the result is identical to fixing $\sigma'^2 = \sigma^2$ and learning η by empirical Bayes.

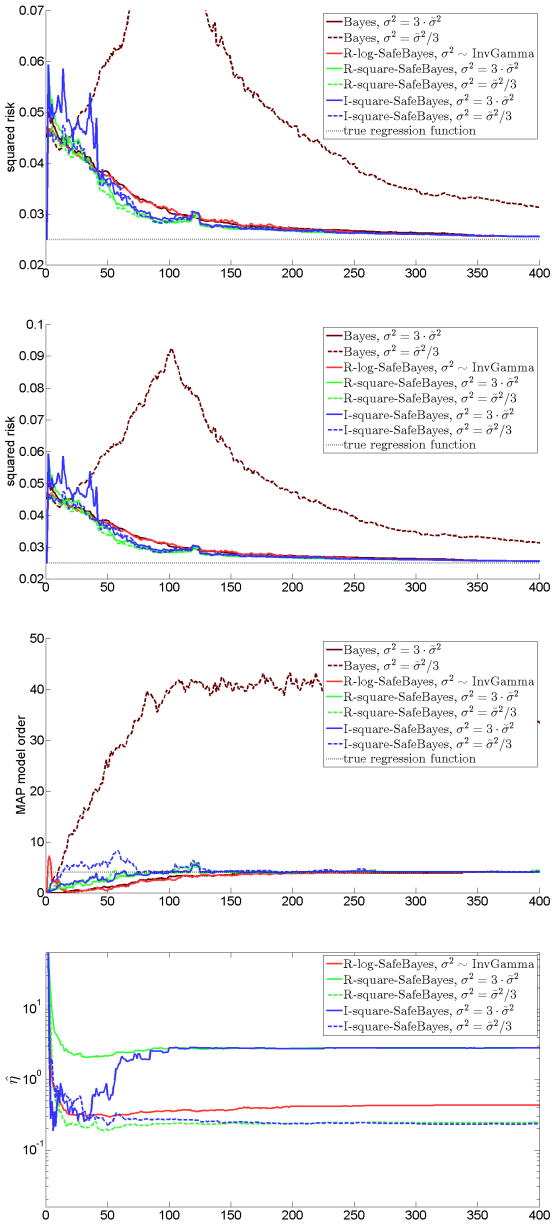

Figure 5.1: Bayesian model averaging, fixed σ^2, for the model-wrong experiment of Figure 3.3 with $p_{\max} = 50$. The second graph is a scaled version of the first. Since fixed σ^2 implies fixed self-confidence ratio, the self-confidence graph is not shown. For clarity in the η-graph we do not show standard deviations of the η's.

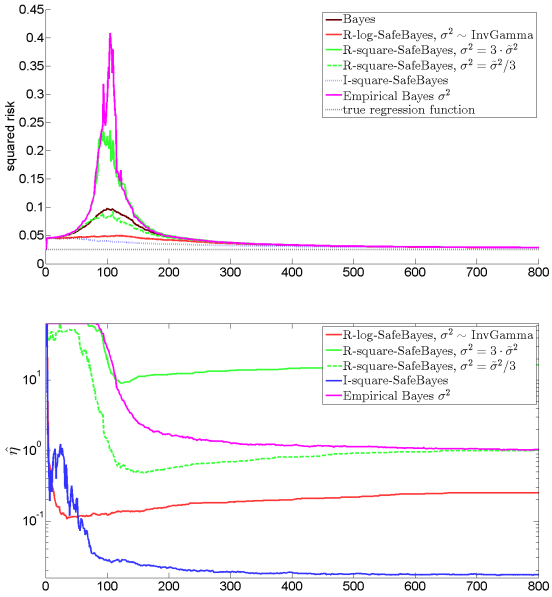

Figure 5.2: Bayesian ridge regression: Same graphs as in Figure 3.7, for fixed σ^2 and the model-wrong experiment conditioned on $p := p_{\max} = 50$. Note the difference in scale for the risk in this figure and Figure 5.1.

($p = 4$) containing the pseudo-truth, and for this submodel, the final term in (3.23) (which is approximately linear in p) is much smaller than for $p = 50$ and does have not such a strong influence.

