On Measuring Sensitivity to Parametric Model Misspecification

Author(s): Paul Gustafson

Source: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 2001, Vol. 63, No. 1 (2001), pp. 81-94

Published by: Wiley for the Royal Statistical Society

Stable URL: https://www.jstor.org/stable/2680635

J. R. Statist. Soc. B (2001)
63, *Part* 1, *pp.* 81–94

On measuring sensitivity to parametric model misspecification

Paul Gustafson

University of British Columbia, Vancouver, Canada

[Received September 1998. Final revision September 2000]

Summary. In settings where parametric inference is inconsistent under model misspecification, the discrepancy between correct and misspecified inferences is compared with the discrepancy between correct and misspecified models. To make the comparison tractable, large sample and small misspecification approximations are employed. The ratio of the approximate discrepancy between inferences to the approximate discrepancy between models is regarded as a relative measure of sensitivity to model misspecification. The maximum ratio over a family of correct distributions is determined as a measure of worst case sensitivity. As well, the distribution producing this maximum can be examined, to see how a particular combination of a parametric family and estimand is susceptible to model misspecifications.

Keywords: Bayes estimates; Large sample theory; Maximum likelihood estimates; Model misspecification

1. Introduction

Statisticians recognize that in reality no models are really correct. In many parametric situations, model misspecifications give rise to asymptotically inconsistent estimators of parameters. The goal of this work is to quantify how bad an estimator can be, relative to the degree of model misspecification. The rationale for such an investigation is that misleading inferences arise when a bad answer is produced from an only mildly misspecified model. We are less concerned about bad inferences arising from grossly misspecified models, as such models are likely to be screened out by diagnostic checks.

We quantify the sensitivity of a parametric family and estimand to model misspecifications by comparing the Kullback–Leibler (KL) divergence between correct and misspecified inferences with the KL divergence between correct and misspecified models. More specifically, the divergence between inferences can be based on either sampling distributions of maximum likelihood estimators or posterior distributions. For tractability local approximations to the KL divergence are employed, in a somewhat similar spirit to Cook (1986). The maximum ratio of approximate divergences over a class of correct models can be computed as a worst case measure of sensitivity to misspecifications. As well, the correct model achieving the maximum can be examined to gain insight into the kind of departures from the model which are damaging. This methodology is described fully in the next section. Then Section 3 details an application to some common parametric models for survival data. The paper concludes with a discussion in Section 4. Some technical comments are relegated to Appendix A and Appendix B.

Address for correspondence: Paul Gustafson, Department of Statistics, University of British Columbia, 333-6356 Agricultural Road, Vancouver, British Columbia, V6T 1Z2, Canada.
E-mail: gustaf@stat.ubc.ca

 1369–7412/01/63081

2. Quantifying sensitivity

2.1. General framework

Let $\mathcal{F} = \{f(x|\theta)\colon \theta \in \Theta\}$ be a parametric family of densities used to model independent and identically distributed random variables $X_1, \ldots, X_n$. To study situations where $\mathcal{F}$ is a misspecified model, we construct a second family of densities, $\mathcal{G} = \{g(x|\gamma)\colon \gamma \in \Gamma\}$, around a base-line density $f(x|\theta^*)$ belonging to $\mathcal{F}$. In particular, $\mathcal{G}$ is constructed so that $g(x|0) = f(x|\theta^*)$, and, as γ moves away from 0, $g(x|\gamma)$ moves away from $\mathcal{F}$ in some specified way. We consider what happens when data modelled with $\mathcal{F}$ are in fact generated from a member of $\mathcal{G}$.

To establish the notation, upper and lower case letters are used to distinguish distributions and densities respectively. Let $p = \dim(\theta)$ and $q = \dim(\gamma)$ be the dimensions of $\mathcal{F}$ and $\mathcal{G}$, with

$$s(\theta;\, x) = \partial[\log\{f(x|\theta)\}]/\partial\theta$$

and

$$t(\gamma;\, x) = \partial[\log\{g(x|\gamma)\}]/\partial\gamma$$

being the respective score vectors, and

$$I_F(\theta) = E_\theta\{s(\theta;\, X)^{\mathrm{T}}\, s(\theta;\, X)\}$$

and

$$I_G(\gamma) = E_\gamma\{t(\gamma;\, X)^{\mathrm{T}}\, t(\gamma;\, X)\}$$

being the corresponding Fisher information matrices. We require some regularity conditions for incorrect model asymptotics. To avoid distraction a discussion of regularity is relegated to Appendix A.

We focus on inference about a scalar estimand of interest, viewed as a functional T of the distribution generating the data. Common examples of T are a moment or quantile of the distribution. Thus the modeller uses a likelihood and possibly a prior on θ to yield *reported* inferences about $h_{\mathrm{R}}(\theta) = T\{F(\cdot|\theta)\}$, whereas in reality $h_{\mathrm{C}}(\gamma) = T\{G(\cdot|\gamma)\}$ is the true value of the estimand, i.e. *correct* inferences would arise if the modeller used a model other than $\mathcal{F}$ which contains the true distribution $G(\cdot|\gamma)$.

