On the relationship between model uncertainty
and inferential/predictive uncertainty

BY DAVID DRAPER
Department of Applied Mathematics and Statistics, University of California,
Santa Cruz, USA

SUMMARY

Inference and prediction are usually based on a model relating known and
unknown quantities. The model itself typically contains features that are not
known with certainty. In good applied work this is often dealt with by expanding a
single “best” or standard model in ways that are sensitive to the model uncertainty,
in effect embedding the single model in a larger family of models of which it is a
special case. This represents a net gain in model uncertainty, and one might expect
that uncertainty about the desired inference or prediction would increase as a
result. But this need not be so; inferential/predictive uncertainty can either go up
or down as model uncertainty increases. In this paper I discuss a common example
in which inferential/predictive uncertainty decreases when model uncertainty goes
up.

Key words: Calibration; Distributional uncertainty; Location-scale models; Minimal Fisher
information; Modeling strategies; Serial correlation.

1. Introduction

We use models constantly in our inferential and predictive work. Most of the
time, one or more features of such models are arrived at after a search (typically
guided by the data) among possible modeling choices. Often we deal with the
uncertainty in the modeling process implied by this search by ignoring it: we
find the best model (in some sense) and carry out our inferences and predictions
conditional on this model, as if it were “correct.” Sometimes this produces well-
calibrated answers (for instance, when there is little ambiguity about sensible
modeling choices); sometimes it does not.

When we follow through to see if our inferential or predictive statements were
right about as often as we asserted they would be, we find most frequently that lack
of calibration is in the direction of insufficient conservatism—in other words, we
had more uncertainty than we were willing to admit. One possible explanation for
this is the underpropagation of modeling uncertainty just mentioned. One would
ordinarily expect that acknowledging greater modeling uncertainty—for example,
by making a standard modeling choice a special case of a larger family of models
and enlisting the help of the data to search among this larger family for a plausible
model, rather than just assuming the standard model is “correct”—would lead to
an increase in inferential or predictive uncertainty. But this need not be so. In
this paper I examine a common class of examples in which an increase in modeling uncertainty leads to a decrease in inferential uncertainty.

Figure 1 is a normal quantile-quantile plot of 100 weighings of a checkweight called NB10, made by workers at the U.S. National Bureau of Standards in 1962–63 under conditions as close to IID as possible (Freedman, Pisani, Purves & Adhikari, 1991; the units are micrograms below the nominal weight of 10g). How much does NB10 weigh? Figure 1 shows that it is plausible in answering this question to assume a symmetric location-scale model, but the form of the error distribution is less clear. The standard choice is Gaussian; with $\mu$ as the true weight and little or no prior information, the posterior distribution for $\mu$ based on an assumption of Gaussian errors is close to normal with mean 404.6 and standard deviation 0.65. But the solid line in Figure 1 is the target shape for the plot if the data were in fact Gaussian, and there is clear evidence for heavier tails. If one were to instead adopt (say) a $t_k$ model for the errors and treat the degrees of freedom $k$ as unknown—which corresponds to an increase in modeling uncertainty—it turns out that the posterior SD drops to 0.46, a 29% decrease. Is this a fluke or an example of a general phenomenon?

2. Increasing model uncertainty can decrease inferential uncertainty

Consider $Y_i, i = 1, \ldots, n$, IID (given $(\mu, \sigma)$) from a symmetric location-scale family $Y_i = \mu + \sigma e_i$, where the $e_i$ are assumed to have two finite moments and
Support \((-\infty, \infty)\). Suppose as above that the exact form of the density \(f(y|\mu, \sigma)\) of the \(Y_i\) is not known a priori, and interest focuses on the effect, on uncertainty assessments about \(\mu\), implied by this modeling uncertainty about \(f\). Without loss of generality take \(E(e_i) = 0\) and \(V(e_i) = 1\) so that \(\mu\) and \(\sigma^2 = V(Y_i)\) have the same meaning in all models to be compared.

For large \(n\) and little or no prior information about \(\mu\), uncertainty assessments for \(\mu\) will be based on the Fisher information for location. The posterior distribution for \(\mu\) is approximately normal with mean given by the maximum likelihood estimator (MLE) \(\hat{\mu}\) and variance \(\hat{I}^{-1}(f)\sigma^2/n\), where

\[
I(f) = \int_{-\infty}^{\infty} \frac{[f'(x)]^2}{f(x)} \, dx.
\]

(1)

Here \(\hat{I}\) is observed information and \(f\) is the (mean 0, SD 1) density of the normalized errors \(e_i\).