5.1.2 Slightly informative prior

Again we only consider model-wrong experiments. Within each model, we now use the following prior parameters: $\bar{\beta}_0 = \mathbf{0}$ and $\Sigma_0 = 10^3\mathbf{I}$ for the multivariate normal distribution on β; and $a_0 = 1$ and $b_0 = \sigma^{*2}a_0$ (as before) for the inverse gamma distribution on σ^2 (where σ^{*2} is the true marginal variance of noise in our data, as defined in Section 3.5.1.2). We repeated the model-wrong experiment of Section 3.5.3 with $p_{\max} = 50$ with this slightly informative prior and obtained similar results to those obtained using our original informative prior with $\Sigma_0 = \mathbf{I}$: Bayes performs badly roughly between samples 90 and 130 and has some risk spikes before that so that its overall performance is comparable to before, while R-log-SafeBayes and I-log-SafeBayes both obtain good risks. We omit the pictures as they show no surprises.

We also repeated the model-wrong experiment for ridge regression (Section 3.5.4). Here the effect of the new prior on Bayes' performance is similar: the square-risk now peaks at a larger value, but in a smaller range of sample sizes. However, the effect of changing the learning rate is different in this ex-

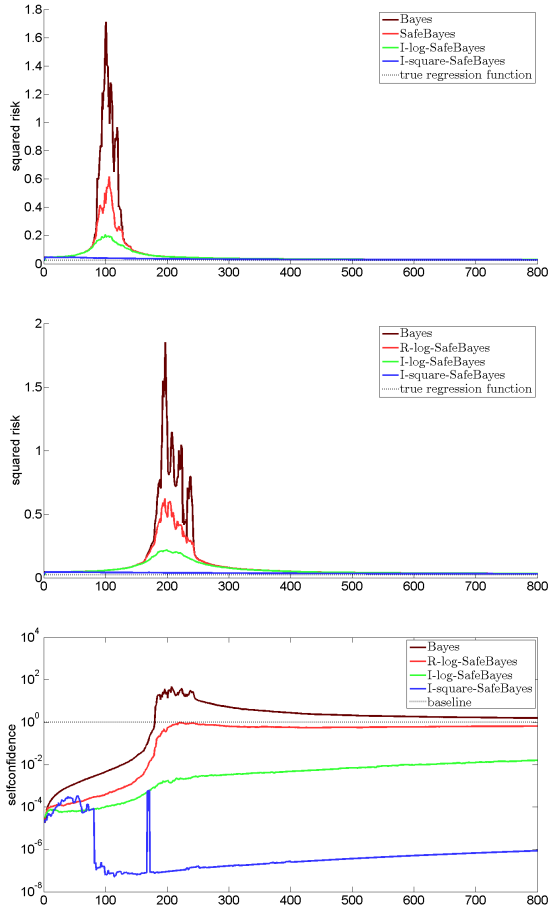

Figure 5.3: Top two graphs: square-risk for two different ridge experiments. In both experiments the slightly informative prior of Section 5.1.2 is used. In the first experiment $p = 50$, in the second $p = 100$; otherwise the experiments are just as the 'wrong model experiment' of Section 3.5.4, Figure 3.7, but we also included performance of I-square-SafeBayes. Final graph shows self-confidence for the $p = 100$ case for Bayes and SafeBayes, on a logarithmic scale because of the range of values involved.

periment than what we have seen before: here one can take η *very* small and still get good results. So in a sense, the problematic behaviour of Bayes has a trivial solution here: just pick a very small but fixed η. R-log-SafeBayes was too conservative in this, I-log-SafeBayes did fine. R-log-SafeBayes became competitive again however, if we used the discounting version described in Section 5.2.1 below.

In Figure 5.3 we repeat the pictures for ridge regression (Section 3.5.4) with

this slightly informative prior, because they give additional insight. Note that the phenomenon is now much more 'temporary'. In the beginning, it seems that there is a sort of cancellation between the influence of the irrelevant variables and standard Bayes behaves fine. However, if we increase the number of irrelevant variables, the problem (while starting at a later sample) takes longer to recover from.