As a final preliminary, the KL divergence between densities is

$$\mathrm{KL}\{f_0 \parallel f_1\} = \int \log\left\{\frac{f_0(x)}{f_1(x)}\right\} \mathrm{d}F_0(x).$$

2.2. Inference and model discrepancy

We deal first with maximum likelihood estimation, when model $\mathcal{F}$ is used but in fact $G(\cdot|\gamma)$ is the distribution generating the data. For a sample of size n, the reported estimator is $h_{\mathrm{R}}(\hat{\theta}_n)$, where $\hat{\theta}_n$ maximizes the likelihood based on $f(\cdot|\theta)$. The sampling distribution of this estimator can be approximated by a normal distribution, denoted

$$\mathrm{INF}^{\mathrm{ML}}_{\mathrm{R},n} \equiv N\{\mu^{\mathrm{ML}}_{\mathrm{R}}(\gamma),\, n^{-1}\, \nu^{\mathrm{ML}}_{\mathrm{R}}(\gamma)\},$$

so that $n^{1/2}\{h_{\mathrm{R}}(\hat{\theta}_n) - \mu^{\mathrm{ML}}_{\mathrm{R}}(\gamma)\}$ converges in distribution to the $N\{0,\, \nu^{\mathrm{ML}}_{\mathrm{R}}(\gamma)\}$ distribution as n increases. It is well known (see Appendix A for references) that $\mu^{\mathrm{ML}}_{\mathrm{R}}(\gamma) = h_{\mathrm{R}}\{\tilde{\theta}(\gamma)\}$, where

$$\tilde{\theta}(\gamma) = \arg\min_{\theta}[\mathrm{KL}\{g(\cdot|\gamma) \parallel f(\cdot|\theta)\}], \tag{1}$$

i.e. estimates of θ converge to the value of θ at which the family $\mathcal{F}$ is closest to the true distribution in the KL sense. The form of $\nu_{\mathrm{R}}^{\mathrm{ML}}(\gamma)$ is also readily available (see Appendix A and the references therein), though it will not be required for our first-order investigation.

In a similar spirit, let

$$\mathrm{INF}_{\mathrm{C},n}^{\mathrm{ML}} \equiv N\{\mu_{\mathrm{C}}^{\mathrm{ML}}(\gamma),\, n^{-1}\nu_{\mathrm{C}}^{\mathrm{ML}}(\gamma)\}$$

denote a normal approximation to the sampling distribution of the maximum likelihood estimator if a correct model is used. We do not specify *which* correct model; we simply assume that some regular parametric family containing $g(\cdot|\gamma)$ is employed. The important point is that a correct specification implies consistency, so $\mu_{\mathrm{C}}^{\mathrm{ML}}(\gamma) = h_{\mathrm{C}}(\gamma)$. Because the specific correct model is not given, $\nu_{\mathrm{C}}^{\mathrm{ML}}(\gamma)$ cannot be determined. This will not, however, be an impediment to the following analysis.

We quantify the discrepancy between the correct and reported large sample inferences by the KL divergence between $\mathrm{INF}_{\mathrm{C},n}^{\mathrm{ML}}$ and $\mathrm{INF}_{\mathrm{R},n}^{\mathrm{ML}}$. The KL divergence between the two approximating normal distributions is readily calculated, giving

$$\mathrm{KL}\{\mathrm{INF}_{\mathrm{C},n}^{\mathrm{ML}} \parallel \mathrm{INF}_{\mathrm{R},n}^{\mathrm{ML}}\} = n\, a^{\mathrm{ML}}(\gamma) + \tilde{a}^{\mathrm{ML}}(\gamma) \tag{2}$$

where

$$a^{\mathrm{ML}}(\gamma) = \frac{\{\mu_{\mathrm{R}}^{\mathrm{ML}}(\gamma) - \mu_{\mathrm{C}}^{\mathrm{ML}}(\gamma)\}^2}{2\,\nu_{\mathrm{R}}^{\mathrm{ML}}(\gamma)} \tag{3}$$

and

$$\tilde{a}^{\mathrm{ML}}(\gamma) = \frac{1}{2}\left[\frac{\nu_{\mathrm{C}}^{\mathrm{ML}}(\gamma)}{\nu_{\mathrm{R}}^{\mathrm{ML}}(\gamma)} - \log\left\{\frac{\nu_{\mathrm{C}}^{\mathrm{ML}}(\gamma)}{\nu_{\mathrm{R}}^{\mathrm{ML}}(\gamma)}\right\} - 1\right].$$

Note in particular that the numerator of equation (3) is the squared asymptotic bias.

Now consider Bayesian analysis. Any reported inference would be based on the posterior distribution of $h_{\mathrm{R}}(\theta)$ induced by model $\mathcal{F}$. Let $\mu_{\mathrm{R},n}^{\mathrm{B}}(\gamma)$ be, say, the mode of this distribution for sample size n, and let $\nu_{\mathrm{R}}^{\mathrm{B}}(\gamma)$ be such that the posterior distribution of $n^{1/2}\{h_{\mathrm{R}}(\theta) - \mu_{\mathrm{R},n}^{\mathrm{B}}(\gamma)\}$ converges to the $N\{0,\, \nu_{\mathrm{R}}^{\mathrm{B}}(\gamma)\}$ distribution as n increases. Formally this can be viewed as convergence in distribution for almost every data sequence generated under $G(\cdot|\gamma)$. Thus the posterior distribution of the estimand $h_{\mathrm{R}}(\theta)$ can be approximated by

$$\mathrm{INF}_{\mathrm{R},n}^{\mathrm{B}} \equiv N\{\mu_{\mathrm{R},n}^{\mathrm{B}}(\gamma),\, n^{-1}\nu_{\mathrm{R}}^{\mathrm{B}}(\gamma)\}.$$

Similarly, let

$$\mathrm{INF}_{\mathrm{C},n}^{\mathrm{B}} \equiv N\{\mu_{\mathrm{C},n}^{\mathrm{B}}(\gamma),\, n^{-1}\nu_{\mathrm{C}}^{\mathrm{B}}(\gamma)\}$$

be an asymptotic approximation to the posterior distribution of the estimand under a correct model containing $g(\cdot|\gamma)$. Again it is not necessary to specify which correct model is used.

