SORT 34 (1) January-June 2010, 3-20

Small-sample inference about variance and its transformations*

N. T. Longford

SNTL Statistics Research and Consulting

Abstract

We discuss minimum mean squared error and Bayesian estimation of the variance and its common transformations in the setting of normality and homoscedasticity with small samples, for which asymptotics do not apply. We show that permitting some bias can be rewarded by greatly reduced mean squared error. We apply borderline and equilibrium priors. The purpose of these priors is to reduce the onus on the expert or client to specify a single prior distribution that would capture the information available prior to data inspection. Instead, the (parametric) class of all priors considered is partitioned to subsets that result in the preference for different actions. With the family of conjugate inverse gamma priors, this Bayesian approach can be formulated in the frequentist paradigm, describing the prior as being equivalent to additional observations.

MSC: 62F10 and 62F15

Keywords: Borderline prior, equilibrium prior, expected loss, gamma distribution, mean squared error, plausible prior

1. Introduction

We consider the problem in which a small random sample from a normal distribution, $\mathcal{N}(\mu, \sigma^2)$, is observed and we would like to estimate the variance σ^2 or its transformation, such as σ, $1/\sigma^2$ or $1/\sigma$, or to know whether σ^2 exceeds (or falls short of) a specified threshold σ_{R}^2. We study two approaches: minimum mean squared error (MSE) estimation, to which we refer as *efficient* estimation, and application of (Bayesian) priors. We use only the conjugate family of priors, both for computational simplicity and

* N. T. Longford, SNTL and Departament d'Economia i Empresa, Universitat Pompeu Fabra, Ramon Trias Fargas 25–27, 08005 Barcelona, Spain; email: NTL@sntl.co.uk.

Received: July 2009

Accepted: February 2010

because their representation in terms of additional observations can greatly aid the process of eliciting prior information from an expert. We find a frequentist interpretation of the (Bayesian) posterior distribution which makes the Bayesian approach accessible to frequentist analysis. In the motivating problem, two alternative actions, A and B, are contemplated; A is appropriate when $\sigma^2 < \sigma^2_{\mathrm{R}}$ and B when $\sigma^2 > \sigma^2_{\mathrm{R}}$. There is some prior information, but the analyst's client is either not available or elicitation of a single prior from him or her is unlikely to be constructive.

We are concerned only with analysis of small samples. In large samples, asymptotics apply and maximum likelihood (ML) estimation is satisfactory. The prior information has a diminishing impact and a nonlinear transformation of a parameter is estimated by the same transformation of the ML estimator of the parameter. In small samples, the prior has a non-trivial impact, and efficiency is not maintained by nonlinear transformations. Therefore, efficient (frequentist) estimation of σ^2, σ, $1/\sigma^2$ and $1/\sigma$ are, in principle, distinct problems, and the prior for a Bayesian analysis has to be selected with integrity and care.

The next section deals with efficient estimation. Section 3.1. introduces borderline priors and Section 3.2. equilibrium priors and the related solutions. Equilibrium priors incorporate the losses due to making an incorrect decision (choosing A when $\sigma^2 > \sigma^2_{\mathrm{R}}$ or B when $\sigma^2 < \sigma^2_{\mathrm{R}}$). The perspective of Section 2. is entirely frequentist, while Section 3. might appear at first as entirely Bayesian, exploiting prior information. However, the Bayesian analysis has a frequentist interpretation, with the prior regarded as additional observations. The concluding section summarises the proposed methods.

2. Efficient small-sample estimation

Suppose $y_1, \ldots, y_n$ is a random sample from a normal distribution $\mathcal{N}(\mu, \sigma^2)$. The variance σ^2 is commonly estimated by the corrected mean squares,

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \hat{\mu})^2 \,, \tag{1}$$

where $\hat{\mu} = (y_1 + \cdots + y_n)/n$ is the sample mean. The estimator $\hat{\sigma}^2$ has a scaled χ^2 distribution with $n-1$ degrees of freedom:

$$(n-1)\,\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1} \,.$$

The χ^2_k distribution has the density function

$$f(x) = \frac{1}{\Gamma\left(\frac{k}{2}\right)} \left(\frac{1}{2}\right)^{\frac{k}{2}} x^{\frac{k}{2}-1} \exp\left(-\frac{x}{2}\right) .$$

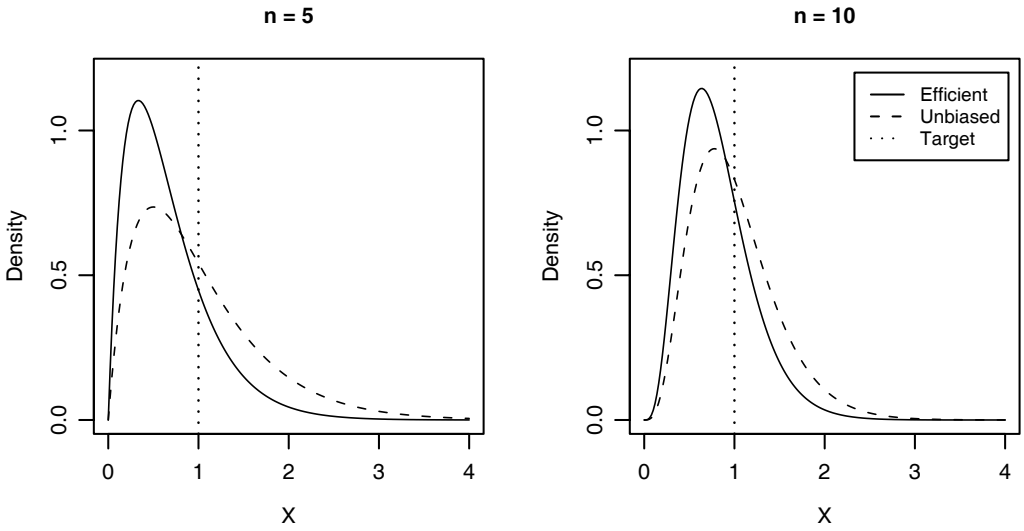

Figure 1: *The densities of the estimators $\hat{\sigma}^2$ and $\tilde{\sigma}^2$ for $n = 5$ and $n = 10$. Both panels are based on the setting $\sigma^2 = 1$.*

The correction for the degree of freedom lost, by using the divisor $n-1$ in (1), is generally regarded as important because the estimator $\hat{\sigma}^2$ is unbiased. However, if we do not insist on unbiasedness we obtain a more efficient estimator as $\tilde{\sigma}^2 = c^*\hat{\sigma}^2$, with $c^* = (n-1)/(n+1)$, that is, with divisor $n+1$ in (1); see Markowitz (1968) and Stuart (1969). The MSE reduction from $2\sigma^4/(n-1)$ to $2\sigma^4/(n+1)$, by $100(1-c^*)\% = 200/(n+1)\%$, converges to zero as $n \to +\infty$, but for small n it is far from trivial. The densities of $\hat{\sigma}^2$ and $\tilde{\sigma}^2$, based on samples of sizes $n = 5$ and 10, are drawn in Figure 1 for target $\sigma^2 = 1$. From the diagram it is difficult to judge that $\hat{\sigma}^2$ (dashes) is less efficient than $\tilde{\sigma}^2$ (solid line) because their densities are distinctly asymmetric and have different shapes. However, the distribution of $\hat{\sigma}^2$ has a thicker right-hand tail than $\tilde{\sigma}^2$, which corresponds to greater probability of large positive estimation errors $\hat{\sigma}^2 - \sigma^2$.