5.1.3 Prior as advised by Raftery et al.

In Raftery et al. (1997), some guidelines for choosing priors in regression models are given. Letting $\bar{\beta}_0$ denote the prior mean, one of their recommendations is that the prior densities for $\beta = \bar{\beta}_0$ and $\beta = \bar{\beta}_0 + \mathbf{1}$ should differ by a factor of at most $\sqrt{10}$. The prior density on β marginalized over σ^2 follows a multivariate t-distribution, and the factor in question varies with the dimensionality of β, so that models of larger order are given less informative priors. In our case, we find that the resulting prior is always less informative than our original prior, and for model $\mathcal{M}_{10}$ and above (i.e. β of dimension 11 or larger), it becomes even less informative than the prior introduced in the previous section.

For the prior on σ^2, Raftery et al. advise that the density should vary by no more than a factor 10 in a region of σ^2 from some small value to the sample variance of y. For our choice of hyperparameters $a_0 = 1$, $b_0 = 1/40$, the mode of $\pi(\sigma^2)$ is at $b_0/(a_0 + 1) = 1/80$, and the density is within a factor 10 of this maximum in the approximate region $(0.0037, 0.0941)$. For the correct model experiments, the actual variance of Y is 0.065; for the wrong model experiments, it is 0.045 (with a larger variance for 'good' points and zero variance for 'easy' points). For both experiments, the factor-10 condition holds between $\mathrm{Var}(Y)/12$ and $\mathrm{Var}(Y)$. We conclude that this prior satisfies the guidelines in Raftery et al. quite well.

We will refer to the prior described above as Raftery's prior (even though it is really a different prior for each model order). Using this prior, we found the following experimental results.

In the model-wrong setting of Section 3.5.3 (model selection/averaging), with our original prior replaced by Raftery's prior, Bayes performs somewhat *better* than R-log-SafeBayes (except on very small sample sizes). However, I-log-SafeBayes performs as well as Bayes, and so does the R-log-SafeBayes variant that discounts half of the initial sample when choosing the learning rate (see Section 5.2.1).

This might suggest that Raftery's prior could be used to accomplish the same kind of safety against wrong models as SafeBayes provides, at least in a model selection context. To test this, another experiment was performed where the fraction of 'easy' points was increased to 75%. In this experiment, the misbehaviour of Bayes seen in Section 3.5.3 returned worse than before, with risks a factor 20 larger than before, whereas the SafeBayes methods continued to work fine. This suggests that Raftery's prior can not be relied on if the severeness of misspecification is unknown.

If Raftery's prior is used for model selection with a correct model, Bayes and the SafeBayes variants perform well, and very similarly to each other.

For ridge regression, the results with Raftery's prior for both the correct and the incorrect model experiment are very similar to those with the slightly informative prior, except that the peak in the risks is higher for all methods.

5.1.4 The g-prior

Another prior we experimented with was the g-prior, which is a popular choice in model selection contexts (Zellner, 1986; Liang et al., 2008). For all definitions we refer to the latter paper. In contrast to all other priors we considered, the g-prior depends on the design matrix $\mathbf{X}_n$, and hence can only be used in settings where this matrix, and hence the eventual sample size of interest n, is given once and for all. For this reason, we decided to depict in Figure 5.4, for each value of n, the risk obtained when predicting the $(n+1)$-th data point with the posterior calculated from the g-prior corresponding to the first n covariates $(x_1, \ldots, x_n)$ and observed data y^n. This is subtly different from our previous graphs (e.g. Figures 3.3–3.6) that show how the risk evolves as n increases in a *single* run of the experiment, averaged over 30 runs.

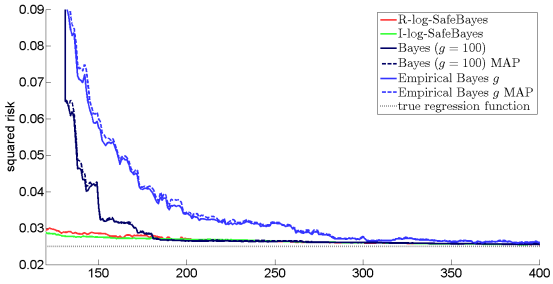

Figure 5.4: Risk as a function of sample size (starting at the first sample size at which the g-prior is defined) for model averaging and selection based on the g-prior in the model-wrong experiment of Figure 3.3 both with $g = 100$ and with g chosen by Empirical Bayes at each sample size