In keeping with the spirit of equation (2), the discrepancy between the correct and reported large sample inferences is quantified by the KL divergence between the correct and reported approximate posteriors, $\mathrm{INF}_{\mathrm{C},n}^{\mathrm{B}}$ and $\mathrm{INF}_{\mathrm{R},n}^{\mathrm{B}}$. In particular,

$$\mathrm{KL}\{\mathrm{INF}_{\mathrm{C},n}^{\mathrm{B}} \parallel \mathrm{INF}_{\mathrm{R},n}^{\mathrm{B}}\} = n\, a_n^{\mathrm{B}}(\gamma) + \tilde{a}^{\mathrm{B}}(\gamma) \tag{4}$$

where

$$a_n^{\mathrm{B}}(\gamma) = \frac{\{\mu_{\mathrm{R},n}^{\mathrm{B}}(\gamma) - \mu_{\mathrm{C},n}^{\mathrm{B}}(\gamma)\}^2}{2\,\nu_{\mathrm{R}}^{\mathrm{B}}(\gamma)} \tag{5}$$

and

$$\tilde{a}^{\mathrm{B}}(\gamma) = \frac{1}{2}\left[\frac{\nu_{\mathrm{C}}^{\mathrm{B}}(\gamma)}{\nu_{\mathrm{R}}^{\mathrm{B}}(\gamma)} - \log\left\{\frac{\nu_{\mathrm{C}}^{\mathrm{B}}(\gamma)}{\nu_{\mathrm{R}}^{\mathrm{B}}(\gamma)}\right\} - 1\right].$$

In the light of the asymptotics governing $\mathrm{INF}_{\mathrm{R}}$, it is natural to quantify the discrepancy between correct and assumed models by the minimum divergence between the true distribution and members of the assumed parametric family, i.e.

$$\begin{aligned}\mathrm{KL}\{\mathrm{MOD}_{\mathrm{C}} \parallel \mathrm{MOD}_{\mathrm{R}}\} &= b(\gamma) \\ &= \int \log\left[\frac{g(x|\gamma)}{f\{x|\tilde{\theta}(\gamma)\}}\right] \mathrm{d}G(x|\gamma).\end{aligned} \tag{6}$$

2.3. Relative sensitivity

The ratio of the inference discrepancy (2) or (4) to the model discrepancy (6) is viewed as an appropriate measure of sensitivity to model misspecification for the family $\mathcal{F}$ and the estimand T. A value of γ which makes this ratio large corresponds to a damaging true distribution in that relatively the answer is bad, and yet the misspecification is not easy to detect. We restrict our attention to the case of inconsistency under misspecification, i.e. $\mu_{\mathrm{R}}^{\mathrm{ML}}(\gamma) \neq \mu_{\mathrm{C}}^{\mathrm{ML}}(\gamma)$ when $\gamma \neq 0$. In such cases, to leading asymptotic order the ratio is given by

$$\lim_{n\to\infty}\left(n^{-1}\,\frac{\mathrm{KL}\{\mathrm{INF}_{\mathrm{C},n} \parallel \mathrm{INF}_{\mathrm{R},n}\}}{\mathrm{KL}\{\mathrm{MOD}_{\mathrm{C}} \parallel \mathrm{MOD}_{\mathrm{R}}\}}\right) = \frac{a(\gamma)}{b(\gamma)}, \tag{7}$$

where $a(\gamma) = a^{\mathrm{ML}}(\gamma)$ in the maximum likelihood case and $a(\gamma) = a^{\mathrm{B}}(\gamma) = \lim_{n\to\infty}\{a_n^{\mathrm{B}}(\gamma)\}$ in the Bayes case. Since $\mu_{\mathrm{R},n}^{\mathrm{B}} = \mu_{\mathrm{R}}^{\mathrm{ML}} + O(n^{-1/2})$ and $\mu_{\mathrm{C},n}^{\mathrm{B}} = \mu_{\mathrm{C}}^{\mathrm{ML}} + O(n^{-1/2})$, we have

$$a^{\mathrm{B}}(\gamma) = \frac{\{\mu_{\mathrm{R}}^{\mathrm{ML}}(\gamma) - \mu_{\mathrm{C}}^{\mathrm{ML}}(\gamma)\}^2}{2\,\nu_{\mathrm{R}}^{\mathrm{B}}(\gamma)}, \tag{8}$$

which differs from $a^{\mathrm{ML}}(\gamma)$ as given in equation (3) only in the denominator.

Next we apply small misspecification approximations, namely $a^{\mathrm{ML}}(\gamma) = \gamma^{\mathrm{T}} A^{\mathrm{ML}} \gamma + O(\gamma^3)$ or $a^{\mathrm{B}}(\gamma) = \gamma^{\mathrm{T}} A^{\mathrm{B}} \gamma + O(\gamma^3)$, and $b(\gamma) = \gamma^{\mathrm{T}} B \gamma + O(\gamma^3)$, where A^{ML}, A^{B} and B are second-derivative matrices of $a^{\mathrm{ML}}(\gamma)$, $a^{\mathrm{B}}(\gamma)$ and $b(\gamma)$, each evaluated at $\gamma = 0$. In fact, whereas $a^{\mathrm{ML}}(\gamma)$ and $a^{\mathrm{B}}(\gamma)$ are not in general equal except when $\gamma = 0$, it turns out that $A^{\mathrm{ML}} = A^{\mathrm{B}}$, so we can write A unambiguously.

Thus for large n and small γ our approximate sensitivity measure is the ratio

$$R(\gamma) = \frac{\gamma^{\mathrm{T}} A \gamma}{\gamma^{\mathrm{T}} B \gamma}.$$

Of particular interest is the worst case sensitivity, defined as

$$R_{\mathrm{MAX}} = \max_{\gamma \neq 0}\left(\frac{\gamma^{\mathrm{T}} A \gamma}{\gamma^{\mathrm{T}} B \gamma}\right). \tag{9}$$

It turns out that R_{MAX} is relatively easy to compute. Moreover, we can examine the γ which achieves the maximum, to indicate the sort of distribution for which inference based on $\mathcal{F}$ and T is likely to be unreliable.