Estimates of variances are used in a variety of roles, and are often involved in nonlinear functions. For example, variance ratios v^2/σ^2 are estimated when comparing two variances using their (independent) estimators and the standardised value in meta-analysis (Sutton *et al.*, 2000; Longford, 2010) is defined as μ/σ, where μ is the (average) treatment effect and σ the standard deviation of the study-specific treatment effects. The efficiency of $\tilde{\sigma}^2$ is eroded by a nonlinear transformation, so $\tilde{\sigma}^2$ should not be substituted for σ^2 in a nonlinear expression, unless the sampling variation of $\tilde{\sigma}^2$ is very small. For example, neither $1/\hat{\sigma}^2$ nor $1/\tilde{\sigma}^2$ is an efficient or unbiased estimator of the precision $1/\sigma^2$. The respective expectation and variance of $1/\hat{\sigma}^2$ are $(n-1)/(n-3)/\sigma^2$ when $n > 3$ and $2(n-1)^2/\{(n-3)^2(n-5)\sigma^4\}$ when $n > 5$. These expressions are obtained by relating the relevant integrand to another χ^2 distribution or by differentiating the moment generating function; see Stuart (1969) and Stuart and Ord (1994, Chapter 16).

Substituting $\hat{\sigma}^2$ or $\tilde{\sigma}^2$ for σ^2 when it is (a factor) in a denominator is ill-advised when $n < 6$ because the resulting statistic has infinite variance.

We consider first estimators $c/\hat{\sigma}^2$ of $1/\sigma^2$. For $n > 5$, their MSEs are

$$\begin{aligned} & \frac{1}{\sigma^4}\left[\frac{2c^2(n-1)^2}{(n-3)^2(n-5)} + \left\{\frac{c(n-1)}{n-3} - 1\right\}^2\right] \\ = \; & \frac{1}{\sigma^4}\left\{c^2\frac{(n-1)^2}{(n-3)(n-5)} - 2c\frac{n-1}{n-3} + 1\right\}, \end{aligned}$$

so their minimum is attained for $c^* = (n-5)/(n-1)$. The minimum attained is $2/\{(n-3)\sigma^4\}$, smaller than the MSE of the naive estimator $1/\hat{\sigma}^2$, equal to $2(n+3)/\{(n-3)(n-5)\sigma^4\}$, or the MSE of the unbiased estimator $(n-3)/\{(n-1)\hat{\sigma}^2\}$, equal to $2/\{(n-5)\sigma^4\}$, so long as $n > 5$.

Although $c^*/\hat{\sigma}^2$ is much more efficient than $1/\hat{\sigma}^2$ for $n = 6, \ldots, 10$, it does not address the problem of infinite variance for $n \leq 5$. This problem is resolved by the estimator $1/(d+\hat{\sigma}^2)$ for a positive constant d, but the optimal value of d cannot be derived analytically. (A closed form expression for the MSE of this estimator involves incomplete gamma functions.) We explore this estimator by simulations in the next section.

The variance is often used in a linear function of σ or $1/\sigma$. Efficient estimators of these quantities in the respective classes of estimators $c\hat{\sigma}$ and $c/\hat{\sigma}$ are derived similarly to the efficient estimators of $\hat{\sigma}^2$ and $1/\hat{\sigma}^2$. Let

$$U_n = \frac{\sqrt{2}}{\sqrt{n-1}}\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}.$$

Then $\mathrm{E}(\hat{\sigma}) = \sigma U_n$ and $\mathrm{var}(\hat{\sigma}) = \sigma^2\left(1 - U_n^2\right)$, so the MSE of $c\hat{\sigma}$ is

$$\sigma^2\left\{(1-cU_n)^2 + c^2\left(1-U_n^2\right)\right\} = \sigma^2\left(1 - 2cU_n + c^2\right).$$

This function of c attains its minimum for $c^* = U_n$, and the minimum attained is $\sigma^2(1-U_n^2)$.

For estimating $1/\sigma$ we introduce the constants

$$V_n = \frac{\sqrt{n-1}}{\sqrt{2}}\frac{\Gamma\left(\frac{n-2}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} = \frac{\sqrt{n-1}}{\sqrt{n-2}}\frac{1}{U_{n-1}}.$$

Then

$$\begin{aligned} \mathrm{E}\left(\frac{1}{\hat{\sigma}}\right) &= \frac{V_n}{\sigma} \\ \mathrm{var}\left(\frac{1}{\hat{\sigma}}\right) &= \frac{1}{\sigma^2}\left(\frac{n-1}{n-3} - V_n^2\right). \end{aligned}$$

Hence the MSE of $c/\hat{\sigma}$ is

$$\frac{1}{\sigma^2}\left\{c^2\frac{n-1}{n-3}-2cV_n+1\right\},$$

and so the estimator of $1/\sigma$ efficient in the class of estimators $c/\hat{\sigma}$ is $\hat{\sigma}^{-1}V_n(n-3)/(n-1)$, so long as $n>3$. The corresponding MSE is $\{1-V_n^2(n-3)/(n-1)\}/\sigma^2$. Estimators of the form $c/(d+\hat{\sigma})$, with a positive offset d, may be more efficient; they have finite variances for any sample size n.

2.1. Estimating a reciprocal with an offset

We explore next estimating the reciprocal $1/\sigma$ by $1/(d+\hat{\sigma})$. We do this by simulations because we have no convenient expression for the moments of $1/(d+\hat{\sigma})$. Figure 2 displays the empirical biases and root-MSEs of the estimators $1/(d+\hat{\sigma})$ for $d\in(0,0.5)$ and $\sigma^2=0.1,0.25,0.5$ and 1.0, based on a sample of size $n=4$. The values of d for which the estimator is unbiased and for which it attains minimum MSE are marked by vertical ticks at the bottom of the respective panels.

Unbiasedness and minimum MSE are attained for different values of d. The minimum MSE is attained for $d^*=0.13,0.21,0.29$ and 0.41 when $\sigma^2=0.1,0.25,0.5$ and 1.0, respectively. Although d^* varies substantially with σ^2, the root-MSEs become more and more flat as σ^2 increases. Therefore, the choice of d is less critical for large σ^2, and