It is a fact (see Kagan, Linnik & Rao, 1973, and the Appendix below) that in this situation the off-the-shelf Gaussian choice for \(f\) is conservative: \(I(f)\) is minimized by the standard Gaussian distribution. This means that if one were to place the Gaussian in a larger family of densities \(f_\beta\) in which it is a special case \((\beta = 0, \text{say})\), and compare two modeling strategies,

- **Strategy 1:** I assert that the \(Y_i\) are Gaussian, which corresponds to placing all my prior mass on \(\beta = 0\); or

- **Strategy 2:** I express little prior knowledge of \(\beta\) and await the data's information about plausible values for it,

the second strategy would admit greater modeling uncertainty than the first, and yet—at least for large \(n\)—would lead to smaller uncertainty assessments for \(\mu\).

The \(t_k\) family mentioned above (with \(k > 2\); take \(\beta = 1/k\) to place the Gaussian at 0) is one instance of this model; it has been studied in location-scale problems by Lange, Little & Taylor (1989). Another example, this time including distributions with tails both heavier and lighter than those of the Gaussian, is the generalized power-exponential distributions examined by Box & Tiao (1962),

\[
f(x|\beta) = c \exp \left\{ -\frac{1}{2} |x|^{2/(1+\beta)} \right\},
\]

(2)

where \(c\) is a normalizing constant. Lange, Little & Taylor note, in both of these models, that in large samples scale \(\sigma\) and shape \(\beta\) may be confounded (although insisting that \(V(Y_i) = \sigma^2\) in all models permits direct examination of the effect of \(\beta\) on the posterior variance of \(\mu\) given the data), but that \((\sigma, \beta)\) will be uncorrelated with location \(\mu\), so that (at least with large \(n\)) one pays no price in Strategy 2, in uncertainty about \(\mu\), for one's uncertainty about \(\beta\).
**Model uncertainty**

![Graph](image)

Fig. 2. $I(f)$ for the $t_k$ family as a function of the degrees of freedom $k$.

<table>
<thead>
<tr>
<th>$k$</th>
<th>2.1</th>
<th>2.5</th>
<th>3.0</th>
<th>4.0</th>
<th>5.0</th>
<th>6.0</th>
<th>7.0</th>
<th>8.0</th>
<th>10.0</th>
<th>20.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c_k^*$</td>
<td>0.28</td>
<td>0.56</td>
<td>0.71</td>
<td>0.84</td>
<td>0.89</td>
<td>0.93</td>
<td>0.95</td>
<td>0.96</td>
<td>0.97</td>
<td>0.99</td>
</tr>
</tbody>
</table>

**Table 1: Values of $c_k^*$ for selected degrees of freedom $k$ in the $t_k$ model**

In the $t_k$ family with $k > 2$ $I(f)$ has the simple expression

$$I(t_k^*) = \frac{k(k+1)}{(k+3)(k-2)},$$

where $t_k^*$ is the scaled $t$-distribution with $k$ degrees of freedom and variance 1 (cf. Taylor, 1992). Figure 2 plots $I(t_k^*)$ for $k$ from 2.25 to 8, and Table 1 gives some values of the multiplier $c_k^* = I^{-\frac{1}{2}}(t_k^*)$ in the expression

$$\left( \text{posterior SD for } \mu \right)_{\text{assuming } t_k} = c_k^* \left( \text{posterior SD for } \mu \right)_{\text{assuming normality}}$$

for selected values of $k$. Table 1 indicates that noticeable decreases in inferential uncertainty from that implied by the Gaussian will only occur in the $t_k$ model in datasets for which the posterior distribution for $k$ concentrates most of its mass on $k < 5$ or so. For the NB10 data the MLE of $k$ is 3.0 (with a standard error of 0.86), and the observed decrease in inferential uncertainty when moving from Strategy 1 to 2 agrees closely with the relevant value from Table 1: the ratio of posterior SDs is $0.46/0.65 = 0.71$. 
In the power-exponential family no closed-form expression for $I(f)$ is available, but numerical comparisons may still be made. Figure 3 plots $I(f)$ over the relevant range of $\beta$, from near $-1$ (in the limit as $\beta \to -1$ one obtains the uniform distribution) to $+1$ (the double-exponential distribution), and Table 2 gives values of the multiplier $c_\beta^*$ analogous to $c_\beta^*$ in Table 1. In the limit as $\beta \to -1$, $I(f)$ becomes infinite, reflecting the fact that inferences about a location parameter with uniform errors are an order of magnitude (in powers of $n$ on the variance scale) more accurate than with $\beta > -1$.