Of course $R(\gamma)$ and R_{MAX} require the determination of A and B. Towards this end, note from equation (1) that $\tilde{\theta}(\gamma)$ is the value of θ solving

$$\int s_i(\theta;\, x)\, \mathrm{d}G(x|\gamma) = 0, \qquad i = 1, \ldots, p. \tag{10}$$

Implicit differentiation of equation (10) gives

$$E_{G(\cdot|\gamma)}[s'\{\tilde{\theta}(\gamma);\, X\}]\, \tilde{\theta}'(\gamma) + E_{G(\cdot|\gamma)}[s\{\tilde{\theta}(\gamma);\, X\}^{\mathrm{T}}\, t(\gamma;\, X)] = 0,$$

which, when evaluated at $\gamma = 0$, yields

$$\tilde{\theta}'(0) = I_F^{-1}(\theta^*)\, E_{\theta^*}\{s(\theta^*;\, X)^{\mathrm{T}}\, t(0;\, X)\}. \tag{11}$$

This expression is needed to compute both A and B.

In the case of A, differentiation of equation (3) or (8) leads to $A = aa^{\mathrm{T}}$, where

$$a = \frac{1}{\sigma}\{h'_{\mathrm{R}}(\theta^*)\, \tilde{\theta}'(0) - h'_{\mathrm{C}}(0)\}^{\mathrm{T}}, \tag{12}$$

with $\sigma^2 = \nu_{\mathrm{R}}^{\mathrm{ML}}(0) = \nu_{\mathrm{R}}^{\mathrm{B}}(0) = h'_{\mathrm{R}}(\theta^*)\, I_F^{-1}(\theta^*) h'_{\mathrm{R}}(\theta^*)^{\mathrm{T}}$ being the asymptotic variance in the absence of misspecification ($\gamma = 0$).

In the case of B, differentiating equation (6) gives

$$B = I_G(0) - \tilde{\theta}'(0)^{\mathrm{T}}\, I_F(\theta^*)\, \tilde{\theta}'(0). \tag{13}$$

Note that $\gamma^{\mathrm{T}}\, I_G(0)\gamma$ approximates $\mathrm{KL}\{g(\cdot|\gamma) \,\|\, f(\cdot|\theta^*)\}$, so the second term in equation (13) adjusts for the fact that $\tilde{\theta}(\gamma)$ rather than θ^* indexes the closest member of $\mathcal{F}$ to $g(\cdot|\gamma)$.

If A and B were both positive definite, then a standard linear algebra result would give R_{MAX} as the largest eigenvalue of $B^{-1}A$. This result has been used in different Bayesian sensitivity analyses by McCulloch (1989) and Clarke and Gustafson (1998). In the present context B is positive definite, but $A = aa^{\mathrm{T}}$ is non-negative definite and of rank 1. Since the only eigenvector of $B^{-1}aa^{\mathrm{T}}$ which corresponds to a non-zero eigenvalue is $B^{-1}a$, we have

$$R_{\mathrm{MAX}} = a^{\mathrm{T}} B^{-1} a,$$

with the maximum achieved by taking $\gamma = B^{-1}a$ in equation (9).

2.4. *Invariance properties*

It is instructive to verify that the sensitivity measure R_{MAX} behaves sensibly under transformations. As an expanded notation, let $R_{\mathrm{MAX}}(\theta^*;\, X,\, \theta,\, T)$ be the maximum sensitivity (9) based on the observation of X parameterized by θ with estimand $T(F_X)$, when the base-line parameter value is θ^*. Furthermore, write $g_{X,\theta^*}(x|\gamma)$ rather than $g(x|\gamma)$ to denote the dependence of $\mathcal{G}$ on the chosen scale of the data and the base-line parameter value. Let $Y = s_1(X)$, $\phi = s_2(\theta)$ and $U(F_Y) = s_3\{T(F_{s_1^{-1}(Y)})\}$, where $s_1(\cdot)$, $s_2(\cdot)$ and $s_3(\cdot)$ are smooth invertible transformations. Define $\phi^* = s_2(\theta^*)$. Provided that the construction of $\mathcal{G}$ transforms in the obvious way,

$$g_{X,\,\theta^*}(x|\gamma) = g_{Y,\phi^*}\{s_1(x)|\gamma\} \left| \frac{\partial}{\partial x} s_1(x) \right|,$$

then $R_{\mathrm{MAX}}(\theta^*;\ X,\ \theta,\ T) = R_{\mathrm{MAX}}(\phi^*;\ Y,\ \phi,\ U)$, i.e. the sensitivity measure is invariant under a transformation of the data, the parameter and the estimand. The proof of this result is straightforward and therefore omitted.

From the invariance property described above, some of the behaviour of R_{MAX} in the presence of a scale parameter or location–scale parameters can be deduced. This will be helpful in interpreting the examples of Section 3. First consider a scale structure, with

$$f(x|\theta_1,\ \theta_2) = \theta_2^{-1} f(x/\theta_2|\theta_1,\ 1).$$

If the estimand T satisfies $T(F_{cX}) = r_c\{T(F_X)\}$ for a function r_c depending on c, then it is straightforward to show that $R_{\mathrm{MAX}}(\theta^*;\ X,\ \theta,\ T)$ does not depend on θ_2^*. A similar result obtains in the location–scale case. If

$$f(x|\theta_1,\ \theta_2) = \theta_2^{-1} f\{(x - \theta_1)/\theta_2|0,\ 1\}$$

and $T(F_{a+bX}) = r_{a,b}\{T(F_X)\}$, then in fact $R_{\mathrm{MAX}}(\theta^*;\ X,\ \theta,\ T)$ does not depend on θ^* at all.