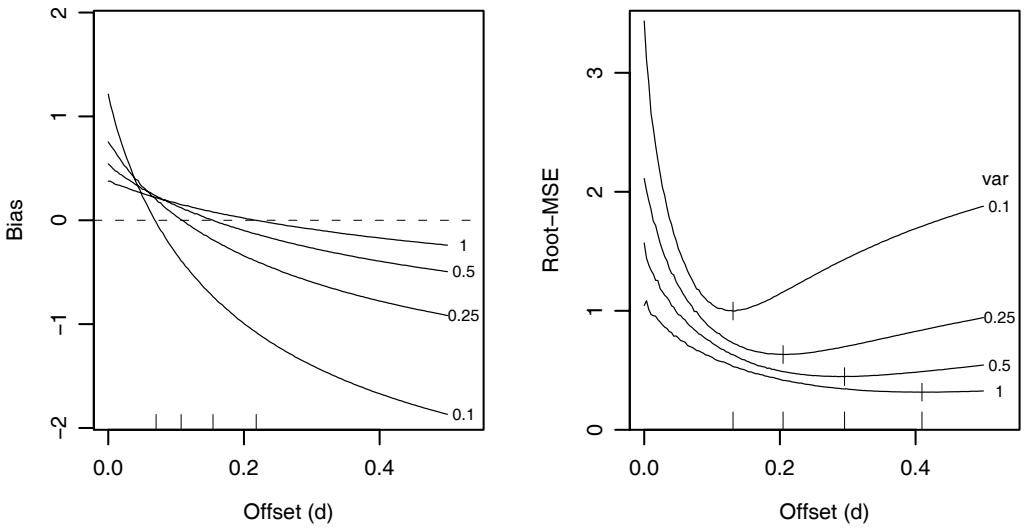

Figure 2: *The bias and root-MSE of the estimator $1/(d+\hat{\sigma})$ of $1/\sigma$ as functions of the offset d, with $n=4$ (3 degrees of freedom). The variances σ^2 are indicated at the right-hand margins. The ticks at the bottom of each panel indicate the value of d for which the estimator is unbiased (left-hand panel) and for which it attains minimum MSE (right-hand panel). Based on 100 000 replications.*

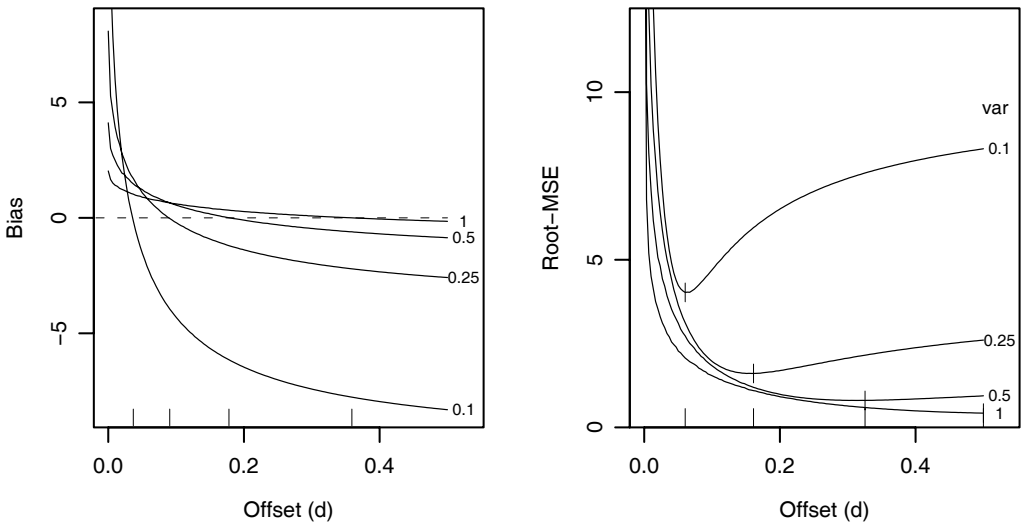

Figure 3: *The bias and root-MSE of the estimator* $1/(d+\hat{\sigma}^2)$ *of* $1/\sigma^2$ *as functions of the offset d, with* $n = 4$ *(3 degrees of freedom). Based on 100 000 replications. The same layout is used as in Figure 2.*

should be informed principally by the smallest plausible value of σ^2. This is a better strategy than using the value $\hat{d}^* = d^*(\hat{\sigma}^2)$ that would be optimal if our estimate were exact. If we can rule out small values of σ^2, a value $d > d^*$ is a safe choice because the root-MSE increases very slowly for $d > d^*$.

Figure 3 presents the biases and root-MSEs of the estimator $1/(d+\hat{\sigma}^2)$ of $1/\sigma^2$. It highlights how excessive bias and MSE are avoided by choosing a positive offset d. We arrive at the same general conclusion that if small values of σ^2 can be ruled out, then it is safe to choose a value d that is sufficiently large, because the root-MSEs are flat functions of d for d greater than the optimum offset.

Figures 4 and 5 display the biases and root-MSEs of the respective estimators $1/(d+\hat{\sigma})$ and $1/(d+\hat{\sigma}^2)$ of $1/\sigma$ and $1/\sigma^2$ for sample sizes $n = 6, 11$ and 21. They confirm that the root-MSE is a flat function of d for large σ^2. The precise choice of d is less important for greater variances σ^2, but the estimator $1/(d+\hat{\sigma})$ is very inefficient when too large a value of d is selected, especially when the variance σ^2 is small. Table 1 summarises the results for $\sigma^2 = 1$. The results for different values of σ^2 are obtained by replacing d with d/σ and applying the appropriate rescaling to the bias and MSE. The naive estimator of $1/\sigma$ is perceptibly inefficient even for $n = 21$, and the unbiased estimator is even more inefficient. The difference between the root MSEs of the two estimators that are efficient in the respective classes $c/\hat{\sigma}$ and $1/(d+\hat{\sigma})$ is about 8% for $n = 21$, and much more for smaller n.

One drawback of the estimator $1/(d^*+\hat{\sigma})$ is that the ideal offset d^* depends on σ^2. Therefore, the estimators with an offset can be compared more equitably with $c^*/\hat{\sigma}$ by finding the range of values d for which $1/(d+\hat{\sigma})$ is more efficient than $c^*/\hat{\sigma}$.

