CREDIBILITY APPROXIMATIONS FOR BAYESIAN PREDICTION OF SECOND MOMENTS

By William S. Jewell*

University of California, Berkeley, California

AND

Rene Schnieper†

Department of Mathematics, ETH-Zentrum, Zurich

Abstract

Credibility theory refers to the use of linear least-squares theory to approximate the Bayesian forecast of the mean of a future observation; families are known where the credibility formula is exact Bayesian. Second-moment forecasts are also of interest, for example, in assessing the precision of the mean estimate. For some of these same families, the second-moment forecast is exact in linear and quadratic functions of the sample mean. On the other hand, for the normal distribution with normal-gamma prior on the mean and variance, the exact forecast of the variance is a linear function of the sample variance and the squared deviation of the sample mean from the prior mean. Bühlmann has given a credibility approximation to the variance in terms of the sample mean and sample variance.

In this paper, we present a unified approach to estimating both first and second moments of future observations using linear functions of the sample mean and two sample second moments; the resulting least-squares analysis requires the solution of a 3×3 linear system, using 11 prior moments from the collective and giving joint predictions of all moments of interest. Previously developed special cases follow immediately. For many analytic models of interest, 3-dimensional joint prediction is significantly better than independent forecasts using the "natural" statistics for each moment when the number of samples is small. However, the expected squared-errors of the forecasts become comparable as the sample size increases.

0. INTRODUCTION

In applications of Bayesian prediction, it is often difficult or extravagant to compute the entire predictive distribution; for example, the underlying likelihood and prior densities may be empirical, with only a few moments known with any degree of reliability. Also, the decision structure may depend only upon the first few moments, instead of upon the total shape of the predictive density. Finally,

* This research has been partially supported by the Air Force Office of Scientific Research (AFSC), USAF, under Grant AFOSR-81-0122 with the University of California.

† The research was supported by a grant from the Swiss National Science Foundation.

the need for repeated recalculation of forecasting formulae may argue for simple, easy-to-compute results.

A case in point is actuarial science, where the *fair premium* (predictive mean) is the point estimator of basic importance. To this may be added *fluctuation loadings*, which are given functions of the predictive second moment, the variance, or the standard deviation (see, e.g., GERBER, 1980). *Credibility theory* is the name given by actuaries to approximations of Bayesian predictors by formulae that are linear in the data, chosen to minimize quadratic Bayes risk. Thus, *credibility formulae* are linear least-squares predictors, and are akin to the classical estimators of that type, and to the linear filters used in electrical engineering.

The main emphasis of credibility theory thus far has been on approximating the predictive mean, under a wide variety of different model assumptions (see, *inter alia*, NORBERG, 1979; JEWELL, 1980). For many simple models used in practice, the linear credibility predictor of the mean is exactly the Bayesian conditional mean; in other situations, the credibility formula is usually quite robust.

1. BASIC MODEL AND NOTATION

Consider the usual Bayesian setup, in which a random observable, $\tilde{x}$, depends upon an unknown parameter, $\tilde{\theta}$, through a (discrete or continuous) *likelihood density*, $p(x|\theta)$. In the experiment of interest, $\tilde{\theta}$ is fixed at some unknown and unobservable value θ, but the parameter has a known *prior density*, $p(\theta)$. The conditional moments of $\tilde{x}$, given θ are:

$$m_i(\theta) = \mathscr{E}\{(\tilde{x})^i|\theta\}, \qquad (i = 1, 2, \ldots). \tag{1.1}$$

If we were to attempt to predict x prior to observing any data, and without knowing θ, we would have to use the *marginal density* of $\tilde{x}$, $p(x) = \mathscr{E}\{p(x|\tilde{\theta})\} = \int p(x|\theta)p(\theta)\, d\theta$, which has *prior-to-data (marginal) moments*:

$$m_i = \mathscr{E}\{m_i(\tilde{\theta})\} = \mathscr{E}\{(\tilde{x})^i\}. \tag{1.2}$$

For convenience in the sequel, we also define higher order cross-moments about the origin, such as:

$$m_{ij} = \mathscr{E}\{m_i(\tilde{\theta})m_j(\tilde{\theta})\}; \qquad m_{ijk} = \mathscr{E}\{m_i(\tilde{\theta})m_j(\tilde{\theta})m_k(\tilde{\theta})\}; \qquad \text{etc.}, \tag{1.3}$$

explicitly permitting the indices to be repeated, e.g., $m_{11} = \mathscr{E}\{(m_1(\tilde{\theta}))^2\}$. Thus, from the four conditional moments $\{m_i(\theta); i = 1, 2, 3, 4\}$ we can form *eleven* marginal moments of order four or less:

$$\mathscr{M} = \{m_1; m_2, m_{11}; m_3, m_{21}, m_{111}; m_4, m_{31}, m_{22}, m_{211}, m_{1111}\}. \tag{1.4}$$

Three central moments of order two deserve special symbols:

$$e = \mathscr{E}\mathscr{V}\{\tilde{x}|\tilde{\theta}\} = m_2 - m_{11}; \qquad d = \mathscr{V}\mathscr{E}\{\tilde{x}|\tilde{\theta}\} = m_{11} - m_1^2; \tag{1.5}$$

$$c = \mathscr{V}\{\tilde{x}\} = e + d = m_2 - m_1^2,$$

https://doi.org/10.2143/AST.15.2.2015022 Published online by Cambridge University Press

where double operators and their corresponding operands are to be interpreted "inside-out". Central moments of higher order can also be defined.

Now suppose that n independent observations, $\mathcal{D} = \{x_1, x_2, \ldots, x_n\}$, are drawn from the same likelihood density, $p(x|\theta)$, with θ fixed, but unknown. From Bayes' law, the *posterior-to-data parameter density* is:

$$p(\theta|\mathcal{D}) \propto \prod_{u=1}^{n} p(x_u|\theta)p(\theta), \tag{1.6}$$

and knowing this enables us to calculate the *posterior-to-data predictive density* for the next observation, $\tilde{x}_{n+1}$, as:

$$p(x_{n+1}|\mathcal{D}) = \int p(x_{n+1}|\theta)p(\theta|\mathcal{D})\, d\theta. \tag{1.7}$$

This is, in fact, the predictive density for *any* future observation, assuming that θ does not change, and that no more information is available. From our viewpoint, given $\mathcal{D}$, the $\{\tilde{x}_{n+1}, \tilde{x}_{n+2}, \tilde{x}_{n+3}, \ldots\}$ are *exchangeable random variables*; for example, the joint predictive density of $(\tilde{x}_{n+1}, \tilde{x}_{n+2})$ is:

$$p(x_{n+1}, x_{n+2}|\mathcal{D}) = \int p(x_{n+1}|\theta)p(x_{n+2}|\theta)p(\theta|\mathcal{D})\, d\theta. \tag{1.8}$$

(1.7) and (1.8) also have predictive moments analogous to (1.2), (1.3):

$$m_1(\mathcal{D}) = \mathscr{E}\{\tilde{x}_{n+1}|\mathcal{D}\}; \qquad m_2(\mathcal{D}) = \mathscr{E}\{\tilde{x}_{n+1}^2|\mathcal{D}\};$$
$$m_{11}(\mathcal{D}) = \mathscr{E}\{\tilde{x}_{n+1}\tilde{x}_{n+2}|\mathcal{D}\}; \qquad \text{etc.}, \tag{1.9}$$

that can, in principle, be calculated exactly; however, analytic solutions almost always require that $p(x|\theta)$ and $p(\theta)$ be chosen from among *natural conjugate families*. We now consider how approximate results can be obtained for the predictive moments in (1.9).

2. CREDIBLE MEAN FORMULAE

Consider first the problem of calculating or approximating $m_1(\mathcal{D})$. For many years, actuaries (in a different terminology) have been assuming that this "experience-rated premium" was linear in the data, as summarized in the sample mean, $\bar{x} = \sum x_u/n$ (it is clear from exchangeability arguments that each of the samples, x_u, should be weighted the same). Using heuristic reasoning, they argued for the approximation:

$$m_1(\mathcal{D}) = \mathscr{E}\{\tilde{x}_{n+1}|\mathcal{D}\} \approx f_1^*(\mathcal{D}) = (1 - z_1)m_1 + z_1\bar{x}, \tag{2.1}$$

i.e., the forecast, $f_1^*(\mathcal{D})$, should be a convex combination of the "manual" (prior) mean, m_1, and the "experienced" mean, $\bar{x}$. The "credibility factor", z_1, that weights these two means is, they argued:

$$z_1 = \frac{n}{n_{01} + n}, \tag{2.2}$$

https://doi.org/10.2143/AST.15.2.2015022 Published online by Cambridge University Press

where the "credibility time constant", n_{01}, was to be chosen empirically. This heuristic formula, used for many years, was considerably strengthened by BÜHLMANN (1967), who showed that the *best* linear formula (in the least-squares sense) to approximate the predictive mean $m_1(\mathcal{D})$ was precisely the credibility formula, $f_1^*(\mathcal{D})$, but with the time constant computed explicitly from the prior second moments:

$$n_{01} = \frac{e}{d} = \frac{m_2 - m_{11}}{m_{11} - m_1^2} = \frac{m_2 - m_1^2}{m_{11} - m_1^2} - 1 = \frac{c}{d} - 1. \tag{2.3}$$

Thus, a credibility predictor to approximate $\mathscr{E}\{\tilde{x}_{n+1}|\mathcal{D}\}$ needs only the first three components of (1.4), $\{m_1; m_2, m_{11}\}$, instead of the complete shape of the prior and likelihood densities.