2.5. A construction for $\mathcal{G}$ in the continuous univariate case

In the case of continuous univariate X_i, one way to create a rich family of densities around the base-line density $f(x|\theta^*)$ is to create a family around the standard uniform distribution and then to apply the probability integral transform. Specifically, let $\tilde{g}(u|\gamma)$ be a density on (0, 1) given by

$$\tilde{g}(u|\gamma) = \exp\left\{\sum_{i=1}^{q} \gamma_i\, b_i(u) - c(\gamma)\right\}, \tag{14}$$

where

$$c(\gamma) = \log\left[\int_0^1 \exp\left\{\sum_{i=1}^{q} \gamma_i\, b_i(u)\right\} \mathrm{d}u\right]$$

is a normalizing constant and $b_1(\cdot), \ldots, b_q(\cdot)$ are basis functions from which the log-density is constructed. We define $g(x|\gamma)$ as the density of $X = F^{-1}(U|\theta^*)$, where $U \sim \tilde{g}(\cdot|\gamma)$. Consequently the members of $\mathcal{G}$ are given as

$$g(x|\gamma) = \exp\left[\sum_{i=1}^{q} \gamma_i\, b_i\{F(x|\theta^*)\} - c(\gamma)\right] f(x|\theta^*).$$

The examples of the next section are based on taking $\tilde{\mathcal{G}} = \{\tilde{g}(u|\gamma)\colon \gamma \in \Gamma\}$ to be all distributions on (0, 1) with log-densities that are natural cubic splines on equally spaced knots i/q, $i = 0, \ldots, q$. We use the B-spline basis to span this family, so $b_i(\cdot)$ is the cubic B-spline basis function centred at i/q. For identifiability b_0 is omitted from the basis, so $\Sigma_{i=1}^{q} \gamma_i\, b_i(0) = 0$ for any choice of γ. Further details on the use of this family $\mathcal{G}$ in the calculation of equations (11)–(13) are given in Appendix B.

3. Examples

We consider three parametric models for univariate survival data. Each family is parameterized by $\theta = (\alpha,\ \beta)$, where α is a shape parameter and β is a scale parameter, i.e.

$$f(x|\alpha,\ \beta) = \beta^{-1} f(x/\beta|\alpha,\ 1).$$

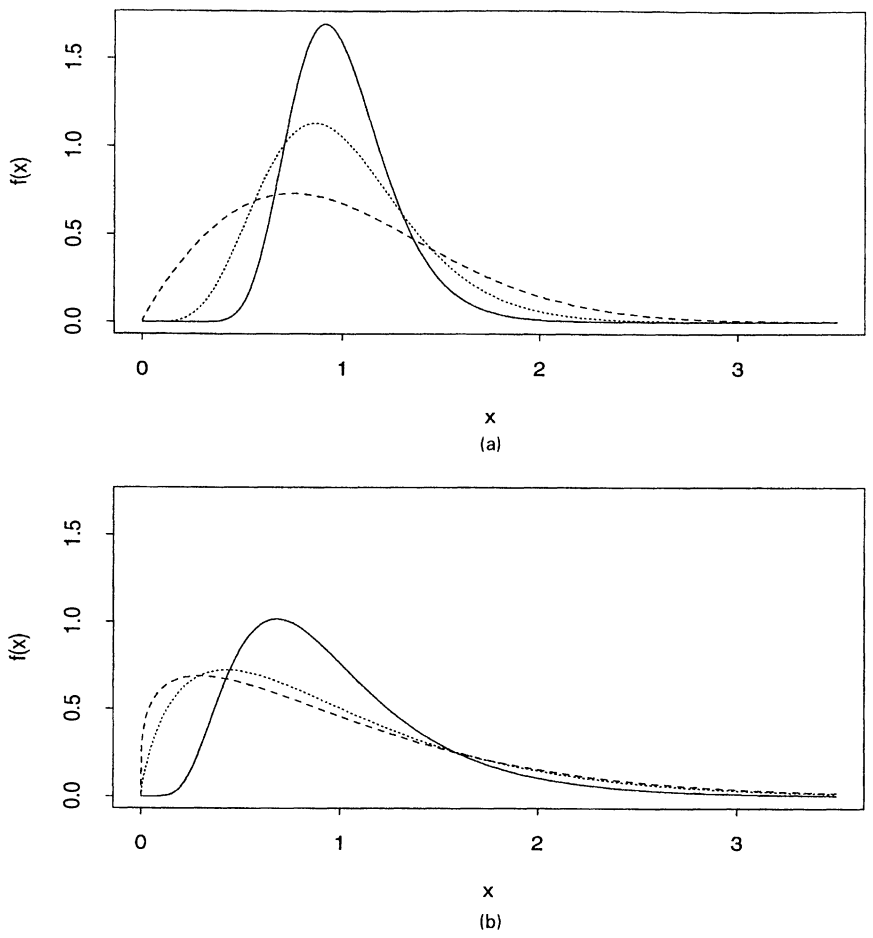

Fig. 1. Base-line densities for the three survival time families log-normal (———), gamma (·····) and Weibull (- - - -): in (a), the base-line parameters are chosen so that $\mathrm{sk}(\alpha) = 0.55$ and $E_\theta(X) = 1$ for each distribution; in (b), $\mathrm{sk}(\alpha) = 0.6$ and $E_\theta(X) = 1$ are used to determine the base-line parameters

The specific families are gamma, Weibull and log-normal, with respective densities

$$f_{\mathrm{GM}}(x|\alpha, 1) = \frac{1}{\Gamma(\alpha)} x^{\alpha-1} \exp(-x),$$

$$f_{\mathrm{WB}}(x|\alpha, 1) = \alpha x^{\alpha-1} \exp(-x^{\alpha}),$$

$$f_{\mathrm{LN}}(x|\alpha, 1) = \left(\frac{\alpha}{2\pi}\right)^{1/2} \frac{1}{x} \exp\left\{-\frac{\alpha}{2} \log(x)^2\right\},$$

for $x > 0$ and $\alpha > 0$. The shape parameters for the three families are not directly comparable, so we define $\mathrm{sk}(\alpha) = \Pr_\alpha\{X < E(X)\}$ as a measure of skewness that is applicable to all three families.