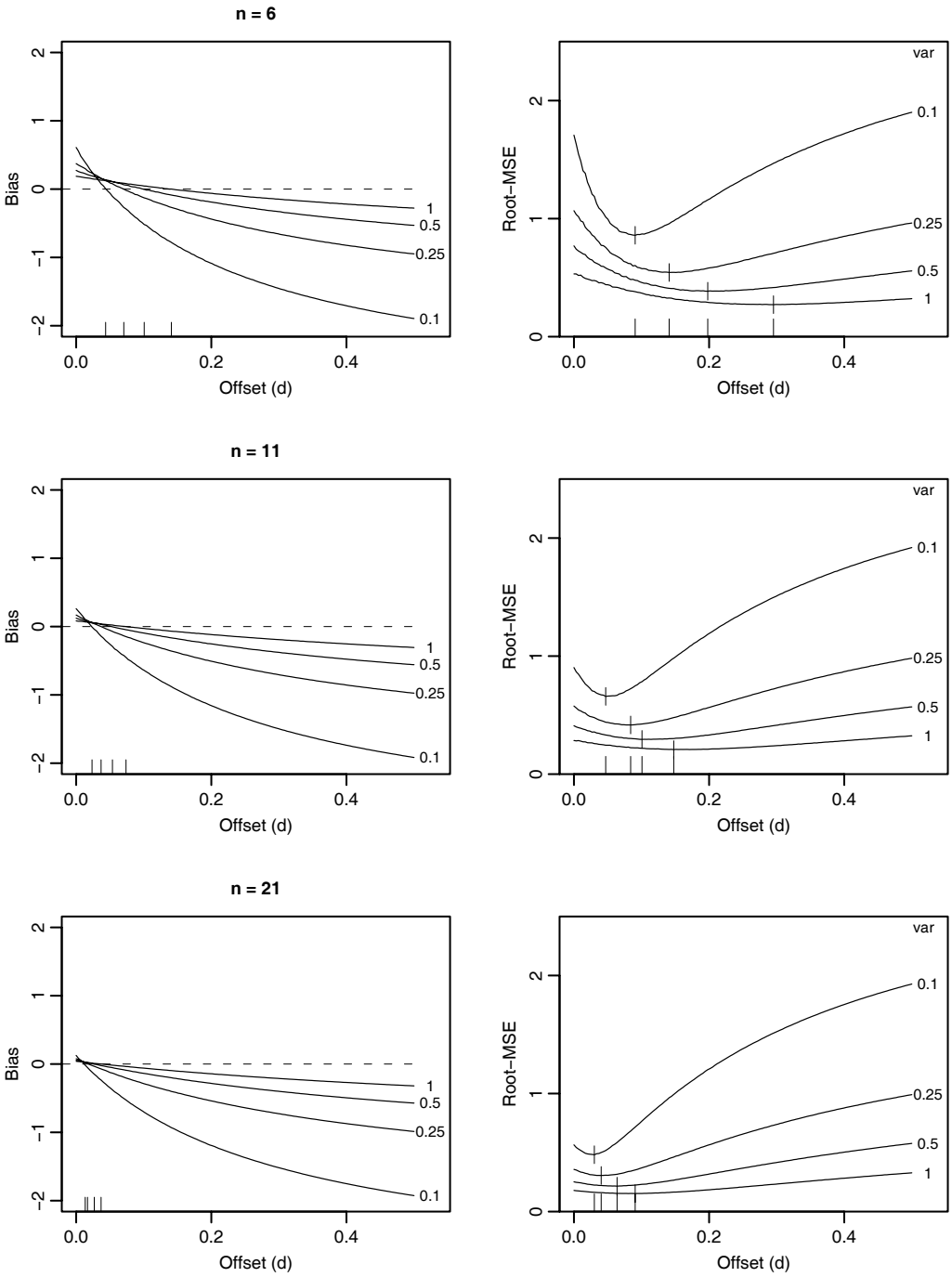

Figure 4: *The bias and root-MSE of the estimator* $1/(d+\hat{\sigma})$ *of* $1/\sigma$ *as functions of the offset* d*, with* $n = 6, 11$ *and 21 (*$n-1$ *degrees of freedom). Based on 50 000 replications.*

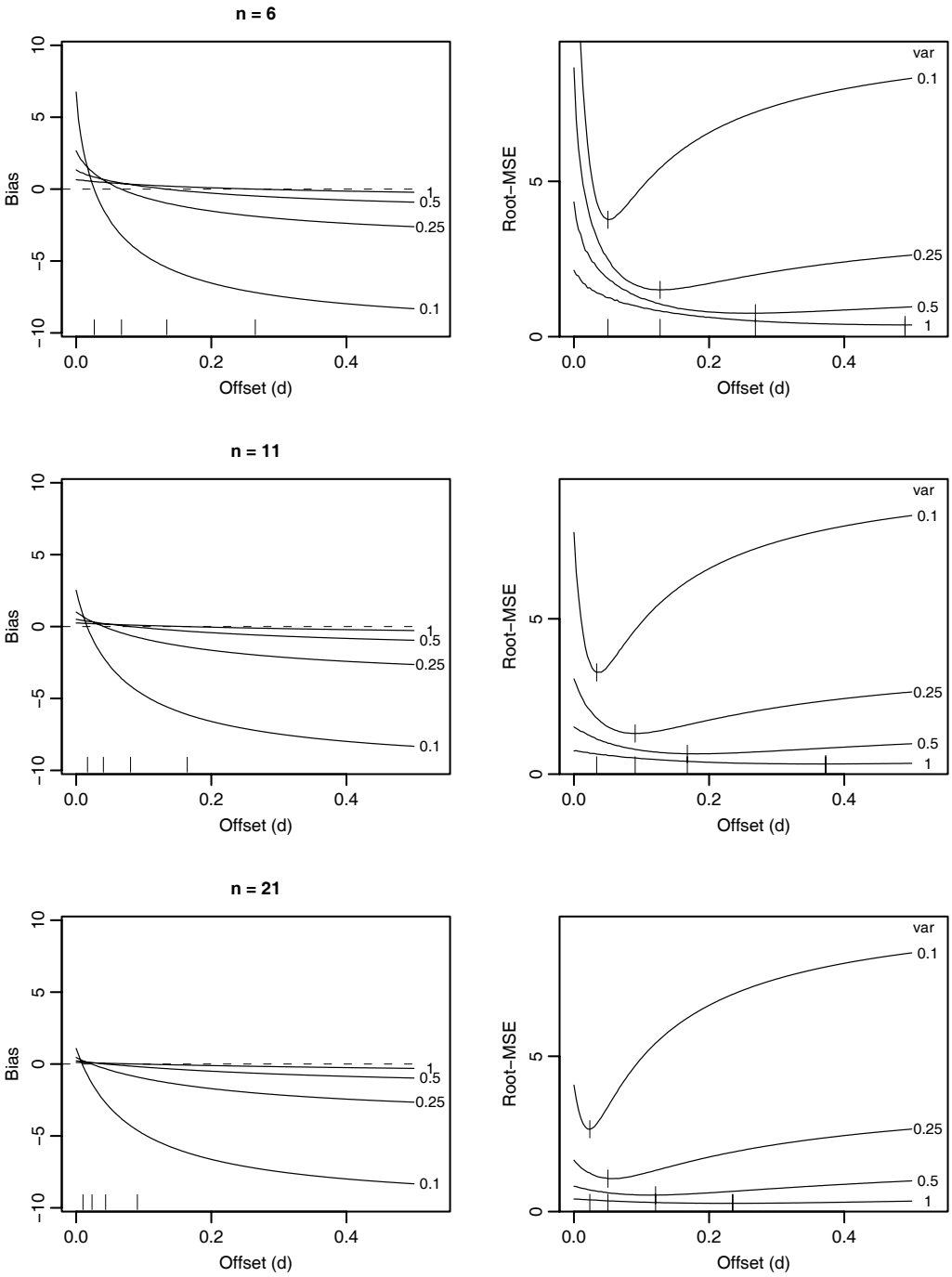

Figure 5: *The bias and root-MSE of the estimator* $1/(d+\hat{\sigma}^2)$ *of* $1/\sigma^2$ *as functions of the offset* d*, with* $n = 6, 11$ *and 21. Based on 50 000 replications.*

Table 1: *Properties of the alternative estimators of the reciprocal standard deviation* $1/\sigma$ *for sample sizes* $n = 6, 11$ *and 21;* $\sigma^2 = 1$.

Estimator		Bias			root-MSE		
		Sample size			Sample size		
		6	11	21	6	11	21
Naive	$\frac{1}{\hat{\sigma}}$	0.189	0.084	0.040	0.537	0.287	0.179
Unbiased	$\frac{1}{V_n\hat{\sigma}}$	0.000	0.000	0.000	0.633	0.392	0.262
Efficient in $\frac{c}{\hat{\sigma}}$	$\frac{V_n}{\hat{\sigma}}\frac{n-3}{n-1}$	−0.151	−0.060	−0.027	0.389	0.246	0.165
Efficient in $\frac{1}{d+\hat{\sigma}}$	$\frac{1}{d^*+\hat{\sigma}}$	−0.124	−0.072	−0.052	0.280	0.208	0.153

When $\sigma^2 = 1.0$, these ranges are 0.084 – 0.690, 0.045 – 0.312 and 0.025 – 0.151 for the respective sample sizes $n = 6, 11$ and 21. When $n = 21$, the optimal offset for $\sigma^2 = 0.1$ is $d^* = 0.030$. Therefore, when σ^2 is in fact equal to 1.0, but we base the value of d^* erroneously on $\sigma^2 = 0.1$, we still obtain an estimator that is more efficient than $c^*/\hat{\sigma}$.