In fact, Bailey, Mayerson, and others had already shown in the 1950s that (2.1), (2.2), (2.3) was exactly $m_1(\mathcal{D})$ for many "natural" $p(x|\theta)$ and $p(\theta)$ used in Bayesian modelling. JEWELL (1974*a*) then showed that, if the likelihood were a member of the *simple exponential family* (for which $\bar{x}$ is the sufficient statistic) over some space χ:

$$p(x|\theta) = \frac{a(x)\, e^{-\theta x}}{\gamma(\theta)}, \qquad (x \in \chi) \tag{2.4}$$

and $p(\theta)$ were the *natural conjugate prior* to (2.4):

$$p(\theta) = \frac{[\gamma(\theta)]^{-n_{01}}\, e^{-\theta x_{01}}}{g(n_{01}, x_{01})}, \qquad (\theta \in \Theta) \tag{2.5}$$

over the maximal range Θ for which the normalization $g(n_{01}, x_{01})$ is finite, then, under a certain regularity condition (JEWELL, 1975), (2.1) is *exact*, with the hyperparameters n_{01} in (2.3) and (2.5) identical, and with $x_{01} = m_1 n_{01}$.

A simple argument also shows that, if the exponent θx in (2.4) is replaced by, say, $\theta t(x)$, then the credibility form (2.1) again provides an exact prediction for $\mathscr{E}\{t(\tilde{x}_{n+1})|\mathcal{D}\}$ as a linear combination of the prior mean of the statistic, $\mathscr{E}\{t(\tilde{x})\}$, and the sample mean of the statistic $\sum t(x_u)/n$, with appropriate redefinition of (2.3). For this and other reasons that will become clearer below, we feel that (2.1) is a robust formula in most cases.

3. EXACT RESULTS FOR SECOND MOMENTS

We now consider exact results that are known for the predictive moments, $m_1(\mathcal{D})$, $m_2(\mathcal{D})$, and $m_{11}(\mathcal{D})$, concentrating on the most-studied case, the simple exponential family.

It is well known that the combination (2.4), (2.5) is *closed under sampling*, so that, posterior-to-data $\mathcal{D}$, the hyperparameters in (2.5) are replaced by:

$$n_{01} \leftarrow n_{01} + n; \qquad x_{01} \leftarrow x_{01} + n\bar{x}. \tag{3.1}$$

https://doi.org/10.2143/AST.15.2.2015022 Published online by Cambridge University Press

Since $m_1 = x_{01}/n_{01}$, it follows that the updated first moment is:

$$\mathscr{E}\{\tilde{x}_{n+1}|\mathscr{D}\} = m_1(\mathscr{D}) = \frac{x_{01}+n\bar{x}}{n_{01}+n} = (1-z_1)m_1 + z_1\bar{x}, \tag{3.2}$$

which is simply (2.1), (2.2). It is also clear that the marginal second moments must also involve only n_{01} and x_{01}, and that the predictive second moments must be a function of only the sufficient statistic, $\bar{x}$, but no further statement can be made about dependencies in general. JEWELL (1974*a*) tabulates $d = d(n_{01}, x_{01})$ for six of the examples given below, whence one can easily get $e = n_{01}d(n_{01}, x_{01})$, $c = (n_0+1)d(n_{01}, x_{01})$, and hence:

$$m_2(\mathscr{D}) = (n_{01}+1+n)d(n_{01}+n, x_{01}+n\bar{x}) + m_1^2(\mathscr{D}) \tag{3.3}$$

$$m_{11}(\mathscr{D}) = d(n_{01}+n, x_{01}+n\bar{x}) + m_1^2(\mathscr{D}),$$

and, from these, the updated versions of the central moments c and d:

$$\mathscr{V}\{\tilde{x}_{n+1}|\mathscr{D}\} = m_2(\mathscr{D}) - m_1^2(\mathscr{D}), \tag{3.4}$$

$$\mathscr{C}\{\tilde{x}_{n+1}; \tilde{x}_{n+2}|\mathscr{D}\} = m_{11}(\mathscr{D}) - m_1^2(\mathscr{D}).$$

EXAMPLE 1. Let $p(x|\theta)$ be Bernoulli (π) and $p(\pi)$ be Beta $(x_{01}, n_{01}-x_{01})$, $(\theta = \ln(\pi^{-1}-1))$, then:

$$d(n_{01}, x_{01}) = \frac{x_{01}(n_{01}-x_{01})}{n_{01}^2(n_{01}+1)}. \tag{3.5}$$

EXAMPLE 2. Let $p(x|\theta)$ be Geometric (π), and $p(\pi)$ be Beta $(x_{01}, n_{01}+1)$, $(\theta = \ln \pi^{-1})$, then:

$$d(n_{01}, x_{01}) = \frac{x_{01}(x_{01}+n_{01})}{n_{01}^2(n_{01}-1)}. \tag{3.6}$$

EXAMPLE 3. Let $p(x|\theta)$ be Poisson (π) and $p(\pi)$ be Gamma (x_{01}, n_{01}), $(\theta = \ln \pi^{-1})$, then:

$$d(n_{01}, x_{01}) = \frac{x_{01}}{n_{01}^2}. \tag{3.7}$$

EXAMPLE 4. Let $p(x|\theta)$ be Exponential (θ), and $p(\theta)$ be Gamma $(n_{01}+1, x_{01})$, then:

$$d(n_{01}, x_{01}) = \frac{x_{01}^2}{n_{01}^2(n_{01}-1)}. \tag{3.8}$$

EXAMPLE 5. Let $p(x|\theta)$ be Normal (π, s_0^2), s_0^2 known, and $p(\pi) =$ Normal $(x_{01}/n_{01}, s_0^2/n_{01})$, $\theta = -\pi/s_0^2$, then:

$$d(n_{01}, x_{01}) = \frac{s_0^2}{n_{01}} \quad \text{(independent of } x_{01}\text{)}. \tag{3.9}$$

https://doi.org/10.2143/AST.15.2.2015022 Published online by Cambridge University Press

Thus, in these examples from JEWELL (1974*a*), $d(n_{01}, x_{01})$, $m_1(\mathcal{D})$, $m_2(\mathcal{D})$, and $m_{11}(\mathcal{D})$ are all linear, quadratic, or constant in x_{01} and hence in $\bar{x}$ as well.

MORRIS (1982) refers to simple exponential likelihoods (2.4) in which $m_2(\theta)$ is at most a quadratic polynomial in $m_1(\theta)$ as QVF-NEF; he shows that the only members of this family are the five examples above, plus Example 6, below, plus all of the related members found through linear translation and convolution (Binomial, Pascal, Gamma, etc.).

EXAMPLE 6. The last member of this group is the Hyperbolic Secant density:

$$p(x|\theta) = (\cos\theta)\frac{e^{-\theta x}}{2\cosh(\pi x/2)}, \qquad \chi = [-\infty, +\infty], \Theta = \left[-\frac{\pi}{2}, +\frac{\pi}{2}\right]$$

for which

$$d(n_{01}, x_{01}) = \frac{x_{01}^2 + n_{01}^2}{n_{01}^2(n_{01}-1)}. \tag{3.10}$$

This likelihood seems to be useful only in certain random-walk problems.

We should mention also that it is easy to construct members of the simple exponential family in which the mean is a complicated function of $\bar{x}$, for example, by truncating the range of any of the above distributions.

To obtain dependency on $\bar{x}$ *and* other statistics, we must turn to two-parameter families, of which the most popular is the normal density with both the mean, $\tilde{\mu}$, and the precision, $\tilde{\omega}$, as random quantities.