As a first scenario consider base-line parameter values chosen such that $\mathrm{sk}(\alpha) = 0.55$ and $E_\theta(X) = 1$ for all three parametric families. The resulting base-line densities appear in Fig. 1(a). For the three families and three different estimands we compute R_{MAX}, using the spline-based family $\mathcal{G}$ described in Section 2.5 with $q = 20$ knots. Numerical integration is

Table 1. R_{MAX} for three parametric families and three estimands†

Scenario	*Estimand*	*Results for the following parametric families:*		
		Log-normal	*Gamma*	*Weibull*
I	0.55 quantile	0.53	0.54	0.49
	Mean	0.00071	—	0.0019
	Variance	0.16	0.14	0.020
II	0.6 quantile	0.52	0.57	0.50
	Mean	0.0092	—	0.00095
	Variance	0.52	0.34	0.12

†In scenario I the base-line parameter values are those corresponding to Fig. 1(a), whereas those in scenario II correspond to Fig. 1(b).

used as necessary to compute equation (11) and subsequently equations (12) and (13). The estimands considered are the 0.55 quantile of X, the mean of X and the variance of X. The base-line value of the first two estimands is 1 for all three families. The R_{MAX}-values appear in the upper half of Table 1. The table entry for estimating a gamma mean is omitted, since it is easy to see from the score equations that inference is consistent in this case. For the log-normal and Weibull families, an estimation of the mean is much less susceptible to model misspecifications than estimations of the variance or quantiles are. Differences across families are slight for quantile estimation, but for variance estimation the Weibull model is much less susceptible to misspecification than the other two models are.

As a second scenario consider base-line parameter values chosen so that $\mathrm{sk}(\alpha) = 0.6$ and $E_\theta(X) = 1$ for all three parametric families. The corresponding base-line densities appear in Fig. 1(b). Now the 0.6 quantile is considered as an estimand, so again the base-line values of the first two estimands coincide under all three families. The R_{MAX}-values, given in the lower half of Table 1, are similar to those in the previous scenario, except that the estimation of the variance is now somewhat more susceptible to model misspecifications.

We also examine the eigenvector $\gamma_{MAX} = B^{-1}a$ that gives rise to each R_{MAX}-value. In the second scenario, the corresponding worst case densities on the uniform scale, $\tilde{g}(u|\gamma_{MAX})$ as in equation (14), appear in Fig. 2. For presentation, each eigenvector is standardized according to $\gamma^T B\gamma = 0.05^2$, to represent a fixed amount of divergence from the assumed model $\mathcal{F}$. There is still a sign indeterminacy, however, and so two densities appear in each case. For each estimand, the worst case densities on the uniform scale are quite similar for the three families. For estimating the 0.6 quantile, the worst case densities change rapidly near 0.6, which is not surprising. For the other estimands, however, the maximum sensitivity is attained with a very smooth redistribution of mass. This highlights a fundamental characteristic of our approach to sensitivity. In contrast with some classical robustness theories, worst case densities are not degenerate, or near degenerate. We attribute this to the relative nature of our sensitivity measure.

In the light of the invariance property for scale parameters discussed in Section 2.4, the R_{MAX}-values in Table 1 are invariant to changes in the base-line scale parameter β^*. To study the effect of changes in the base-line shape parameter α^*, R_{MAX} is computed for a variety of α^*-values, for each family and three estimands (the median, mean and variance). The results are plotted in Fig. 3, using $\mathrm{sk}(\alpha^*)$ as the abscissa for cross-family comparability. For estimation of the median we see that R_{MAX} depends on the base-line shape parameter for the gamma family, but not for the log-normal or Weibull families. This is predictable on noting that the transformation $Y = \log(X)$ yields a location–scale family for the last two families. With respect to Y we are dealing with the estimand $T(F) = \exp\{F^{-1}(\frac{1}{2})\}$ which satisfies