Estimators of the precision $1/\sigma^2$ can be assessed similarly. The offset estimator $1/(d^* + \hat{\sigma}^2)$ is more efficient than $c^*/\hat{\sigma}^2$ even when $n = 21$ (root-MSEs 0.270 *versus* 0.333), but the largest error that we can afford in estimating or guessing the value of d^* is much smaller than for estimating $1/\sigma$. For example, the estimators $1/(d + \hat{\sigma}^2)$ are more efficient than $c^*/\hat{\sigma}^2$ for $0.064 < d < 0.490$. The ideal offset when $\sigma^2 = 0.25$ is $d^\dagger = 0.057$, outside this range, so $1/(d^\dagger + \hat{\sigma}^2)$ is less efficient than $c^*/\hat{\sigma}^2$. In contrast, for $\sigma^2 = 0.30$ we have $d^\dagger = 0.072$, so the offset estimator is more efficient than $c^*/\hat{\sigma}^2$. The gains in efficiency by using offset d in $1/(d + \hat{\sigma}^2)$ to estimate $1/\sigma^2$ are in general not as great as by using $1/(d + \hat{\sigma})$ for estimating $1/\sigma$.

We explored estimators $1/(d + \hat{\sigma})^2$, but found them uniformly less efficient than $1/(d + \hat{\sigma}^2)$. Estimators in the class $1/(d + c\hat{\sigma}^2)$ would be more efficient if the constants c and d were set optimally. However, having to set (or estimate) two constants is likely to be too difficult a task in most settings.

3. Decision about σ^2 with prior information

To estimate σ^2 better than by $\hat{\sigma}^2$, we draw on the prior information in a Bayesian approach. We want to cater for the setting in which no party could be called upon to declare a single prior distribution for σ^2. An expert (client) may not be available at all, the process of elicitation may reach an impasse, or the expert might feel uncomfortable with the declaration of any single prior because some similar priors might equally well

be declared, and yet they would lead to appreciably different posterior distributions. See Garthwaite, Kadane and O'Hagan (2005) for a review of methods of elicitation and related issues, such as the uncertainty about the prior. Given the strong impact of a prior on the posterior distribution, and the substantial uncertainty about the prior, drawing any conclusions from the details of any particular posterior distribution is poorly justified. We therefore focus on the tails of the posterior in the context of the problem with a discrete choice. Suppose we have two options, actions A and B; A is preferred when $\sigma^2 < \sigma^2_{\mathrm{R}}$ and B is preferred otherwise. The *reference* variance σ^2_{R} is given. If there is no obvious value of σ^2_{R}, the method described below can be applied to a small number of distinct values of σ^2_{R}.

We consider only the inverse gamma distributions as possible priors for σ^2; their densities are

$$f(s) = \frac{1}{\Gamma(r)}\theta^r s^{-r-1}\exp\left(-\frac{\theta}{s}\right), \tag{2}$$

where $\theta > 0$ and $r > 0$ are parameters, called the shape and inverse scale, respectively. We regard this class of distributions as sufficiently rich for representing the prior information. Convenience is an important factor in this choice; inverse gamma is the conjugate distribution for (scaled) χ^2, the distribution of the estimator $\hat{\sigma}^2$. The expectation of the inverse gamma is $\theta/(r-1)$, so long as $r > 1$, and its variance is $\theta^2/\{(r-1)^2(r-2)\}$, so long as $r > 2$.

We prefer the parametrisation in terms of the precision $\tau = 1/\sigma^2$, the double-shape $q = 2r$ and the scale $\lambda = 2\theta/q$, because it facilitates an easier interpretation and helps the client make the relevant choices regarding the prior distribution. The prior density for τ that corresponds to (2) is the gamma

$$f(\tau) = \frac{1}{\Gamma\left(\frac{q}{2}\right)}\left(\frac{q\lambda}{2}\right)^{\frac{1}{2}q}\tau^{\frac{1}{2}q-1}\exp\left(-\frac{q\lambda\tau}{2}\right).$$

The posterior density of τ is

$$f(t \mid \hat{\sigma}^2 = y) = C(k,q,\lambda)\,t^{\frac{1}{2}(k+q)-1}\exp\left\{-\frac{t}{2}(ky+q\lambda)\right\},$$

where C is the normalising constant. The corresponding distribution is scaled χ^2 with $k+q$ degrees of freedom and the scaling $ky+q\lambda$. In the standard Bayesian approach, all inferential statements about σ^2 are based on this distribution. For example, its expectation $(ky+q\lambda)/(k+q)$ may be quoted as an estimate, and its variance as a measure of uncertainty about σ^2, akin to the (frequentist) sampling variance.

In the frequentist perspective, the impact of the prior on the posterior is equivalent to adding q degrees of freedom (random draws from $\mathcal{N}(\mu,\sigma^2)$ or elementary observations) with a contribution of λ per degree of freedom to the corrected sum of squares,

increasing it from $k\hat{\sigma}^2$ to $k\hat{\sigma}^2 + q\lambda$. (Of course, we have to overlook that q may be fractional.) We can regard $\hat{\sigma}^2$ and λ as two independent (elementary) estimators of σ^2. Then the posterior expectation is a composite estimator of σ^2; it combines the two elementary estimators with weights proportional to the associated degrees of freedom.

3.1. Borderline priors

Without being able or willing to commit ourselves to a single prior when the prior would have a strong impact on the posterior distribution, it is not feasible or meaningful to study the entire posterior. Instead, we focus on the tails of the posterior, addressing the concern that the variance σ^2 may be greater (or smaller) than an *a priori* set reference value σ^2_{R}. We can motivate this by adopting the following decision rule. If σ^2_{R} lies in the $100\alpha\%$ right-hand tail of the posterior distribution for σ^2, that is,

$$\mathrm{P}\left(\sigma^2 > \sigma^2_{\mathrm{R}} \,|\, \hat{\sigma}^2\right) < \alpha\,,$$

for a given probability α, we take action A; otherwise we take action B. This is similar to a Bayesian version of hypothesis testing, although we treat the two actions symmetrically and consider both very small and very large values of α (e.g., 0.05 and 0.95).

We want to cater for settings in which the process of elicitation has not been concluded with a single prior (or has not taken place at all), but a set of plausible priors has been agreed (or was declared by the analyst). Such a set may be a rectangle given by the ranges $\lambda \in (\lambda_{\mathrm{L}}, \lambda_{\mathrm{H}})$ and $q \in (q_{\mathrm{L}}, q_{\mathrm{H}})$, or, more generally, a convex set in the parameter space for (λ, q). Since there is no single (prior) distribution that faithfully reflects the prior information, we invert the standard Bayesian solution and seek priors that would yield the so-called *borderline* posteriors. These are posteriors for which the $100(1-\alpha)$ percentile is equal to σ^2_{R}. The corresponding priors are also called borderline.