That these are large-sample results is emphasized by Figure 4, reproduced with permission from Box & Tiao (1962). They fit the power-exponential model for various values of $\beta$ to Darwin's data on the heights of self- and cross-fertilized plants ($n = 15$). The rightmost solid curve is the posterior distribution for $\mu$ obtained by assuming the Gaussian ($\beta = 0$); other choices for $\beta$ from $-0.9$ to $+0.9$ are also displayed. The large-sample results presented here would lead one to expect that the Gaussian posterior would have the greatest variance, but with this dataset values of $\beta$ near $-0.3$ yield posteriors with slightly larger inferential uncertainty than that implied by the Gaussian. This phenomenon would vanish

Table 2: Values of $c_\beta^*$ for selected shape parameters $\beta$ in the power-exponential model

<table>
<thead>
<tr>
<th>$\beta$</th>
<th>$-0.9$</th>
<th>$-0.8$</th>
<th>$-0.6$</th>
<th>$-0.4$</th>
<th>$-0.2$</th>
<th>$0.0$</th>
<th>$0.2$</th>
<th>$0.4$</th>
<th>$0.6$</th>
<th>$0.8$</th>
<th>$1.0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c_\beta^*$</td>
<td>0.39</td>
<td>0.56</td>
<td>0.78</td>
<td>0.91</td>
<td>0.98</td>
<td>1.00</td>
<td>0.98</td>
<td>0.94</td>
<td>0.87</td>
<td>0.79</td>
<td>0.71</td>
</tr>
</tbody>
</table>

Fig. 3. $I(f)$ for the generalized power-exponential family.
with larger \( n \). (More precisely, in general with \( Y = (Y_1, \ldots, Y_n) \))

\[
V(\mu|Y) = V_\beta[E(\mu|Y, \beta)] + E_\beta[V(\mu|Y, \beta)].
\] (5)

As \( n \) increases both terms on the right-hand side go to 0 like \( \frac{c}{n} \), but simulations (not presented here) reveal that the numerator constant \( c \) for the first term is far smaller than that for the second term, which may be approximated by \( V(\mu|Y, \beta = \hat{\beta}_{M\text{LE}}) \leq V(\mu|Y, \beta = 0) \).

The conclusions here also apply to more general location-scale problems, including regression (see, e.g., Lange, Little & Taylor, 1989), and may be summarized with the statement that for whatever reason—historical accident or otherwise—from a model uncertainty point of view the default choice for the underlying error distribution in location problems is conservative. (This result is related to the maximum-entropy property of the Gaussian—see, e.g., Rao, 1973—although it is not straightforward to connect entropy and Fisher information for location algebraically.) This conclusion might have been anticipated in the case of long-tailed data from robustness considerations: the point of robust estimators is to down-weight outliers, and when it is appropriate to do so the result will be sharper inferential statements. It is perhaps less frequently noted that the same effect occurs with light-tailed data, and for a different reason: moving from the Gaussian
to the uniform involves crossing over from inferential uncertainty assessments of the form \( V(\mu|Y) = O(n^{-1}) \) to \( O(n^{-2}) \).

3. Comments

- One natural reaction to the results in Section 2 is to attribute the phenomenon to goodness-of-fit. For example, it might seem intuitively plausible to point out that moving from the Gaussian to the \( t_3 \) model for the NB10 data amounts to switching from a model that does not fit well to one that does, and one may expect to enjoy a decrease in inferential uncertainty as a result. But the fit of a model \( M \) and what may be termed the conditional inherent accuracy (for a given unknown quantity like \( \mu \)) given \( M \) are two different things. This may perhaps be seen most directly by running the NB10 experience in reverse: if the data had followed a Gaussian model and one had begun by instead assuming \( t_3 \) errors, embedding the \( t_3 \) model in the larger \( t_k \) framework would reveal that the Gaussian fit better, and yet the move from \( t_3 \) to Gaussian would involve at most a negligible decline in inferential uncertainty.