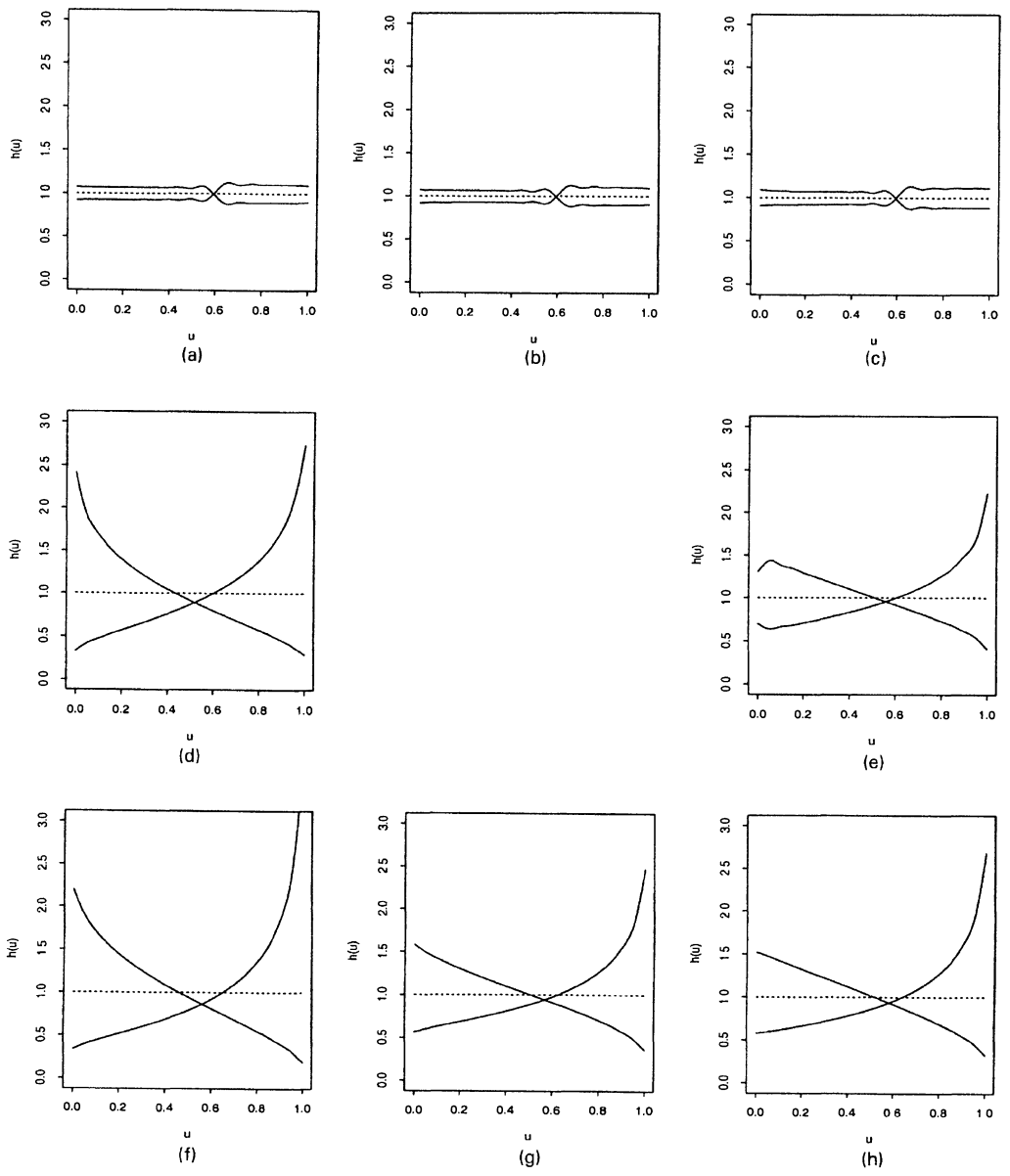

Fig. 2. Worst case distributions yielding the R_{MAX}-values in the lower half of Table 1 (densities are given on the (0, 1) scale; the eigenvectors are scaled via $\gamma^T B\gamma = 0.0025$, to represent a fixed amount of model discrepancy; two densities appear, corresponding to the two possible signs of the eigenvector; ·····, standard uniform density): (a) log-normal, 0.6 quantile; (b) gamma, 0.6 quantile; (c) Weibull, 0.6 quantile; (d) log-normal, mean; (e) Weibull, mean; (f) log-normal, variance; (g) gamma, variance; (h) Weibull, variance

$T(F_{a+bY}) = \exp\{a + b\, T(F_Y)\}$. Consequently the invariance property for location–scale families described in Section 2.4 obtains. Whereas the dependence of R_{MAX} on the base-line shape parameter is modest for estimation of the mean, the across-family differences in sensitivity for estimation of the variance become quite pronounced for more skewed base-line distributions.

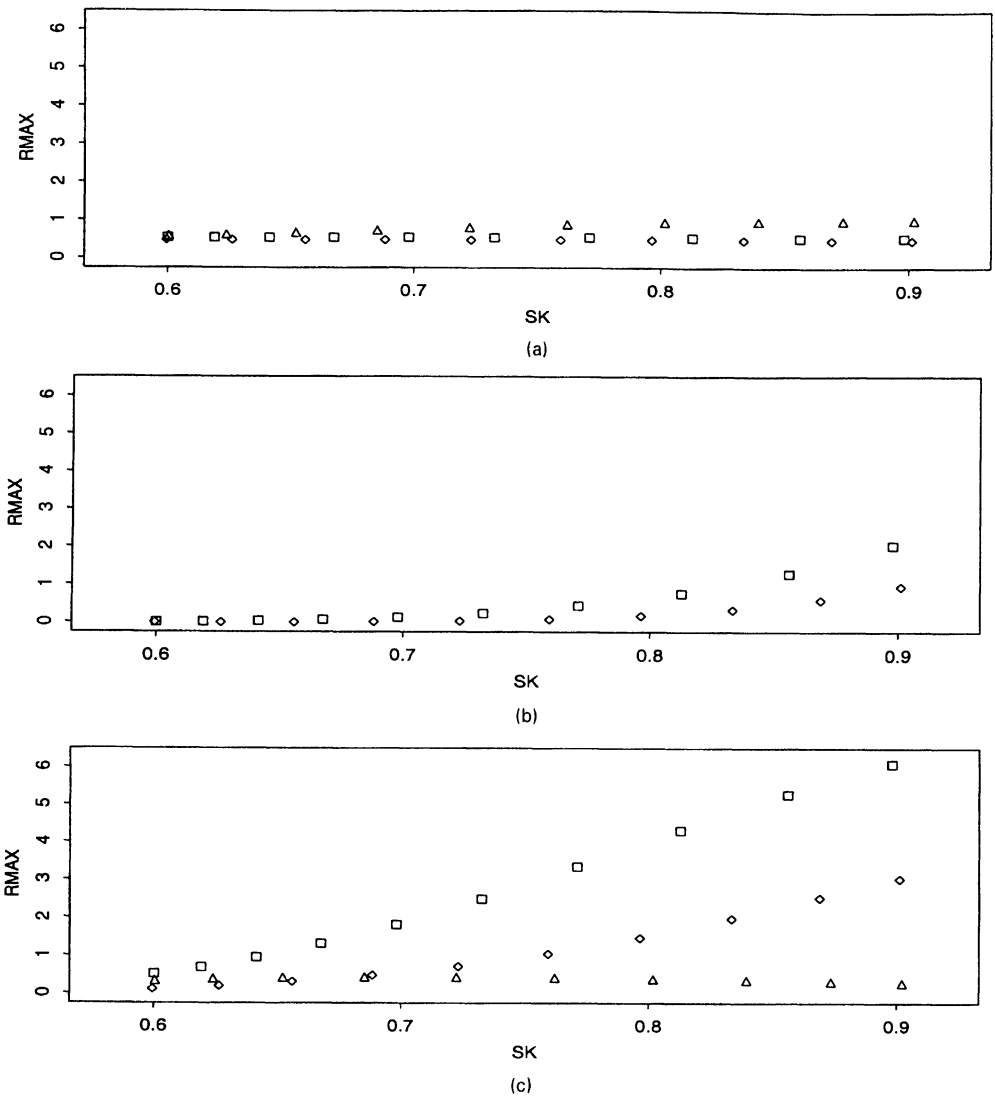

Fig. 3. R_{MAX} for the three estimands (a) median, (b) mean and (c) variance, as a function of the base-line shape parameter (for comparability across families, the abscissa is taken to be the skewness measure sk(α)): □, log-normal; △, gamma; ◇, Weibull

As a final investigation we compare R_{MAX} for different quantiles in Fig. 4. In the light of the aforementioned invariance considerations, R_{MAX} does not depend on the base-line shape parameter in the log-normal and Weibull families. For the gamma family, plots are given for the two base-line shape parameters utilized previously (sk(α) = 0.55 and sk(α) = 0.6). We find that R_{MAX} does not depend heavily on which quantile is being estimated, though it is slightly higher for extreme quantiles as we might expect.