For a given value of the (prior) parameter q, the borderline value of λ, for which (q, λ) defines a borderline prior, is given by the equation

$$\sigma^2_{\mathrm{R}} \frac{(k+q)^2}{k\hat{\sigma}^2 + q\lambda} = F^{-1}_{k+q}(1-\alpha)\,,$$

in which F_h is the distribution function (and F^{-1}_h the quantile function) of the χ^2 distribution with h degrees of freedom. The solution,

$$\lambda_{\mathrm{B}}(q) = \frac{1}{q}\left\{\frac{(k+q)^2\sigma^2_{\mathrm{R}}}{F^{-1}_{k+q}(1-\alpha)} - k\hat{\sigma}^2\right\},$$

is unique, although λ_{B} may be negative for some values of q. For given α and k, $\lambda_{\mathrm{B}}(q)$ is positive for small $q > 0$ when $\hat{\sigma}^2 < k\sigma^2_{\mathrm{R}}/F^{-1}_k(1-\alpha)$.

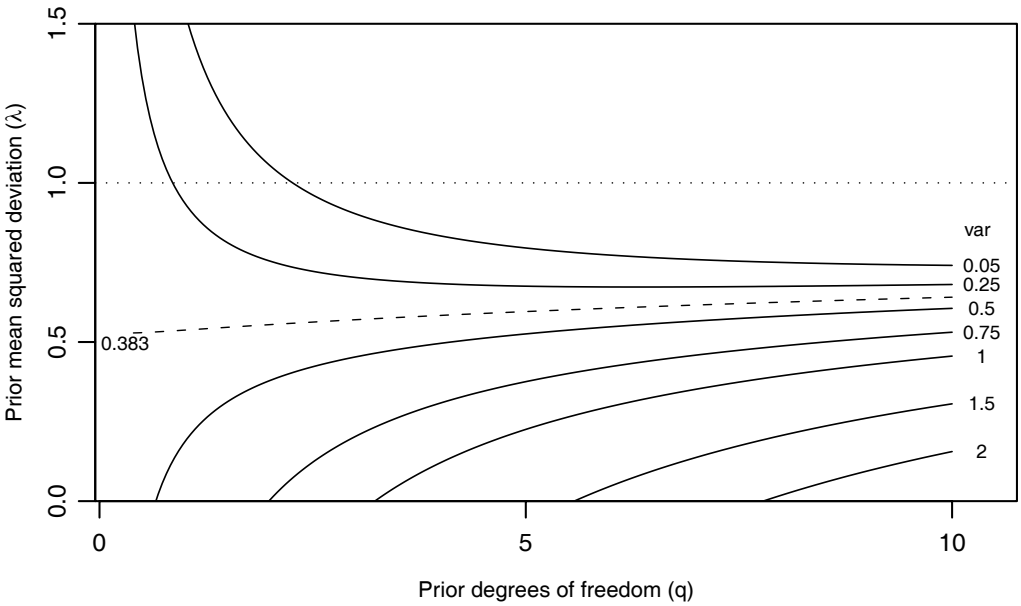

Figure 6: *Borderline priors for the setting with $k = 3$ degrees of freedom, the reference variance $\sigma^2_{\mathrm{R}} = 1$ and $\alpha = 0.05$. The threshold borderline function, for $\hat{\sigma}^2 = 0.383$ is drawn by dashes. The values of σ^2 are indicated at the right-hand margin.*

A set of borderline functions $\lambda_{\mathrm{B}}(q)$ is drawn in Figure 6 for $k = 3$, $\sigma^2_{\mathrm{R}} = 1$, $\alpha = 0.05$ and values of $\hat{\sigma}^2$ indicated at the right-hand margin. The functions are positive for all $q > 0$ when $\hat{\sigma}^2 \leq 0.383$; this threshold value of $\hat{\sigma}^2$ is found by a unidimensional search. All the functions converge to 1.0 as $q \to +\infty$, but the convergence, of the order $O(1/\sqrt{q})$, is very slow. When $\hat{\sigma}^2 < 0.383$, $\lambda_{\mathrm{B}}(q)$ attains very large values for small q, so that $q\lambda_{\mathrm{B}}(q)$ would make a nontrivial contribution to the posterior expectation $k\hat{\sigma}^2 + q\lambda_{\mathrm{B}}(q)$. When $\hat{\sigma}^2 < 0.383$, very small q is associated with large $\lambda_{\mathrm{B}}(q)$ because the prior contains very little information in relation to the data-based estimator $\hat{\sigma}^2$.

The borderline functions for the complemetary setting, with $\alpha = 0.95$, $k = 3$ and $\sigma^2_{\mathrm{R}} = 1$, are displayed in Figure 7. For $\hat{\sigma}^2 \in (7.27, 8.53)$, $\lambda_{\mathrm{B}}(0)$ is positive and yet $\lambda_{\mathrm{B}}(q) < 0$ for some positive values of q. For instance, when $\hat{\sigma}^2 = 8.0$, $\lambda_{\mathrm{B}}(q) < 0$ for $q \in (0.36, 5.60)$. By way of an example, suppose $\hat{\sigma}^2 = 6$ with $k = 3$, the prior parameter q is in the range $(3,5)$ and the prior value of λ does not exceed 0.8 (the shaded box in Figure 7). Then the entire set of plausible prior parameter vectors (q, λ) lies under the borderline function $\lambda_{\mathrm{B}}(q)$, and therefore action A is preferred for every plausible prior; we do not have to hone in on the prior.

If $\hat{\sigma}^2 = 8$, any prior with $q \in (3,5)$ is located above the borderline function, so action B is preferred. Note that it is not sufficient for both the prior λ and the estimate $\hat{\sigma}^2$ to be smaller than the reference σ^2_{R} to conclude with preference for small σ^2, because both sources of information are associated with a lot of uncertainty.

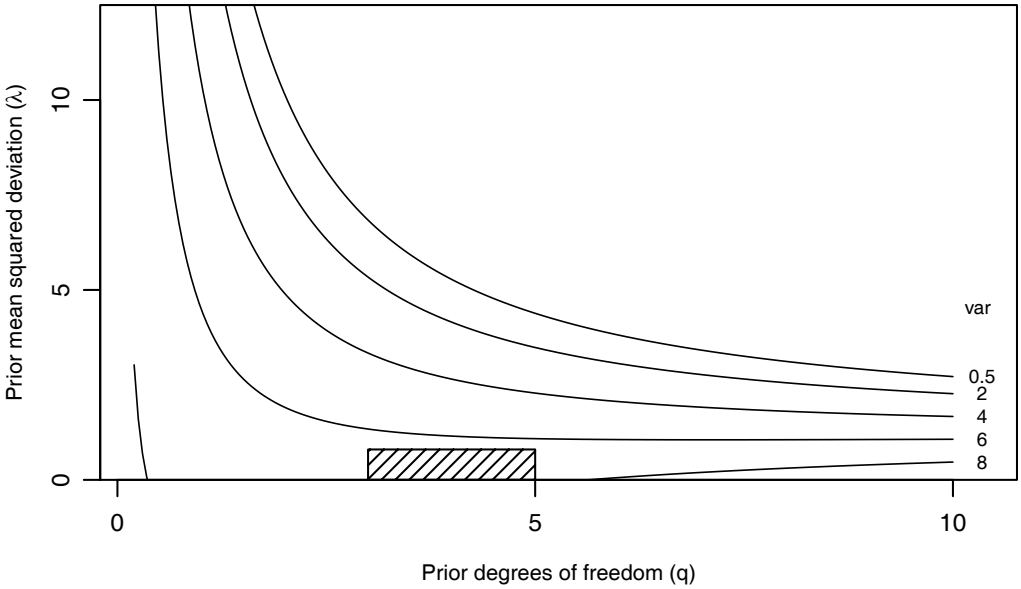

Figure 7: *Borderline priors for the setting with $k = 3$ degrees of freedom, the reference variance $\sigma^2_{\mathrm{R}} = 1$ and $\alpha = 0.95$. The shaded box represents a set of plausible priors.*