The Bayesian formulation of model uncertainty (see, e.g., Lindley, 1982) is clarifying: with \( y \) as an unknown quantity of interest and \( x \) as what is known,

\[
p(y|x) = \int p(y|x, M) p(M|x) dM. \tag{6}
\]

The second term in the product in this integral captures goodness-of-fit, the first conditional inherent accuracy; a retrospectively well-calibrated uncertainty assessment for \( y \) relies on both, and the two terms play different roles in such an assessment.

- The discussion so far has been in the context of inference; the situation for prediction is similar, in that increasing model uncertainty can either increase or decrease predictive uncertainty. The former behavior is much more common in practice than the latter; examples have been given, e.g., by Draper (1995). The latter behavior is possible but requires an implausible scenario, e.g.: an analyst first pulls a poorly-fitting model with low conditional inherent accuracy off the shelf, so to speak, and then, questioning its fit, embeds it in a larger set of models. If one or more of the new models fits better and also has higher conditional inherent accuracy, the result will be a net decrease in predictive uncertainty. It is hard to find practical examples of this phenomenon.

- Finally, it is probably worth noting that the pleasant conclusion of this paper about the effects of distributional model uncertainty on inferential uncertainty is rare: more typically, as with prediction the increase in modeling uncertainty arising from the embedding of a standard model in a larger class
of models leads to an increase in inferential uncertainty. A common example is the assumption of IID observations in location problems, where the measurement process may instead induce a serial correlation. As an instance of this consider an autoregressive process of order 1,

\[ Y_i = \rho Y_{i-1} + (1 - \rho)\mu + e_i, \quad i = 1, \ldots, n, \]  

where \(|\rho| < 1\) and (given \(\sigma_e^2\)) the \(e_i\) are IID \(N(0, \sigma_e^2)\). These assumptions specialize to the Gaussian location-scale model of Section 2 for \(\rho = 0\), and imply a likelihood function given, e.g., by Box & Jenkins (1976). As in Section 2, with large \(n\) and little prior information the posterior distribution for \((\mu, \sigma_e, \rho)\) given the \(Y_i\) is approximately normal with mean given by the MLE \((\hat{\mu}, \hat{\sigma}_e, \hat{\rho})\) and covariance matrix \(\hat{I}^{-1}\), where in this case the observed information matrix \(\hat{I}\) is approximately diagonal. The MLE for \(\mu\) is

\[ \hat{\mu} = \frac{n\bar{Y}_{1:n} - \hat{\rho}(n - 2)\bar{Y}_{2:n-1}}{n - \hat{\rho}(n - 2)}, \]  

where \(\bar{Y}_{i:j}\) denotes the mean of observations \(\{i, i+1, \ldots, j\}\); the posterior variance of \(\hat{\mu}\) is approximately

\[ \frac{\hat{\sigma}_e^2}{(1 - \hat{\rho})(n - \hat{\rho}(n - 2))}. \]

This may be compared with the value one would obtain by assuming \(\rho = 0\) and estimating \(\mu\) with the sample mean,

\[ V(\mu|Y, \rho = 0) = \frac{s^2}{n} = \frac{\hat{\sigma}_e^2}{n(1 - \hat{\rho}^2)}, \]  

where \(s^2\) is the sample variance of the \(Y_i\). For large \(n\) the multiplier \(c^*_\rho\) in the relation

\[ \left( \begin{array}{c} \text{posterior SD for } \mu \\ \text{assuming AR(1)} \end{array} \right) = c^*_\rho \left( \begin{array}{c} \text{posterior SD for } \mu \\ \text{assuming IID} \end{array} \right) \]

has the simple expression \(c^*_\rho = \sqrt{\frac{1 + \rho}{1 - \rho}}\). Some values of \(c^*_\rho\) are given in Table 3, which shows that in the empirically more common case of positive \(\rho\) (see, e.g., Scheffé, 1959, who quotes Student, 1927, as remarking that in years of careful laboratory work he had never observed a negative \(\rho\)) the posterior variance of \(\mu\) given the data is greater, often sharply so, than its value when \(\rho\) is incorrectly assumed to be 0.