The graph is not shown starting at $n = 0$, because of another difference between the g-prior and the priors we used in other experiments:

Because of the same design dependence, with the g-prior, the posterior on β remains a degenerate distribution on an initial segment of outcomes. For example, with $\mathcal{M}_p$ for $p = 50$, the matrix $\mathbf{X}_n^\top \mathbf{X}_n$ is singular until at least 50 *different* design vectors have been observed. For our model-wrong experiment, this means that on average, about a 100 observations are required before the posterior becomes nondegenerate; this explains why Figure 5.4 starts at n a little over a 100.

The experimental results clearly indicate that the g-prior does not deal with our data in a satisfactory way, regardless of the value of g. Of the values of g

we tried (up to 10^4), $g \approx 100$ (shown in the graph) yielded the smallest square-risk around sample size $n = 200$; for larger sample sizes, larger values of g were better, but only slightly. Furthermore (as in fact we expected by analogy to learning η with Empirical Bayes), the value of g found by Empirical Bayes is not optimal for dealing with our data and only makes things worse: larger values of g (which put more weight on the data) would yield smaller risks.

5.2 Experiments on variations on the method

Below we look at a number of other more or less promising alternative approaches to modifying standard Bayes.

5.2.1 An idea to be explored further: Discounting initial observations

Just like standard Bayes, all our SafeBayesian methods are, at heart, *prequential* (Dawid, 1984). All prequential methods suffer to a greater or lesser extent from the *start-up problem* (Van Erven et al., 2007; Wong and Clarke, 2004): sequential predictions based on a model $\mathcal{M}_p$ may perform very badly for the first few samples. While they quickly recover when the sample size gets large, the behaviour on the first few samples may dominate their cumulative prediction error for a while, leading to suboptimal choices for moderate n. We can address this issue in several ways. A very simple method to 'discount' initial observations, apparently first used (implicitly) to modify standard Bayes factors by Lempers (1971, Chapter 6), is to only look at the cumulative sequential prediction error on the second half of the sample, so that the first half of the sample merely functions as a 'warming-up' sample (Catoni, 2012). Without claiming that this is the 'right' method to discount initial observations, we experimented with it to see whether it can further improve the performance of SafeBayes; for simplicity, we concentrated on R-log-SafeBayes.

We found that in most experiments, this new method for determining η performed very similarly to the standard method based on the whole sample, sometimes slightly better and sometimes slightly worse, making it hard to say whether the new method is an improvement or not. Still, there are two experiments in which the new method performed substantially better, namely the experiments with less informative priors of Section 5.1.2 and 5.1.3. Thus we cannot just dismiss the idea of fitting η based on only part of the data or more generally, discounting initial observations, and it would be interesting to explore this further in future work: of course taking half of the data is rather arbitrary, and better choices may be possible. In particular, we may try a variation of *switching* between η's analogously to the switch distribution (Van Erven et al., 2007) to counter the start-up problem.

5.2.2 Other methods for model selection: AIC, BIC, (generalized) cross-validation

We tested the performance of several classic model selection methods on the same data and models as in our main model selection/averaging experiment, Section 3.5.3. We associated with each model $\mathcal{M}_p$ its standard (i.e. $\eta = 1$) Bayes predictive distribution under the prior described in Section 3.5.1 (these generally perform better than the maximum likelihood distributions based on $\mathcal{M}_p$ whose use is more standard here). We then ran leave-one-out cross-validation, 10-fold cross-validation and GCV based on the predictions (posterior means/MAPs $\bar{\beta}_{i,\eta}$) made by these predictive distributions. We also compared the models via AIC and BIC, where for AIC we used the small-sample correction of Hurvich and Tsai (1989).