EXAMPLE 7. Let

$$p(x|\theta) = p(x|\mu, \omega) = \text{Normal}(\mu, \omega^{-1})$$

$$p(\omega) = \text{Gamma}\left(\alpha, \frac{1}{2}\left[x_{02} - \left(\frac{x_{01}^2}{n_{01}}\right)\right]\right),$$

and

$$p(\mu|\omega) = \text{Normal}\left(\frac{x_{01}}{n_{01}}, (n_{01}\omega)^{-1}\right),$$

with α, x_{01}, x_{02}, and n_{01} given hyperparameters. This family is closed under sampling, with updating:

$$\alpha \leftarrow \alpha + \frac{n}{2}; \qquad n_{01} \leftarrow n_{01} + n; \tag{3.11}$$

$$x_{01} \leftarrow x_{01} + \sum x_u; \qquad x_{02} \leftarrow x_{02} + \sum x_u^2;$$

from which we find that (3.2) again holds, and that:

$$d = d(n_{01}, \alpha, x_{01}, x_{02}) = \frac{1}{(2\alpha - 2)}\left[\left(\frac{x_{02}}{n_{01}}\right) - \left(\frac{x_{01}}{n_{01}}\right)^2\right] = \frac{1}{(2\alpha - 2)}c, \tag{3.12}$$

where $c = e + d$ is the prior variance.

https://doi.org/10.2143/AST.15.2.2015022 Published online by Cambridge University Press

For this example, we see that the updating will give exact second-moment predictors that are quadratic in $\bar{x}$ and linear in $\overline{x^2} = \sum x_u^2/n$. Because the normal case is so important to least-squares approximations, we also give the exact results corresponding to (3.4) in terms of the *sample variance*, $s^2 = n^{-1}\sum (x_u - \bar{x})^2$, the sample mean, $\bar{x}$, the prior marginal variance, c, and the credibility factor, z_1:

$$\mathcal{V}\{\tilde{x}_{n+1}|\mathcal{D}\} = (n_{01}+1+n)\,\mathcal{C}\{\tilde{x}_{n+1}; x_{n+2}|\mathcal{D}\} = \left(\frac{n_{01}+1+n}{2\alpha-2+n}\right) \times\left[\left(\frac{2\alpha-2}{n_{01}+1}\right)(1-z_1)c + z_1 s^2 + z_1(1-z_1)(m_1-\bar{x})^2\right]. \tag{3.13}$$

An important simplification occurs if the "natural" choice $2\alpha = n_{01}+3$ is made; note that this does not significantly restrict the choice of the 2-parameter Gamma, but does mean that there are only three distinct hyperparameters in all. (3.13) then simplifies to a generalized credibility formula:

$$\begin{aligned}\mathcal{V}\{\tilde{x}_{n+1}|\mathcal{D}\} &= (n_{01}+1+n)\,\mathcal{C}\{\tilde{x}_{n+1}; \tilde{x}_{n+2}|\mathcal{D}\} \\ &= (1-z_1)c + z_1 s^2 + z_1(1-z_1)(m_1-\bar{x})^2.\end{aligned} \tag{3.14}$$

This result is not new, but rearrangement into credibility form first appeared in JEWELL (1974*a*). The equivalent multidimensional formula appeared in JEWELL (1974*b*, 1983).

(3.14) is, in fact, equivalent to:

$$\mathcal{E}\{\tilde{x}_{n+1}^2|\mathcal{D}\} = m_2(\mathcal{D}) = (1-z_1)m_2 + z_1\overline{x^2}, \tag{3.15}$$

that is, the predictive second moment is exactly in credibility form with the *same* credibility factor as in (2.2), with obvious adjustments to the prior mean and sample mean.

4. LEAST-SQUARES THEORY AND MULTIDIMENSIONAL CREDIBILITY

We now take a temporary detour to display some general results from multidimensional credibility theory that will be used in the next section. Suppose we have a vector-valued version of the Bayesian model of Section 1, in which samples $\mathcal{D} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n\}$ of a vector-valued random variable, $\tilde{\boldsymbol{y}}$, are to be used to predict a random vector, $\tilde{\boldsymbol{w}}$. If we approximate $\mathcal{E}\{\tilde{\boldsymbol{w}}|\mathcal{D}\}$ by a linear function of the vector-valued sample mean, $\bar{\boldsymbol{y}} = \sum \boldsymbol{y}_n/n$, least-squares theory then shows that the best (vector-valued) predictor is:

$$\boldsymbol{f}(\mathcal{D}) = (\mathcal{E}\{\tilde{\boldsymbol{w}}\} - \boldsymbol{Z}\mathcal{E}\{\tilde{\boldsymbol{y}}\}) + \boldsymbol{Z}\bar{\boldsymbol{y}} \approx \mathcal{E}\{\tilde{\boldsymbol{w}}|\mathcal{D}\}, \tag{4.1}$$

where $\boldsymbol{Z}$ is a matrix of appropriate dimensions given by the solution of the *normal system* of equations:

$$\boldsymbol{Z}\boldsymbol{\mathcal{C}}\{\tilde{\boldsymbol{y}}; \tilde{\boldsymbol{y}}\} = \boldsymbol{\mathcal{C}}\{\tilde{\boldsymbol{w}}; \tilde{\boldsymbol{y}}\} \tag{4.2}$$

($\boldsymbol{\mathcal{C}}$ is the matrix covariance operator).

https://doi.org/10.2143/AST.15.2.2015022 Published online by Cambridge University Press

Now suppose $\tilde{\boldsymbol{w}}$ is actually a future observation of the same random vector, $\tilde{\boldsymbol{y}}$, say $\tilde{\boldsymbol{w}} = \tilde{\boldsymbol{y}}_{n+1}$. Then these equations become:

$$\boldsymbol{f}(\mathscr{D}) = (\boldsymbol{I} - \boldsymbol{Z})\boldsymbol{m} + \boldsymbol{Z}\bar{\boldsymbol{y}} \approx \mathscr{E}\{\tilde{\boldsymbol{y}}_{n+1} | \mathscr{D}\} \tag{4.1'}$$

where $\boldsymbol{I}$ is the $n \times n$ unit matrix, $\boldsymbol{Z}$ is the square solution of:

$$\boldsymbol{Z}\left(\boldsymbol{D} + \frac{1}{n}\boldsymbol{E}\right) = \boldsymbol{D}, \tag{4.2'}$$

$\boldsymbol{m}$ is the prior mean vector, obtained from the conditional mean vector:

$$\boldsymbol{m}(\theta) = \mathscr{E}\{\tilde{\boldsymbol{y}} | \theta\}; \tag{4.3}$$

$$\boldsymbol{m} = \mathscr{E}\{\tilde{\boldsymbol{y}}\} = \mathscr{E}\{\boldsymbol{m}(\tilde{\theta})\};$$

and $\boldsymbol{E}$ and $\boldsymbol{D}$ are the two components of within-risk and between-risk covariance, respectively:

$$\boldsymbol{E} = \mathscr{E}\mathscr{C}\{\tilde{\boldsymbol{y}}; \tilde{\boldsymbol{y}} | \tilde{\theta}\}; \qquad \boldsymbol{D} = \mathscr{C}\{\boldsymbol{m}(\tilde{\theta}); \boldsymbol{m}(\tilde{\theta})\}; \tag{4.4}$$

$$\mathscr{C} = \mathscr{C}\{\tilde{\boldsymbol{y}}; \tilde{\boldsymbol{y}}\} = \boldsymbol{E} + \boldsymbol{D}.$$

Thus, the credibility formula of Section 2 extends directly to the multi-dimensional case, with a credibility *matrix*, $\boldsymbol{Z}$, mixing the prior mean, $\boldsymbol{m}$, and the experience mean, $\bar{\boldsymbol{y}}$. The analogy is complete if we assume $\boldsymbol{D}$ has an inverse and rearrange (4.2'):

$$\boldsymbol{Z} = n\boldsymbol{D}(\boldsymbol{E} + n\boldsymbol{D})^{-1} = n(n\boldsymbol{I} + \boldsymbol{N})^{-1}; \qquad \boldsymbol{N} = \boldsymbol{E}\boldsymbol{D}^{-1}, \tag{4.5}$$

where $\boldsymbol{N}$ is now a matrix of time constants. Further details on this extension may be found in JEWELL (1974*b*).