The borderline function divides the space of the prior parameters (q, λ) into the subsets that correspond to the priors which lead to the two decisions. A prior represented by a point under the curve corresponds to preference for values of σ^2 smaller than the reference, and (q, λ) above the curve to preference for values of σ^2 greater than σ^2_{R}. After being presented the borderline curve, an expert (client) has to decide whether any of the borderline priors are plausible. If none are, and the plausible prior parameter vectors (q, λ) are all above (or all below) the curve, we have an unequivocal decision. The advantage of this approach is that we do not have to force the elicitation process to yield a single prior. It suffices to specify a set of plausible priors. Such a set would be non-convex only in some esoteric settings, and it is hard to envisage even a setting in which it would not be a rectangle in (q, λ) or in a different parametrisation. If the borderline curve intersects this plausible set, we cannot choose between the two actions, because for some plausible priors action A, and for others action B, is preferred. There is, therefore, an incentive to reduce the set of plausible priors as much as possible, but not necessarily to a single point, as is required in the standard Bayesian setup.

A single prior has to originate from an expert. This is a serious stumbling block in any secondary analysis when the expert is not available for the necessary dialogue. Also, the expert may not be willing to commit him- or herself to a single prior. The analyst should proceed with the elicitation only as far as it is constructive. While the declaration of a single prior by an analyst on behalf of the client may be rather presumptious, the declaration of a plausible set of priors maintains the integrity of the analysis if this

set reflects the analyst's view of what a (real or hypothetical) client's prior may be. In essence, a solution is sought for every prior that the analyst believes the (absent) expert might choose. We do not want to integrate the posteriors over the plausible priors to obtain a single posterior distribution (Gelman *et al.* 2003), because that corresponds to using a (single) prior when some other priors are also plausible.

The dialogue with the expert is simplified by using a parametrisation for the priors that is easy to interpret. Thus, first we settle on the range of plausible prior degrees of freedom q (the strength or extent of prior information), and then on λ (the range of prior magnitudes of σ^2). This leads to a rectangle of plausible priors that may be reviewed further. The reference variance σ^2_{R} is set to reflect the client's priorities; when there is no clear candidate value, the problem may be solved for several references. The tail probabilities are usually set by convention, motivated by the practice of hypothesis testing.

3.2. Equilibrium priors

A drawback of the analysis with the borderline priors is that the consequences of the errors of the two kinds, choosing one action when the other would be appropriate, are ignored. To adapt the analysis, we have to specify the losses associated with such errors. Suppose the gain when we correctly conclude that $\sigma^2 > \sigma^2_{\mathrm{R}}$ (take the right action B) is greater than correctly concluding that $\sigma^2 < \sigma^2_{\mathrm{R}}$ by $|\,\sigma^2 - \sigma^2_{\mathrm{R}}\,|$, and the loss when we incorrectly conclude that $\sigma^2 > \sigma^2_{\mathrm{R}}$ (action B instead of A) is greater than incorrectly concluding that $\sigma^2 < \sigma^2_{\mathrm{R}}$ by $\rho\,|\,\sigma^2_{\mathrm{R}} - \sigma^2\,|$. The positive constant ρ is called the *penalty ratio*. Denote the posterior mean $\hat{\sigma}^2_{\mathrm{post}} = (k\hat{\sigma}^2 + q\lambda)/(k+q)$. The posterior density of σ^2 is $f_{k+q}\{(k+q)\,z/\hat{\sigma}^2_{\mathrm{post}}\}(k+q)/\hat{\sigma}^2_{\mathrm{post}}$, where f_h is the density of the χ^2 distribution with h degrees of freedom.

Our objective is to find the sign of the expected gain

$$
\begin{aligned}
&\int_0^{\sigma^2_{\mathrm{R}}} f_{k+q}\left\{\frac{(k+q)\,z}{\hat{\sigma}^2_{\mathrm{post}}}\right\}\frac{k+q}{\hat{\sigma}^2_{\mathrm{post}}}\left(\sigma^2_{\mathrm{R}} - z\right)\mathrm{d}z \\
&-\rho\int_{\sigma^2_{\mathrm{R}}}^{+\infty} f_{k+q}\left\{\frac{(k+q)\,z}{\hat{\sigma}^2_{\mathrm{post}}}\right\}\frac{k+q}{\hat{\sigma}^2_{\mathrm{post}}}\left(z - \sigma^2_{\mathrm{R}}\right)\mathrm{d}z \\
= \;& (1-\rho)\sigma^2_{\mathrm{R}}\,F_{k+q}\left\{\frac{(k+q)\,\sigma^2_{\mathrm{R}}}{\hat{\sigma}^2_{\mathrm{post}}}\right\} - (1-\rho)\hat{\sigma}^2_{\mathrm{post}}\,F_{k+q+1}\left\{\frac{(k+q)\,\sigma^2_{\mathrm{R}}}{\hat{\sigma}^2_{\mathrm{post}}}\right\} \\
&+\rho\left(\sigma^2_{\mathrm{R}} - \hat{\sigma}^2_{\mathrm{post}}\right), \qquad (3)
\end{aligned}
$$

derived using the identity $u f_h(u) = h f_{h+1}(u)$ for any positive h and u.

An expression similar to (3) can be derived for the loss functions that are piecewise linear in τ. That is, suppose the claim that $\sigma^2 > \sigma^2_{\mathrm{R}}$ $(= 1/\tau_{\mathrm{R}})$, when it is correct, is

associated with the gain $\tau_R - \tau$, and when it is incorrect, with the loss $\rho(\tau - \tau_R)$. Then the expected gain is

$$(\rho-1)\tau_R F_{k+q}\left(\frac{k+q}{\hat\sigma^2_{\text{post}}\tau_R}\right) - \frac{k+q}{k+q-1}\,\frac{\rho-1}{\hat\sigma^2_{\text{post}}}\,F_{k+q-1}\left(\frac{k+q}{\hat\sigma^2_{\text{post}}\tau_R}\right) + \rho\left(\tau - \frac{k+q}{k+q-1}\,\frac{1}{\hat\sigma^2_{\text{post}}}\right), \tag{4}$$

so long as $k+q>1$.

A prior or posterior is called *equilibrium* if the corresponding expected gain is equal to zero. In parallel with the borderline priors, we can represent the equilibrium priors as a function $\lambda^{(0)}(q)$, and discuss whether any of these priors are plausible. If all the plausible priors lie beneath this function, then action A, appropriate when $\sigma^2 < \sigma^2_R$, is associated with a positive expected gain; if all the plausible priors are above the function, then action B ($\sigma^2 > \sigma^2_R$) is associated with positive expected gain for every plausible prior.