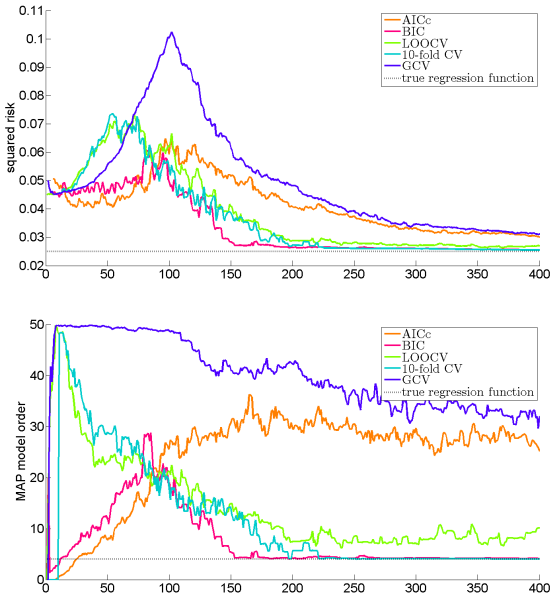

Figure 5.5: Square-risk and selected model order for five different model selection methods. The risks in this graph are risks of single models selected by each method (similar to the MAP risks shown for Bayes and SafeBayes).

We see in Figure 5.5 that AIC and generalized cross-validation have risks and selected model orders similar to those of standard Bayes, though they do not recover as well as Bayes when the sample size increases. Of the other three methods, BIC and 10-fold cross-validation find the optimal model and have smaller risks towards the end than leave-one-out cross-validation, which continues to select larger-than-optimal models with substantial probability. Note that none of the methods can compete with SafeBayes on sample sizes below 150: SafeBayes's risk goes down immediately after the start of the experiment

while for all the other methods it goes up first. Also, SafeBayes finds the optimal model quickly without first trying much larger models.

5.2.3 Other methods for learning η: Cross-validation on log-loss and on squared loss

As indicated in the introduction and Section 3.4.2, finding $\hat{\eta}$ by I-square-SafeBayes is somewhat similar to finding $\hat{\eta}$ by leave-one-out cross-validation with the squared-error loss, the difference being that I-square-SafeBayes finds the optimal η for predicting each point based on past data data points rather than the optimal η for predicting each point based on all other data points. Since the leave-one-out method is often employed in ridge regression, it seemed of interest to try out here as well. Figure 5.6 shows that LOO-cross validation indeed performs very similarly in terms of square-risk to R-log-SafeBayes (and to I-log- and I-square-SafeBayes, which are not depicted here but are similar to R-log-SafeBayes). However, LOO-cross validation is consistently a bit worse in terms of self-confidence; we do not have a clear explanation for this phenomenon.

Perhaps more interestingly, in Figure 5.7 we show what happens if we use LOO-cross validation based on the log-loss of the Bayes predictive distribution, which may seem a reasonable procedure from a 'likelihoodist' perspective. Here we see dismal behaviour, the reason being the hypercompression phenomenon of Section 4.1.3: cross-validation will select a model that, at the given sample size, has small log-risk, but because of hypercompression this model can sometimes perform very badly in terms of all the associated prediction tasks such as square-risk and reliability.

5.3 Experiments on variations of the truth

Other distributions of covariates In all experiments described in Section 3.5 and the earlier sections of this chapter, the covariates $(X_{i1}, X_{i2}, \ldots)$ were sampled independently from a 0-mean multivariate Gaussian. We repeated most of our experiments with $X_{i1}, X_{i2}, \ldots$ that were sampled independently uniformly from $[-1, 1]$, and, as already indicated in the introduction, with polynomials, $X_{ij} = P_j(S_i)$ for P_j the Legendre polynomial of degree j and $S_i \in [-1, 1]$ uniform. This did not change the results in any substantial way, so we do not report on it further.

Fewer easy and 'less-easy' points If the fraction of 'easy' points is reduced, one would expect the performance of standard Bayes to improve. This is confirmed by an experiment where each data point had a probability of only $1/4$ to be $(0, 0)$. Here Bayes still has some trouble finding the optimal model, but the square-risk, MAP model order, and time taken to recover are all much reduced compared to the original experiment in Section 3.5.3 where half the data points