The accuracy of any forecast $\boldsymbol{f}(\mathscr{D})$ for $\boldsymbol{y}_{n+1}$ is measured by the diagonal terms of the expected squared-error matrix:

$$\boldsymbol{\Phi} = \mathscr{E}\{[\tilde{\boldsymbol{y}}_{n+1} - \boldsymbol{f}(\tilde{\mathscr{D}})][\tilde{\boldsymbol{y}}_{n+1} - \boldsymbol{f}(\tilde{\mathscr{D}})]'\}; \tag{4.6}$$

note that the expectation is over all possible joint values of $(\tilde{\boldsymbol{y}}_{n+1}; \tilde{\mathscr{D}})$. However, since the latter are independent, given θ, $\boldsymbol{\Phi}$ can be decomposed into:

$$\begin{aligned} \boldsymbol{\Phi} &= \mathscr{E}\{[\tilde{\boldsymbol{y}}_{n+1} - \boldsymbol{m}(\tilde{\theta})][\tilde{\boldsymbol{y}}_{n+1} - \boldsymbol{m}(\tilde{\theta})]'\} + \mathscr{E}\{[\boldsymbol{f}(\tilde{\mathscr{D}}) - \boldsymbol{m}(\tilde{\theta})][\boldsymbol{f}(\tilde{\mathscr{D}}) - \boldsymbol{m}(\tilde{\theta})]'\} \\ &= \boldsymbol{E} + \boldsymbol{\Psi}, \quad \text{say,} \end{aligned} \tag{4.6'}$$

where we see the portion of the mse due to the inherent fluctuation of the observable, and the mse due to the approximation of the true mean, $\boldsymbol{m}(\theta)$, by the approximation, $\boldsymbol{f}(\mathscr{D})$.

We know that the minimum values of the diagonal terms for $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ are attained by picking the Bayesian predictive mean, $\boldsymbol{m}(\mathscr{D}) = \mathscr{E}\{\tilde{\boldsymbol{y}}_{n+1} | \mathscr{D}\}$, which, in general, leads to a nonlinear regression on the data. With a linear forecast (4.1'), (4.2'), it is easy to show that, for any n,

$$\boldsymbol{\Psi} = \boldsymbol{E}[\boldsymbol{E} + n\boldsymbol{D}]^{-1}\boldsymbol{D} = [\boldsymbol{I} - \boldsymbol{Z}]\boldsymbol{D} = \boldsymbol{D}[\boldsymbol{I} - \boldsymbol{Z}']. \tag{4.7}$$

In most cases of interest, all terms of Ψ will approach zero as n approaches infinity for any forecast, so that all forecasts are asymptotically equivalent; in the linear case, it usually happens because $\boldsymbol{Z}$ approaches $\boldsymbol{I}$ (see also (8.6)). Fortunately, a linear predictor also usually has small mean-squared-error also for moderate n, even though $f(\mathscr{D})$ is not exactly the Bayesian predictive mean.

We now examine the use of (4.1'), (4.2') as an *approximation* for our original one-dimensional problem of estimating second moments.

5. ORGANIZING THE LEAST-SQUARES COMPUTATIONS

We return to the main problem of organizing credibility approximations $f_2(\mathscr{D})$ and $f_{11}(\mathscr{D})$ for $m_2(\mathscr{D})$ and $m_{11}(\mathscr{D})$ of arbitrary distributions. In view of the exact results in Section 3, it seems reasonable to restrict the statistics to be used to linear and quadratic functions of the data; however, there are several different ways to select statistics of this type. After a great deal of experimentation, the authors have found that the choices that give the simplest and clearest results are the "natural" first and second moments about the origin:

$$(5.1)\quad t_1(\mathscr{D})=\frac{1}{n}\sum x_u=\bar{x};\quad t_2(\mathscr{D})=\frac{1}{n}\sum x_u^2=\overline{x^2};\quad t_{11}(\mathscr{D})=\frac{1}{n(n-1)}\sum\sum_{u\neq v}x_u x_v;$$

for $n\geqslant 2$. In other words, we set $\bar{y}=[t_1(\mathscr{D});\, t_2(\mathscr{D});\, t_{11}(\mathscr{D})]'=\boldsymbol{t}(\mathscr{D})$ in (4.1). Note that this choice implicitly includes $(\bar{x})^2=[t_2(\mathscr{D})+(n-1)t_{11}(\mathscr{D})]/n$, as well as the sample variance $s^2=[(n-1)/n][t_2(\mathscr{D})-t_{11}(\mathscr{D})]$.

As predictands, we can get all the forecasts of interest *simultaneously* by setting $\tilde{\boldsymbol{w}}=[\tilde{x}_{n+1};\, \tilde{x}_{n+1}^2;\, \widetilde{x_{n+1}x_{n+2}}]'$. Then, to get $\boldsymbol{Z}$ in (4.1), we need only to compute the means in (4.1) and the two covariance matrices in (4.2). This approach is thus similar to credibility regression modelling; see HACHEMEISTER (1974).

For the means, we find easily:

$$(5.2)\qquad \mathscr{E}\{\tilde{\boldsymbol{y}}|\theta\}=\mathscr{E}\{\tilde{\boldsymbol{w}}|\theta\}=\boldsymbol{m}(\theta)=[m_1(\theta);\, m_2(\theta);\, m_{11}(\theta)]',$$

$$\mathscr{E}\{\tilde{\boldsymbol{y}}\}=\mathscr{E}\{\tilde{\boldsymbol{w}}\}=\boldsymbol{m}=[m_1;\, m_2;\, m_{11}]'.$$

(Note that $m_{11}(\theta)=m_1^2(\theta)$.) Computation of the covariance terms is straightforward, but tedious, as they involve all 11 moments of (1.4); we find, for $n\geqslant 2$:

$$(5.3)\qquad \mathscr{C}\{\tilde{\boldsymbol{y}};\tilde{\boldsymbol{y}}\}=\boldsymbol{D}+\frac{1}{n}\boldsymbol{E}(n);\qquad \mathscr{C}\{\tilde{\boldsymbol{w}};\tilde{\boldsymbol{y}}\}=\boldsymbol{D};$$

where $\boldsymbol{D}$ and $\boldsymbol{E}(n)$ are *new* matrices, *analogous* to the matrices in (4.4), but otherwise unrelated. Explicitly, we find:

$$(5.4)\qquad \boldsymbol{D}=\begin{bmatrix} m_{11}-m_1^2 & m_{21}-m_2m_1 & m_{111}-m_{11}m_1 \\ & m_{22}-m_2^2 & m_{211}-m_2m_{11} \\ \text{(symmetric)} & & m_{1111}-m_{11}^2 \end{bmatrix}$$

https://doi.org/10.2143/AST.15.2.2015022 Published online by Cambridge University Press

and

$$\boldsymbol{E}(n) = \boldsymbol{E}_\infty + \frac{1}{n-1}\boldsymbol{E}_1; \tag{5.5}$$

where

$$\boldsymbol{E}_\infty = \begin{bmatrix} m_2 - m_{11} & m_3 - m_{21} & 2(m_{21} - m_{111}) \\ & m_4 - m_{22} & 2(m_{31} - m_{211}) \\ \text{(symmetric)} & & 4(m_{211} - m_{1111}) \end{bmatrix} \tag{5.6}$$

and

$$\boldsymbol{E}_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 2(m_{22} - 2m_{211} + m_{1111}) \end{bmatrix}. \tag{5.7}$$

Once these have been computed, the credibility matrix $\boldsymbol{Z}$ is the solution of:

$$\boldsymbol{Z}\left(\boldsymbol{D} + \frac{1}{n}\boldsymbol{E}(n)\right) = \boldsymbol{D}; \tag{5.8}$$

and the vector forecast $\boldsymbol{f}(\mathscr{D}) = [f_1(\mathscr{D}); f_2(\mathscr{D}); f_{11}(\mathscr{D})]'$ is given by:

$$\boldsymbol{f}(\mathscr{D}) = (\boldsymbol{I} - \boldsymbol{Z})\boldsymbol{m} + \boldsymbol{Z}\boldsymbol{t}(\mathscr{D}), \tag{5.9}$$

which should be compared with (4.1′), (4.2′).