Thus in practice the default IID modeling choice in location problems is often anything but conservative, and it is probably imprudent for IID to
Table 3: Values of $c^*_\rho$ for selected serial correlations $\rho$ in the AR(1) model

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>-0.9</th>
<th>-0.8</th>
<th>-0.6</th>
<th>-0.4</th>
<th>-0.2</th>
<th>0.0</th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c^*_\rho$</td>
<td>0.23</td>
<td>0.33</td>
<td>0.50</td>
<td>0.65</td>
<td>0.82</td>
<td>1.00</td>
<td>1.22</td>
<td>1.53</td>
<td>2.00</td>
<td>3.00</td>
<td>4.36</td>
</tr>
</tbody>
</table>

be the default, especially with small $n$. For example, with $n = 25$ and $\rho = 0.3$, for which an accurate uncertainty assessment for $\mu$ on the standard deviation scale is about 36% larger than its value obtained by pretending there is no serial correlation, $\text{SD}(\rho|Y) \approx \sqrt{\frac{1-\rho^2}{n}} \approx 0.19$, and observed values of $\hat{\rho}$ near their true value might easily be dismissed as spurious. Behaving as if $\rho = 0$ in this case would lower the actual success rate of nominal 90% (say) inferential statements to about 77%.

ACKNOWLEDGMENTS

I am grateful to Sander Greenland, Ken Lange, and Jeremy Taylor for comments on an earlier draft of this paper, and to Erich Lehmann for the Kagan, Linnik & Rao reference.

APPENDIX

Kagan, Linnik & Rao (1973) consider the problem of a single observation $Y$ from a location family $f(x - \mu)$ such that $f$ is continuously differentiable, $V(Y) = \sigma^2 < \infty$, and $|y|f(y) \to 0$ as $|y| \to \infty$, and show that, for fixed known $\sigma^2$, $I(f)$ is minimized by the $N(\mu, \sigma^2)$ distribution. The additional assumption of symmetry of $f$ about $\mu$ in the case of $n$ IID observations permits attention to focus on a more realistic model in which scale is also unknown, in such a way that $I(f)$ is still the relevant functional of $f$ in uncertainty assessment for $\mu$ (see, e.g., Lehmann, 1983). Kagan, Linnik & Rao’s proof is based on the Cauchy-Schwarz inequality; here is an alternative proof using the calculus of variations (also cf. Huber, 1981, who sketches a different variational proof motivated by quantum mechanics).

Alternative Proof that the Gaussian is Least Informative for Location. The problem may be placed into a standard calculus-of-variations framework employing Lagrange multipliers by seeking the extrema of

$$I(f) = \int_{x_1}^{x_2} g(x, f, f') dx, \quad g = (f')^2/f, \quad x_1 = -\infty, \quad x_2 = \infty$$

subject to the constraints

$$\int_{x_1}^{x_2} g_1(x, f, f') dx = 1, \quad g_1 = f,$$  \hspace{1cm} (13)

$$\int_{x_1}^{x_2} g_2(x, f, f') dx = 0, \quad g_2 = xf,$$  \hspace{1cm} (14)

$$\int_{x_1}^{x_2} g_3(x, f, f') dx = 1, \quad g_3 = x^2f,$$  \hspace{1cm} (15)
and boundary conditions \( f(x_1) = f(x_2) = 0 \). Defining

\[
G = g + \lambda_1 g_1 + \lambda_2 g_2 + \lambda_3 g_3,
\]

(16)

the Euler equation \( \frac{d}{dx} \left( \frac{\delta G}{\delta f} \right) - \frac{\delta G}{\delta f} = 0 \), in terms of the score function \( h = f' \), becomes

\[
2h' + h^2 = \lambda_1 + \lambda_2 x + \lambda_3 x^2.
\]

(17)

This is a Ricatti equation (see, e.g., Bender & Orszag, 1978), whose only analytic solution is the linear scores \( h(x) = \alpha + \beta x \)—in other words, the Gaussian \( f = \exp(\int h) \) is an extreme point. That the Gaussian minimizes \( I(f) \) can be seen by noting that \( I \) can be made arbitrarily large by letting \( f \) approach the uniform distribution (as in the Box & Tiao example above).

REFERENCES