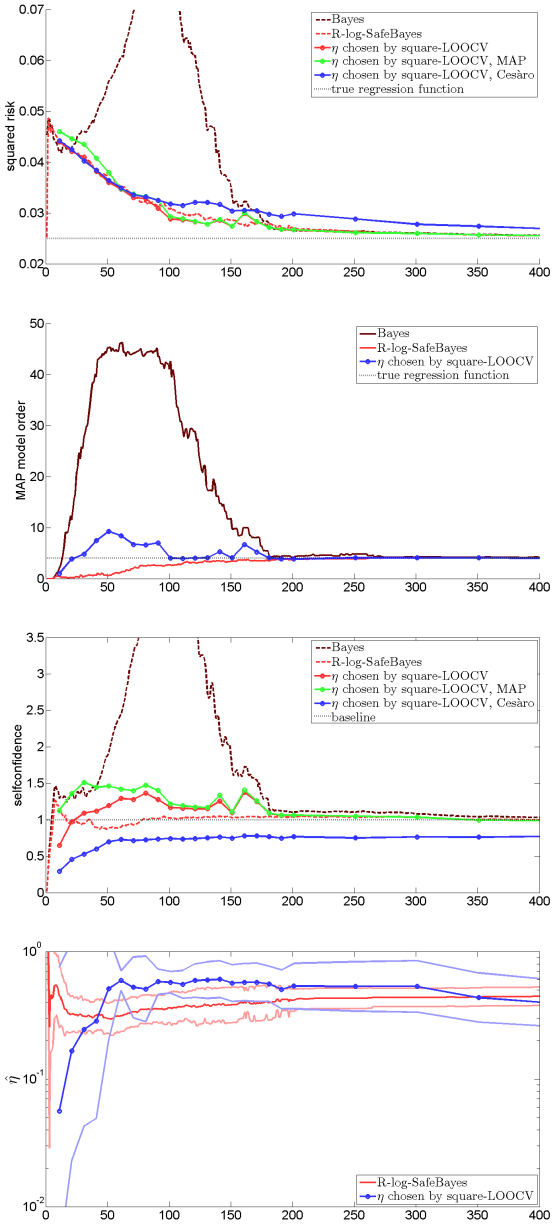

Figure 5.6: Analogue of Figure 3.3 for determining η by leave-one-out cross-validation with square-loss with the wrong-model experiment, $p_{\max} = 50$.

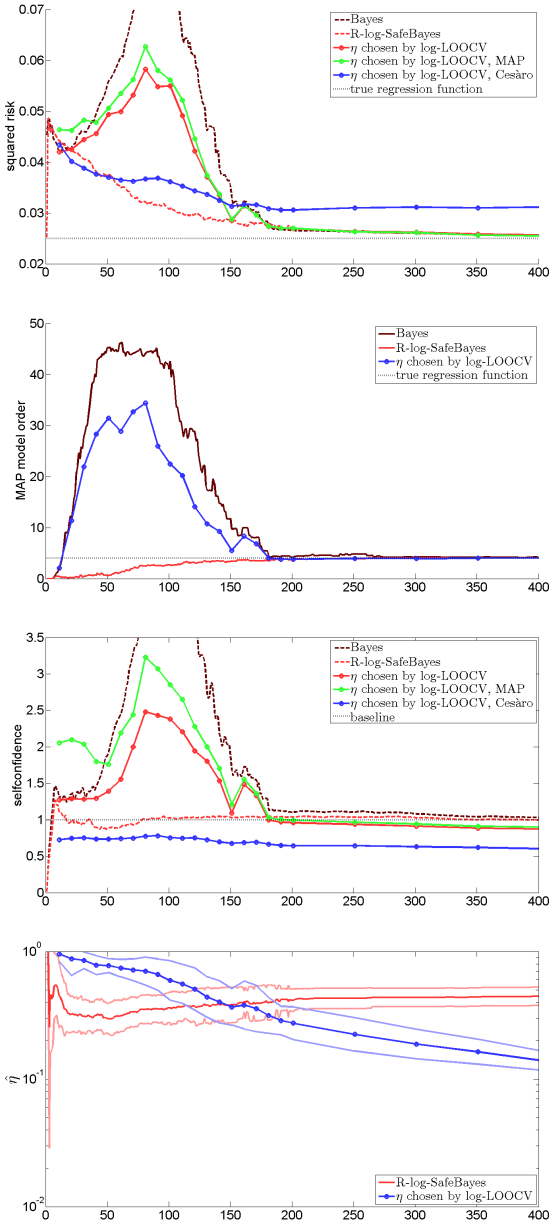

Figure 5.7: Analogue of Figure 3.3 for determining η by leave-one-out cross-validation with log-loss.

were 'easy'. SafeBayes on the other hand showed the same good performance as before.