6. INDEPENDENT FORECASTS USING NATURAL STATISTICS

Before examining the various aspects of the three-dimensional forecast (5.9), it is of interest to consider first how the one-dimensional result (2.1) would generalize if second-moment forecasts were made only in terms of their "natural" statistics, i.e., if the solution to $\boldsymbol{Z}$ were forced to be diagonal. We find:

$$m_2(\mathscr{D}) = \mathscr{E}\{\tilde{x}_{n+1}^2 | \mathscr{D}\} \approx f_2^*(\mathscr{D}) = (1 - z_2)m_2 + z_2 t_2(\mathscr{D}); \tag{6.1}$$

$$z_2 = \frac{n}{n_{02} + n}; \qquad n_{02} = \frac{m_4 - m_{22}}{m_{22} - m_2^2},$$

and, for $n \geq 2$,

$$m_{11}(\mathscr{D}) = \mathscr{E}\{\tilde{x}_{n+1}\tilde{x}_{n+2} | \mathscr{D}\} \approx f_{11}^*(\mathscr{D}) = (1 - z_{11})m_{11} + z_{11}t_{11}(\mathscr{D}); \tag{6.2}$$

$$z_{11} = \frac{n}{n + n_{011}(n)}; \qquad n_{011}(n) = \frac{4(m_{211} - m_{1111}) + \dfrac{2}{n-1}(m_{22} - 2m_{211} + m_{1111})}{m_{1111} - m_{11}^2}.$$

These are to be compared with (2.1), (2.2), (2.3), which, of course, still hold for the first-moment forecast. (Note that asterisks distinguish the independent forecasts f_1^*, f_2^*, and f_{11}^* from the corresponding components of the joint forecast

https://doi.org/10.2143/AST.15.2.2015022 Published online by Cambridge University Press

f, and that z_{11} in (6.2) is *not* the (1, 1)st component of $\boldsymbol{Z}$ in (4.2′).) We will return to analysis of independent forecasts in Section 9, after analyzing the asymptotic behaviors of (5.8), (5.9).

7. LIMITING BEHAVIOR OF THE JOINT FORECAST

The analogy with (4.5) is complete if we can assume that $\boldsymbol{D}$ has an inverse (but see Section 8), for then (5.8) can be rearranged into:

$$\boldsymbol{Z} = n(n\boldsymbol{I} + \boldsymbol{N}(n))^{-1}; \qquad \boldsymbol{N}(n) = \boldsymbol{E}(n)\boldsymbol{D}^{-1}; \tag{7.1}$$

so that we now have a time-varying "time constant":

$$\boldsymbol{N}(n) = \boldsymbol{N}_\infty + \frac{1}{n-1}\boldsymbol{N}_1; \qquad \boldsymbol{N}_\infty = \boldsymbol{E}_\infty\boldsymbol{D}^{-1}; \qquad \boldsymbol{N}_1 = \boldsymbol{E}_1\boldsymbol{D}^{-1}. \tag{7.2}$$

Because of the simple form of $\boldsymbol{E}_1$, it follows that $\boldsymbol{N}_1$ induces correction terms only in the third row of $\boldsymbol{Z}$, that is, in making a prediction of $m_{11}(\mathscr{D})$; furthermore, this correction term vanishes rapidly with increasing n. In fact, one can easily make the asymptotic expansion:

$$\boldsymbol{Z} = \boldsymbol{I} - \frac{1}{n}\boldsymbol{N}_\infty + \frac{1}{n^2}(\boldsymbol{N}_\infty^2 - \boldsymbol{N}_1) + O\left(\frac{1}{n^3}\right), \qquad (n \to \infty), \tag{7.3}$$

so that the correction term $\boldsymbol{N}_1$ introduces changes only of order n^{-2} or smaller.

More importantly, we see that, if $\boldsymbol{D}^{-1}$ exists, then $\boldsymbol{Z} \to \boldsymbol{I}$ as $n \to \infty$; thus our three-dimensional forecasts become "fully credible", that is, the forecasts $f_i(\mathscr{D})$ are ultimately given essentially by their own natural statistics, $t_i(\mathscr{D})$ $(i = 1, 2, 11)$. Asymptotically, then, the joint predictions of Section 5 will be undistinguishable from the independent forms of the last section.

8. REDUCED-RANK D MATRIX

It would be an unusual model for which $\boldsymbol{E}_\infty$ did not have an inverse; however, it is theoretically possible that $\boldsymbol{D}^{-1}$ does not exist. In several of the special cases examined below, $\boldsymbol{D}$ is of rank two because of the close asymptotic relationship between $t_2(\mathscr{D})$ and $t_{11}(\mathscr{D})$. Thus, to perform the inversion in (5.8), we must use the well-known matrix inversion formula which states that, if $\boldsymbol{a}$ and $\boldsymbol{b}$ are $n \times k$ matrices of rank k $(k \leqslant n)$, then:

$$[\boldsymbol{I}_n + \boldsymbol{a}\boldsymbol{b}']^{-1} = \boldsymbol{I}_n - \boldsymbol{a}[\boldsymbol{I}_k + \boldsymbol{b}'\boldsymbol{a}]^{-1}\boldsymbol{b}'. \tag{8.1}$$

If $\boldsymbol{D}$ is of rank two so that, for example, $\boldsymbol{d}^3 = a_{31}\boldsymbol{d}^1 + a_{32}\boldsymbol{d}^2$, where $\boldsymbol{d}^i$ is the ith row of $\boldsymbol{D}$ $(i = 1, 2, 3)$, then $\boldsymbol{D}$ can be written:

$$\boldsymbol{D} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ a_{31} & a_{32} \end{bmatrix} \begin{bmatrix} \boldsymbol{d}^1 \\ \boldsymbol{d}^2 \end{bmatrix} = \boldsymbol{A}\boldsymbol{D}^{12}, \quad \text{say.} \tag{8.2}$$

https://doi.org/10.2143/AST.15.2.2015022 Published online by Cambridge University Press

We find from (8.1) that:

$$\boldsymbol{Z} = \boldsymbol{A}\left[\boldsymbol{\Delta}(n) + \frac{1}{n}\boldsymbol{I}_2\right]^{-1}\boldsymbol{D}^{12}\boldsymbol{E}(n)^{-1}, \tag{8.3}$$

where $\boldsymbol{\Delta}(n)$ is the full-rank 2×2 matrix:

$$\boldsymbol{\Delta}(n) = \boldsymbol{D}^{12}\boldsymbol{E}(n)^{-1}\boldsymbol{A}. \tag{8.4}$$

The important implication of these results is that, when $\boldsymbol{D}$ is of rank two, the limit of $\boldsymbol{Z}(n)$ as $n \to \infty$ is *not* $\boldsymbol{I}_3$, but is:

$$\boldsymbol{Z}(\infty) = \boldsymbol{A}\boldsymbol{\Delta}(\infty)^{-1}\boldsymbol{D}^{12}\boldsymbol{E}_\infty^{-1}. \tag{8.5}$$

Thus, in this case, the $t_i(\mathscr{D})$ are never "fully credible" for the $f_i(\mathscr{D})$, and dependence upon the prior means, m_i $(i = 1, 2, 11)$, and other moments, persists. In fact, $\boldsymbol{Z}(\infty)$ is not even diagonal!

Nevertheless, from (8.3), (8.4), it is easy to show that:

$$(\boldsymbol{I} - \boldsymbol{Z})\boldsymbol{D} = \boldsymbol{A}\left[\boldsymbol{I}_2 - \left[\boldsymbol{I}_2 + \frac{1}{n}\boldsymbol{\Delta}^{-1}(n)\right]^{-1}\right]\boldsymbol{D}^{12}, \tag{8.6}$$

so that, from (9.1), it follows that $\boldsymbol{\Psi}$ will vanish even in this case!

9. COMPARISON OF THREE-DIMENSIONAL FORECASTS WITH INDEPENDENT FORECASTS USING NATURAL STATISTICS

One can show that the second term of (4.7):

$$\boldsymbol{\Psi} = [\boldsymbol{I} - \boldsymbol{Z}]\boldsymbol{D}, \tag{9.1}$$

is still valid when using definitions (5.4) through (5.8). The diagonal terms of this matrix are the mean-squared approximation errors of the (joint) forecasts, call them mse $(f_i(\mathscr{D}))$ $(i = 1, 2, 11)$.

However, there are several arguments in favor of replacing the forecasts (5.9) with their independent counterparts (2.1), (6.1), and (6.2), such as avoiding the numerical inversion of a 3×3 matrix, and requiring only seven moments from the list (1.4). If we let d_{ii} denote the diagonal terms of $\mathscr{D}$, we can show that mean-squared approximation errors of the independent forecast are:

$$\text{mse}\,(f_i^*(\mathscr{D})) = (1 - z_i)d_{ii}, \quad (i = 1, 2); \qquad \text{mse}\,(f_{11}^*(\mathscr{D})) = (1 - z_{11})d_{33}. \tag{9.2}$$

Each of these mse is larger, in general, than the corresponding diagonal terms in (9.1).

However, by making asymptotic expansions for $n \to \infty$, one can show that the corresponding dominant terms in n^{-1} are identical, and that, in the limit, mse $(f_i(\mathscr{D}))$ and mse $(f_i^*(\mathscr{D}))$ differ only by terms of order n^{-2}. Thus, for a large number of samples, we expect little difference in the approximation errors of joint and independent forecasts.