For a given q we find the corresponding equilibrium value of $\lambda^{(0)}(q)$ by the Newton method. Since λ is involved in (3) and (4) only via $\hat\sigma^2_{\text{post}}$, we can find the 'equilibrium' value of $\hat\sigma^2_{\text{post}}$, denoted by $\hat\sigma^2_{\text{equi}}$, and then evaluate $\lambda^{(0)}(q)$ as $\{(k+q)\hat\sigma^2_{\text{equi}} - k\hat\sigma^2\}/q = \hat\sigma^2_{\text{equi}} + k(\hat\sigma^2_{\text{equi}} - \hat\sigma^2)/q$.

Figure 8 displays the equilibrium function $\lambda^{(0)}$ for the setting with $k=3$, $\hat\sigma^2 = 0.25$, $\rho = 20$ (solid line) and $\rho = 5$ (dashes), and the expected gain given by (3). The shaded

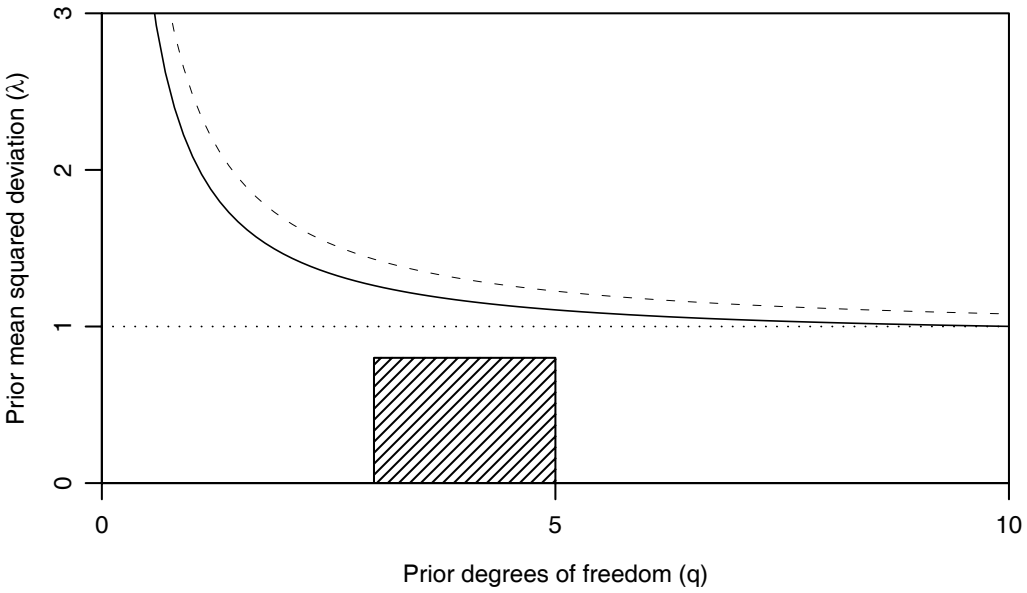

Figure 8: *Equilibrium priors for the setting with $k=3$ degrees of freedom, the reference variance $\sigma^2_R = 1$ and penalty ratios $\rho = 20$ (solid line) and $\rho = 5$ (dashes). The shaded box represents a set of plausible priors.*

box represents a set of plausible priors ($3 < q < 5$ and $0 < \lambda < 0.8$). Since it lies entirely beneath the equilibrium function, it yields an unequivocal conclusion, to take action A, because the expected gain is positive irrespective of which (plausible) prior is a faithful reflection of the prior information. The equilibrium functions in Figure 8 are decreasing in the range $q \in (0, 10)$, and they converge to the reference probability $\sigma_R^2 = 1$ as $q \to +\infty$. However, they are not monotone in $(0, +\infty)$; their values dip under $\sigma_R^2 = 1$. For example, with $\rho = 10$, $\lambda^{(0)}(q)$ attains its minimum of 0.95 at $q \doteq 40$, and with $\rho = 50$ it attains its minimum of 0.78 at $q \doteq 170$.

An aplication of borderline and equilibrium priors in a different small-sample setting is presented in Longford (2009).

4. Conclusion

We explored several alternatives to the established (unbiased) estimator $\hat{\sigma}^2$ of the variance σ^2 in the standard setting of a random sample of small size from $\mathcal{N}(\mu, \sigma^2)$. We demonstrated that estimators of the form $c\hat{\sigma}^2$, and their transformations $g(c\hat{\sigma}^2)$, are superior to $g(\hat{\sigma}^2)$ for functions g equal to the identity and square root, and their reciprocals. The optimal constants c^* are specific to the transformations, but do not depend on σ^2. For the reciprocals, an offset can be applied, as in $1/(d + \hat{\sigma}^2)$ for $d > 0$. The optimal value of d depends on σ^2, but a modicum of error in the value of σ^2 used for the offset $d = d(\sigma^2)$ is tolerated without a substantial loss of efficiency or loss of the superiority over the optimal estimator $c^*/\hat{\sigma}$ or $c^*/\hat{\sigma}^2$.

We introduced the (Bayes) borderline and equilibrium priors for the variance σ^2. Although they require additional specification, a reference variance (σ_R^2) and a tail probability (α) or a loss function (penalty ratio), they choose among the two actions optimally with respect to these specifications. Instead of the standard setting in which a single prior is required, it suffices to specify a (convex) set of (plausible) priors. The analysis avoids an impasse and can maintain its integrity when the process of elicitation fails to conclude with a single prior, or when it does not take place at all. However, specifying a smaller set of plausible priors is advantageous because it is less likely to straddle the borderline or equilibrium curve, when the solution (the decision) is not unequivocal. An outstanding challenge is to combine the advantages of the offset and prior information.

Although a suitable (near-optimal) offset d is found by simulations and the borderline or equilibrium curves are found by iterations, only a modest amount of computing is involved (a few minutes of CPU time for all the simulations). The software developed in `R` is available from the author on request.

Acknowledgements

Research for and preparation of this article were partly supported by the Spanish Ministry of Science and Technology through Grant SEJ2006–13537. Careful reading of the manuscript and helpful suggestions from Jordi Serra i Solanich and two anonymous referees are acknowledged.

References

Garthwaite, P. H., Kadane, J. B., and O'Hagan, A. (2005). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, 100, 680-700.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*, 2nd Ed. Chapman & Hall, CRC, New York.

Longford, N. T. (2009). Analysis of all-zero binomial outcomes with borderline and equilibrium priors. *Journal of Applied Statistics*, 36, 1259-1265.

Longford, N. T. (2010). Estimation of the effect size in meta-analysis with few studies. *Statistics in Medicine*, 29, 421-430.

Markowitz, E. (1968). Minimum mean-square-error of estimation of the standard deviation of the normal distribution. *The American Statistician*, 22, 26.

Stuart A. (1969). Reduced mean-square-error estimation of σ^p in normal samples. *The American Statistician*, 23, 27-28.

Stuart, A., and Ord, K. (1994). *Kendall's Advanced Theory of Statistics*, 6th Ed. Volume I. Distribution Theory. Edward Arnold, London.

Sutton, A. J., Jones, D. R., Abrams, K. R., Sheldon, T. A., and Song, F. (2000). *Methods for Meta-analysis in Medical Research*. Wiley, London, UK.