Two points that might be raised against the use of 'easy' points in our simulations are that they are unlikely to occur in practice, and that if they were to occur, they would be easily detected and dealt with another way. To address this line of argument to some extent, another experiment was performed with a smaller contrast between 'easy' and 'hard' points. Rather than being identically $(0, 0)$, the 'easy' points were random but with smaller variance than the 'hard' points. To be precise, the covariates and noise were both a factor 5 smaller (so that their variances were 25 times smaller). In this experiment, the same phenomena as in Section 3.5.3 occurred, albeit again on a smaller scale (though larger than in the previous, $1/4$-easy experiment).

Different optimal regression functions We experimented with a number of variations of the wrong-model experiment of Section 3.5.3, by changing the underlying 'true' distribution P^*. In each variation, we still tossed, at each i, an independent biased coin to determine whether i would be 'easy' (still probability $1/2$) or 'regular' (probability $1/2$), but in each case we changed the definition of either the 'easy' or the 'regular' instances or both. In all experiments, for the 'regular' instances, only $P^*(Y_i \mid X_i)$ was changed; the marginal distribution of the X_i was still multivariate normal as before. Here is a list of things we tried:

1. For regular instances, set $P^*(Y_i \mid X_i)$ so that $Y_i = 0 + \epsilon_i$ instead of (3.27), with ϵ_i i.i.d. normal as before; easy instances were still set to $(\mathbf{0}, 0)$.

2. For regular instances, (3.27) was replaced by $Y_i = X_{i1} + X_{i2} + X_{i3} + X_{i4} + \epsilon_i$, so the optimal coefficients $\tilde{\beta}_1 \ldots \tilde{\beta}_4$ are ten times as large as in the original experiment; easy instances were still set to $(\mathbf{0}, 0)$.

3. For regular instances, (3.27) was replaced by $Y_i = .1 \cdot (X_{i1} + \ldots + X_{i4}) - .04 + \epsilon_i$ (so the intercept is not 0), and the easy instances were set to $(X_i, Y_i) = (.\mathbf{2}, .04)$, where $.\mathbf{2}$ represents the K-dimensional vector $(.2, \ldots, .2)$. Note that the easy points are on the optimal regression function.

4. For regular instances, (3.27) was replaced by $Y_i = .1 \cdot (X_{i1} + \ldots + X_{i4}) + .5 + \epsilon_i$ so the intercept was again not 0; the easy instances were set to $(\mathbf{0}, .5)$.

We explain each in turn. For the first experiment, all the results were comparable to the results of Experiment 1 in Section 3.5, so we do not list them. For the second experiment, the risks obtained by standard Bayes and SafeBayes were similar to each other. The model order behaviours were similar to what they were before (with standard Bayes selecting large model orders initially), but all methods recovered much more quickly, converging on the optimal model shortly after $n = 50$; presumably this could happen because now the optimal coefficients were substantially larger than the standard deviation in the data.

The third experiment was included to see whether there would be an effect if the 'easy' points would be placed at an arbitrary point rather than the special,

fully symmetric $(\mathbf{0}, 0)$. We added the intercept -0.04 so as to make sure that, for the data we actually observe, $\mathbf{E}_{X,Y\sim P^*}[Y_i] = (1/2) \cdot 0.04 - (1/2) \cdot 0.04 = 0$; thus the Y-values will appear centred around 0, which is standard both in frequentist and Bayesian approaches to regression (for example, both Raftery et al. (1997) and Hastie et al. (2001) preprocess the data so that $\sum_{i=1}^{n} Y_i = 0$). Again, we discerned no difference in the results so did not include any further details.

Finally, the fourth experiment was included just to see what happens if, contrary to standard methodology, we apply the method to Y_i that are *not* (even approximately) centred. In this experiment, standard Bayes did not converge to the optimal model until after $n = 150$ as in the experiment of Section 3.5.3, but its risk and selected model orders were both smaller. The versions of Safe-Bayes worked well as before.