10. PREDICTIVE VARIANCE AND FORECAST ERROR

There are two second-order central moments of special interest: the predictive variance,

$$v(\mathcal{D}) = \mathcal{V}\{\tilde{x}_{n+1}|\mathcal{D}\} = \mathcal{E}\{[\tilde{x}_{n+1} - m_1(\mathcal{D})]^2\} = m_2(\mathcal{D}) - m_1^2(\mathcal{D}) \tag{10.1}$$

and the posterior-to-data mean-squared-forecast-error:

$$\phi(\mathcal{D}) = \mathcal{E}\{[\tilde{x}_{n+1} - f_1(\mathcal{D})]^2|\mathcal{D}\} = v(\mathcal{D}) + [f_1(\mathcal{D}) - m_1(\mathcal{D})]^2. \tag{10.2}$$

If the $f_1(\mathcal{D})$ and $f_2(\mathcal{D})$ obtained previously are exact, then both of the expressions are identical and equal to $f_2(\mathcal{D}) - f_1^2(\mathcal{D})$. If credibility is only an approximation, then this latter expression may still be a good approximation to $v(\mathcal{D})$ (note that we now may be using quadratic functions of the data in $f_1^2(\mathcal{D})$). Comparing $\phi(\mathcal{D})$ and $v(\mathcal{D})$ requires knowing how closely the credibility for the mean approximates the Bayesian predictive mean.

We can proceed a bit further if we rewrite the mean-squared-forecast-error as:

$$\phi(\mathcal{D}) = m_2(\mathcal{D}) - m_{11}(\mathcal{D}) + \mathcal{E}\{[f_1(\mathcal{D}) - m_1(\tilde{\theta})]^2|\mathcal{D}\}, \tag{10.3}$$

and approximate the first two terms by $f_2(\mathcal{D}) - f_{11}(\mathcal{D})$. The third term cannot be estimated directly; however, by averaging once more over all prior values of $\mathcal{D}$, we obtain $\mathcal{E}\{[f_1(\mathcal{D}) - m_1(\tilde{\theta})]^2\} = \text{mse}[f_1(\mathcal{D})]$, which is a natural by-product of our analyses. In summary, then, we would use the following estimators for (10.1) and (10.2):

$$v(\mathcal{D}) \approx f_2(\mathcal{D}) - f_1^2(\mathcal{D}); \tag{10.4}$$

$$\phi(\mathcal{D}) \approx f_2(\mathcal{D}) - f_{11}(\mathcal{D}) + \text{mse}[f_1(\mathcal{D})]. \tag{10.5}$$

BÜHLMANN (1970, p. 100) also considers the problem of estimating the predictive variance. He breaks $v(\mathcal{D})$ into a "variance part" and a "fluctuation part", which, in our notation, are:

$$v(\mathcal{D}) = [m_2(\mathcal{D}) - m_{11}(\mathcal{D})] + [m_{11}(\mathcal{D}) - m_1^2(\mathcal{D})], \tag{10.6}$$

the posterior-to-data version of $c = e + d$ (cf. (1.5)). He then approximates the first part by a one-dimensional credibility forecast using the unbiased sample variance, $\Sigma^2 = (n-1)^{-1} \Sigma (x_u - \bar{x})^2 = t_2(\mathcal{D}) - t_{11}(\mathcal{D})$, i.e.,

$$e(\mathcal{D}) = m_2(\mathcal{D}) - m_{11}(\mathcal{D}) \approx (1 - z_e)(m_2 - m_{11}) + z_e \Sigma^2. \tag{10.7}$$

The credibility factor, z_e, is a complicated function of n, but, by making the simplifying assumption of a "normal excess" (e.g., the kurtosis of $p(x|\theta)$ is that of the normal density for every θ), he obtains a simplified form, $z_e = (n-K)/(n-3)$, where K is a complicated ratio of marginal moments.

The second factor,

$$d(\mathcal{D}) = m_{11}(\mathcal{D}) - m_1^2(\mathcal{D}) = \mathcal{E}\{[m_1(\mathcal{D}) - m_1(\tilde{\theta})]^2|\mathcal{D}\} \tag{10.8}$$

is approximated by: first, replacing $m_1(\mathcal{D})$ by $f_1^*(\mathcal{D})$, and second, averaging over

https://doi.org/10.2143/AST.15.2.2015022 Published online by Cambridge University Press

all prior values of $\mathscr{D}$, obtaining: $d(\mathscr{D}) \approx \text{mse}[f_1^*(\mathscr{D})]$, giving finally:

$$v(\mathscr{D}) \approx (1-z_e)e + z_e[t_2(\mathscr{D}) - t_{11}(\mathscr{D})] + \text{mse}[f_1^*(\mathscr{D})]. \tag{10.9}$$

With our extended use of these statistics, we could presumably improve Bühlmann's analysis by arguing in the same way that:

$$v(\mathscr{D}) \approx f_2(\mathscr{D}) - f_{11}(\mathscr{D}) + \text{mse}[f_1(\mathscr{D})]. \tag{10.10}$$

However, this is exactly the approximation (10.5) for $\phi(\mathscr{D})$, which must be larger than $v(\mathscr{D})$ if mean credibility is not exact! So, we would still prefer (10.4) for the estimate of the variance.

The difficult-to-estimate term, $d(\mathscr{D})$, is, in fact, the posterior-to-data predictive covariance, $\mathscr{E}\{[\tilde{x}_{n+1} - m_1(\mathscr{D})][\tilde{x}_{n+2} - m_1(\mathscr{D})]|\mathscr{D}\}$, which we know must vanish with n as the true value of θ is identified. For instance, with the simple-exponential family of Section 2, we have $d(\mathscr{D}) = e(\mathscr{D})d/(e+nd)$ or $v(\mathscr{D}) = e(\mathscr{D})[1+(d/(e+nd))]$. And, in the general case, if $f_1(\mathscr{D})$ is close to $m_1(\mathscr{D})$, then we know that the average (preposterior) value of $d(\mathscr{D})$ is $\text{mse}[f_1(\mathscr{D})]$, which probably vanishes like $\text{mse}[f_1^*(\mathscr{D})] = ed/(e+nd)$.

So, in short, we doubt if the accuracy issues raised here are important in any realistic application, and expect the errors in using (10.4), (10.5) to be of the same order of magnitude as the errors in the underlying predictions $f(\mathscr{D})$.

11. NUMERICAL EXAMPLES

It should be remembered that important simplifications often occur in $\boldsymbol{D}$ and $\boldsymbol{E}(n)$ for the usual analytic forms assumed for the likelihood and the prior. For instance, where the likelihood is normal, with possibly random mean and variance, we have:

$$m_3(\theta) = 3v(\theta)m(\theta) + m^3(\theta);$$

$$m_4(\theta) = 3v^2(\theta) + 6v(\theta)m^2(\theta) + m^4(\theta);$$

where

$$m(\theta) = m_1(\theta) \quad \text{and} \quad v(\theta) = m_2(\theta) - m_1^2(\theta).$$

From this, we see that all eleven moments in (1.4) can be expressed in terms of moments and cross-moments of $m(\theta)$ (up to order 4) and $v(\theta)$ (up to order 2).

The likelihoods introduced in Examples 1 through 6 of Section 3 have been characterized by Morris (1982) as the natural exponential families with quadratic variance functions, i.e., the variance is at most a quadratic function of the mean. From this, it follows that, for this family, the components of $\boldsymbol{m}(\theta)$ in (5.2) are linearly-dependent functions of the parameter, and that $\boldsymbol{D}$ is singular. For example, if the likelihood is Poisson (π), then $\boldsymbol{m}(\pi) = [\pi; \pi + \pi^2; \pi^2]'$.

We now consider three numerical examples that illustrate these ideas; in all examples, the joint credibility forecasts are *exactly* the Bayesian mean forecast, for all n. (However, we have *not* introduced this prior knowledge into the numerical calculations below!)

EXAMPLE A. Consider Example 7 from Section 3, the Normal (μ, ω^{-1}), with Normal-Gamma prior, and with the following hyperparameters: $x_{01} = 10$; $n_{01} = 10$; $x_{02} = 21$; $\alpha = 6.5$. Note that we have chosen $\alpha = (n_{01}+3)/2$ so that the predictive second moment will be in credibility form (3.14), (3.15).

Numerically, we find the eleven marginal moments to be:

$$\mathcal{M} = \{1, 2.1, 1.1, 4.3, 2.3, 1.3, 12.037, 5.0033, 5.1033, 2.7589, 1.6367\},$$

and the variance components are:

$$d = \mathcal{E}\{(n_0\tilde{\omega})^{-1}\} = 0.1; \qquad e = \mathcal{E}\{\tilde{\omega}^{-1}\} = 1.0.$$

The independent time constants of Section 6 are

$$n_{01} = n_{02} = 10, \text{ and } n_{011}(n) = 10.52 + (5.73/(n-1)).$$

For $n = 2$, 10, 100, and 10,000, Figure 1 shows the credibility matrix $\boldsymbol{Z}$ for the three-dimensional forecasts of $(4.2)'$, together with their corresponding mean-squared errors, the diagonal terms from (9.1). Also shown are the corresponding independent forecast factors of (2.2), (6.1), (6.2), arranged in matrix format for easy visual comparison (and thus making (5.9) a general forecast formula, even with $\boldsymbol{Z}$ diagonal); the corresponding mse's are from (9.2).

	Joint Prediction				Independent Prediction			
n	$\boldsymbol{Z}$			mse $f_1(\mathcal{D})$ $f_2(\mathcal{D})$ $f_{11}(\mathcal{D})$	$\begin{pmatrix} z_1 & 0 & 0 \\ 0 & z_2 & 0 \\ 0 & 0 & z_{11} \end{pmatrix}$			mse $f_1^*(\mathcal{D})$ $f_2^*(\mathcal{D})$ $f_{11}^*(\mathcal{D})$
2	0.16667	0.00000	0.00000	0.08333	0.16667	0.00000	0.00000	0.08333
	0.00000	0.16667	0.00000	0.57778	0.00000	0.16667	0.00000	0.57778
	0.25641	0.02564	0.01282	0.35840	0.00000	0.00000	0.10959	0.37991
10	0.50000	0.00000	0.00000	0.05000	0.50000	0.00000	0.00000	0.05000
	0.00000	0.50000	0.00000	0.34667	0.00000	0.50000	0.00000	0.34667
	0.47619	0.04762	0.21429	0.21862	0.00000	0.00000	0.47265	0.22500
100	0.90909	0.00000	0.00000	0.00909	0.90909	0.00000	0.00000	0.00909
	0.00000	0.90909	0.00000	0.06303	0.00000	0.90909	0.00000	0.06303
	0.16380	0.01638	0.81081	0.04061	0.00000	0.00000	0.90433	0.04082
10,000	0.99900	0.00000	0.00000	0.00010	0.99900	0.00000	0.00000	0.00010
	0.00000	0.99900	0.00000	0.00069	0.00000	0.99900	0.00000	0.00069
	0.00200	0.00020	0.99780	0.00045	0.00000	0.00000	0.99895	0.00045

FIGURE 1. Numerical results for Example A, Normal-Gamma-Normal.

We remark that:

(1) Because of previous results, $m_1(\mathcal{D}) = f_1(\mathcal{D}) = f_1^*(\mathcal{D})$ and $m_2(\mathcal{D}) = f_2(\mathcal{D}) = f_2^*(\mathcal{D})$, since $\alpha = n_{01}+3$. Thus, the upper part of $\boldsymbol{Z}$ is diagonal, with $z_{11} = z_{22}$

https://doi.org/10.2143/AST.15.2.2015022 Published online by Cambridge University Press

equal to the independent prediction factors. We also know that $m_{11}(\mathcal{D}) = f_{11}(\mathcal{D})$ (but not equal to $f_{11}^*(\mathcal{D})$, in general); here it is of interest to see how long a heavier weight is attached to $t_1(\mathcal{D})$ instead of the natural statistic, $t_{11}(\mathcal{D})$.

(2) The mse's for the first two components are, of course, the same for both predictions. As might be expected, predicting second moments gives larger mse's than the mse for $f_1(\mathcal{D})$; however, the relative rate of decrease with n is about the same. Furthermore, there is only about a 6% increase in mse for using $f_{11}^*(\mathcal{D})$ over the exact $f_{11}(\mathcal{D})$.

(3) Both credibility factors approach the identity matrix as n approaches infinity, as the statistics in $\boldsymbol{t}(\mathcal{D})$ become "fully credible".

EXAMPLE B. Consider Example 4 from Section 3, the Exponential (θ), with Gamma prior, with hyperparameters: $x_{01} = 10$; $n_{01} = 10$. The marginal moments are:

$$\mathcal{M} = \{1, 2.2222, 1.1111, 8.3333, 2.7778, 1.3889, 47.619,$$
$$11.905, 7.9365, 3.9683, 1.9841\}.$$

The hyperparameters were chosen to make $m_1 = 1.0$ and $n_{01} = 10.0$, as in Example A, but now, due to the change in distributions, we have $n_{02} = 13.24$, and $n_{011}(n) = 10.59 + (5.29/(n-1))$.

Figure 2 shows again the results for $n = 2$, 10, 100, and 10,000, in a format similar to that of fig. 1.

Notice the following:

(1) As in Example A, $m_1(\mathcal{D}) = f_1(\mathcal{D}) = f_1^*(\mathcal{D})$; however, now both $f_2(\mathcal{D})$ and $f_{11}(\mathcal{D})$ use all three statistics, particularly $t_1(\mathcal{D})$ and $t_{11}(\mathcal{D})$. Now, as $n \to \infty$, we find the surprising result that $2t_{11}(\mathcal{D})$ is the preferred predictor for $m_2(\mathcal{D})$, rather than the "natural" estimator, $t_2(\mathcal{D})$; they both have the same expectation, but the former has smaller variance.

(2) In fact, we can make the following stronger statements. As a consequence of the exponential assumption only, $m_2(\theta) = 2m_1^2(\theta)$ for all θ, so that $m_2(\mathcal{D}) = 2m_{11}(\mathcal{D})$ for *any* prior. Assumption of a Gamma prior makes both predictions linear functions of $\boldsymbol{t}(\mathcal{D})$, and, in fact, we see from fig. 2 that $z_{2j} = 2z_{3j}$ ($j = 1, 2, 3$), so that $f_2(\mathcal{D}) = 2f_{11}(\mathcal{D})$ for all $\mathcal{D}$!

(3) The mse's for independent predictions of the two second moments are, of course, larger than in the joint predictions, and worst for $f_2^*(\mathcal{D})$, as it is forced into using $t_2(\mathcal{D})$, rather than $t_{11}(\mathcal{D})$ as its sole predictor. This gives a relative degradation which climbs about 20%, but, at the same time, all mse's are decreasing with n at about the same relative rate. Substituting $t_{11}(\mathcal{D})$ for the "natural" predictor of $f_2^*(\mathcal{D})$ would, of course, reduce the mse to four times that of $f_{11}^*(\mathcal{D})$, which at its worst value ($n = 2$), is only about 5% larger than the joint prediction.

(4) The non-convergence of $\boldsymbol{Z}$ to the identity matrix is the consequence of the previously-discussed fact that $\boldsymbol{D}$ is singular. However, since $m_2 = 2m_{11}$, $2t_1(\mathcal{D})$ is ultimately "fully credible" as $n \to \infty$, i.e., no dependence upon prior moments remains in $f_2(\mathcal{D})$ in the limit. We have already proven this directly in (8.6).

EXAMPLE C. Consider Example 3 from Section 3, the Poisson (π), with Gamma prior, and hyperparameters: $X_{01} = 10$; $n_{01} = 10$. The marginal moments are:

$$\mathcal{M} = \{1, 2.1, 1.1, 5.62, 2.42, 1.32, 18.336, 6.776, 5.456, 3.036, 1.716\}.$$

The hyperparameters were again chosen to make $m_1 = 1.0$ and $n_{01} = 10.0$, but now $n_{02} = 12.31$, and $n_{011}(n) = 10.43 + (4.35/(n-1))$.

Figure 3 tabulates the results for $n = 2$, 10, 100, and 10,000 in the same format as previous examples.

	Joint Prediction				Independent Prediction			
n	Z			mse $f_1(\mathcal{D})$ $f_2(\mathcal{D})$ $f_{11}(\mathcal{D})$	$\begin{pmatrix} z_1 & 0 & 0 \\ 0 & z_2 & 0 \\ 0 & 0 & z_{11} \end{pmatrix}$			mse $f_1^*(\mathcal{D})$ $f_2^*(\mathcal{D})$ $f_{11}^*(\mathcal{D})$
2	0.16667	0.00000	0.00000	0.09259	0.16667	0.00000	0.00000	0.09259
	0.60606	0.03030	0.03030	2.52525	0.00000	0.13127	0.00000	2.60465
	0.30303	0.01515	0.01515	0.63131	0.00000	0.00000	0.11184	0.66573
10	0.50000	0.00000	0.00000	0.05556	0.50000	0.00000	0.00000	0.05556
	1.05263	0.05263	0.47368	1.54553	0.00000	0.43038	0.00000	1.70786
	0.52632	0.02632	0.23684	0.38638	0.00000	0.00000	0.47222	0.39560
100	0.90909	0.00000	0.00000	0.01010	0.90909	0.00000	0.00000	0.1010
	0.33361	0.01668	1.65138	0.28728	0.00000	0.88312	0.00000	0.35044
	0.16681	0.00834	0.82569	0.07182	0.00000	0.00000	0.90382	0.07209
10,000	0.99900	0.00000	0.00000	0.00011	0.99900	0.00000	0.00000	0.00011
	0.00399	0.00020	1.99601	0.00317	0.00000	0.99868	0.00000	0.00396
	0.00200	0.00010	0.99800	0.00079	0.00000	0.00000	0.99894	0.00079

FIGURE 2. Numerical results for Example B, Gamma-Exponential.

We notice that:

(1) As in Example A and B, the first moment uses only $t_1(\mathcal{D})$, but the second moments use all three statistics, with $t_2(\mathcal{D})$ playing a decreasingly important role. In contrast to Example B, however, we now find that, as $n \to \infty$, $t_1(\mathcal{D}) + t_{11}(\mathcal{D})$ is the preferred predictor for $m_2(\mathcal{D})$, rather than $t_2(\mathcal{D})$.

(2) This is a consequence of the assumption that the likelihood is Poisson, for then $m_2(\theta) = m_1(\theta) + m_1^2(\theta)$ for all θ, so that $m_2(\mathcal{D}) = m_1(\mathcal{D}) + m_{11}(\mathcal{D})$ for *any* prior. It is the assumption of the Gamma prior that makes predictions using only linear functions of $t(\mathcal{D})$ exact, and in fig. 3 we can see that, in fact, $z_{2j} = z_{1j} + z_{3j}$ $(j = 1, 2, 3)$, so that $f_2(\mathcal{D}) = f_1(\mathcal{D}) + f_{11}(\mathcal{D})$ for all $\mathcal{D}$!

(3) The mse's follow the pattern of Example B, with the mse of $f_2^*(\mathcal{D})$ becoming progressively relatively worse than its joint counterpart. Here, however, to improve the prediction error, one would probably have to include both $t_1(\mathcal{D})$ and $t_{11}(\mathcal{D})$, as it is not clear that just one of the latter would be an improvement over using

https://doi.org/10.2143/AST.15.2.2015022 Published online by Cambridge University Press

	Joint Prediction				Independent Prediction			
n	Z			mse $f_1(\mathscr{D})$ $f_2(\mathscr{D})$ $f_{11}(\mathscr{D})$	$\begin{pmatrix} z_1 & 0 & 0 \\ 0 & z_2 & 0 \\ 0 & 0 & z_{11} \end{pmatrix}$			mse $f_1^*(\mathscr{D})$ $f_2^*(\mathscr{D})$ $f_{11}^*(\mathscr{D})$
2	0.16667	0.00000	0.00000	0.08333	0.16667	0.00000	0.00000	0.08333
	0.45833	0.01389	0.01389	0.87472	0.00000	0.13973	0.00000	0.89985
	0.29167	0.01389	0.01389	0.42472	0.00000	0.00000	0.11917	0.44570
10	0.50000	0.00000	0.00000	0.05000	0.50000	0.00000	0.00000	0.05000
	1.02500	0.02500	0.22500	0.52850	0.00000	0.44816	0.00000	0.57723
	0.52500	0.02500	0.22500	0.25850	0.00000	0.00000	0.47806	0.26410
100	0.90909	0.00000	0.00000	0.00909	0.90909	0.00000	0.00000	0.00909
	1.08264	0.00826	0.81818	0.09691	0.00000	0.89036	0.00000	0.11468
	0.17355	0.00826	0.81818	0.04782	0.00000	0.00000	0.90515	0.04799
10,000	0.99900	0.00000	0.00000	0.00010	0.99900	0.00000	0.00000	0.00010
	1.00110	0.00010	0.99790	0.00107	0.00000	0.99877	0.00000	0.00129
	0.00210	0.00010	0.99790	0.00053	0.00000	0.00000	0.99896	0.00053

FIGURE 3. Numerical results for Example C, Gamma-Poisson.

just $t_2(\mathscr{D})$. Furthermore, neither of the other statistics would ever become "fully credible" as $n \to \infty$, as they are not individually equal in expectation to m_2, only in sum. Clearly, the best single statistic to use for $m_2(\mathscr{D})$ in the Poisson case is $t_1(\mathscr{D}) + t_{11}(\mathscr{D})$.

12. COMPUTATIONAL STRATEGIES; CONCLUSION

The last two examples show that some care must be exercised if one wishes to make independent forecasts where $p(x|\theta)$ is assumed to be in the QVF-NEF family, remembering that this also includes (fixed numbers of) convolutions of Examples 1-6, such as the Negative Binomial with fixed shape parameter. One can, of course, use the combination of "natural" statistics appropriate to the assumed likelihood. This is particularly important when we also assume that the natural conjugate prior is appropriate.

On the other hand, for an arbitrary prior, the moments will not be linear functions of the statistics, so that all positions of Z would be non-zero anyway, as would also be the case if all moments were from empirical studies. In these cases, Z would approach the identity matrix as $n \to \infty$, and we expect that the independent forecasts (2.2), (6.1), (6.2) would be equally good (or equally bad) as the joint forecasts. Clearly, more computational experience is needed in making this decision.

The great advantage of the joint forecast is that it can always be used if $n \geq 2$, and, if there is a tendency for certain combinations of statistics to dominate, it

https://doi.org/10.2143/AST.15.2.2015022 Published online by Cambridge University Press

will be revealed automatically. Of course, if $n = 1$, we are forced to use only $t_1(\mathscr{D}) = x_1$ and $t_2(\mathscr{D}) = x_1^2$; the predictive power will be weak anyway, in most practical cases.

In summary, we have presented an easily implemented three-dimensional credibility formula that simultaneously approximates the first and second moments of the Bayesian predictive density. While this approach requires eleven prior moments from the collective, this calculation is simplified when familiar analytic forms are assumed for the likelihood. Previous work has shown that the credibility mean is exact in $t_1(\mathscr{D})$ for a wide class of likelihoods and priors in which the sample mean is the sufficient statistic; here we have shown that the second-moment credibility predictions are also exact for five widely-used likelihoods and their natural conjugate priors, when using the three "natural" statistics in $\boldsymbol{t}(\mathscr{D})$.

For these and other reasons, we believe that these linear prediction formulae will turn out to be robust in other cases where the distributions are empirical, or where the exact predictions are known to be non-linear in the data. We suspect also that, in most cases, it will also be reasonable to use the simplified, independent forecasts, paying due attention to the remarks above. The authors look forward to hearing from those who apply this approach to actual prediction problems.

REFERENCES

BÜHLMANN, H. (1967) Experience Rating and Credibility. *Astin Bulletin* **4**, 199-207.

BÜHLMANN, H. (1970) *Mathematical Methods in Risk Theory.* Springer-Verlag: New York.

GERBER, H. (1980) Introduction to the Mathematical Theory of Risk, The Huebner Foundation, Philadelphia, PA.

HACHEMEISTER, C. A. (1974) Credibility for Regression Models with Application to Trend. In Credibility Theory and Applications, P. M. Kahn (ed.), Academic Press: New York.

JEWELL, W. S. (1974*a*) Credible Means are Exact Bayesian for Simple Exponential Families. *Astin Bulletin* **7**, 237-269.

JEWELL, W. S. (1974*b*) Exact Multidimensional Credibility. *Mitteilungen der Vereinigung schweizerischer Versicherungsmathematiker* **74**, 193-214.

JEWELL, W. S. (1975) Regularity Conditions for Exact Credibility. *Astin Bulletin* **8**, 336-341.

JEWELL, W. S. (1980) Models in Insurance: Paradigms, Puzzles, Communications, and Revolutions. *Transactions 21st International Congress of Actuaries*, Zürich and Lausanne, Volume S, 87-141.

JEWELL, W. S. (1983) Enriched Multinormal Priors Revisited. *Journal of Econometrics* **23**, 5-35.

MORRIS, C. N. (1982) Natural Exponential Families with Quadratic Variance Functions. *The Annals of Statistics* **10**, 65-80.

NORBERG, R. (1979) The Credibility Approach to Experience Rating. *Scandinavian Actuarial Journal*, 181-221.