The choice of $q = 20$ knots used throughout this section is somewhat arbitrary. To assess the dependence of R_{MAX} on the number of knots, the R_{MAX}-values in the lower half of Table 1 were

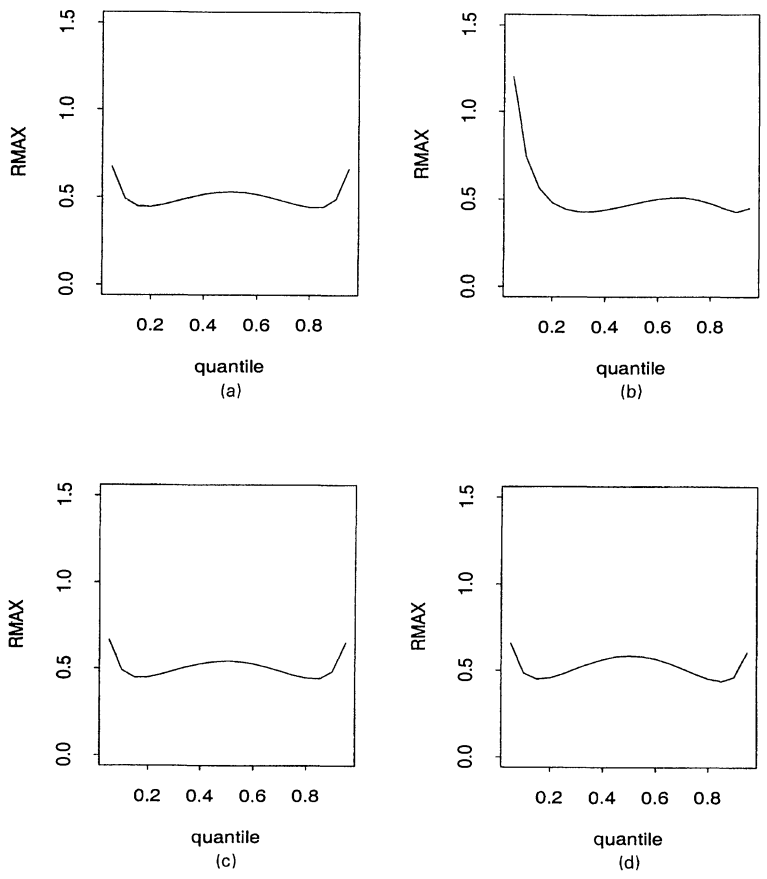

Fig. 4. R_{MAX} for estimating various quantiles: (a) log-normal; (b) Weibull; (c) gamma (sk(α) = 0.55); (d) gamma (sk(α) = 0.6) (for the log-normal and Weibull families, R_{MAX} does not depend on the base-line shape parameter)

recomputed under the alternative choices of $q = 10$ knots and $q = 30$ knots. The results appear in Table 2. We see in fact that R_{MAX} does not depend strongly on q. Analogous recomputations for the upper half of Table 1 (not shown) also support this finding. It is perhaps not surprising; by increasing the number of knots we allow both the inference discrepancy and the model discrepancy to increase, but these increases will offset each other to some extent in the ratio.

4. Discussion

As illustrated in the previous section, a computation of R_{MAX} permits comparisons of sensitivity to model misspecifications across models, across estimands and across base-line parameters within models. This may be of some practical assistance to a statistician facing the choice of a parametric family $\mathcal{F}$ for the analysis of a specific data set. If background knowledge and preliminary diagnostics do not clearly suggest one family over others, then families with small R_{MAX} for the estimand T of interest may be preferred. Of course this would be only one consideration among many that go into the choice of a model.

There are several limitations to our investigation of sensitivity to model misspecifications. One is that the choice of $\mathcal{G}$, the family of correct distributions, will be more problematic when

Table 2. R_{MAX}-values for alternative numbers of knots†

Estimand	*Results for the following parametric families:*		
	Log-normal	*Gamma*	*Weibull*
0.6 quantile	0.48, 0.53	0.52, 0.58	0.46, 0.52
Mean	0.0086, 0.0095	—	0.00091, 0.00096
Variance	0.48, 0.54	0.32, 0.34	0.11, 0.12

†The base-line parameter values are those corresponding to Fig. 1(b). In each cell the first entry is R_{MAX} based on $q = 10$ knots and the second entry is R_{MAX} based on $q = 30$ knots.

X_i is discrete or multivariate. Another is that $\mathcal{G}$ is necessarily parametric, in contrast with some investigations of data robustness which allow nonparametric perturbations of distributions. It is not clear how to arrive at a tractable sensitivity measure akin to R_{MAX} when $\mathcal{G}$ is nonparametric.

The scope of this investigation is also limited to estimands for which inference is inconsistent under model misspecification. There are many situations, such as the estimation of a gamma mean mentioned in Section 3, for which inference is consistent under misspecification. This scenario is more subtle in that $\tilde{a}(\gamma)$ in the inference discrepancy (2) or (4) is no longer a second-order term, since $a(\gamma) = 0$. As a result, the quadratic approximation to equation (2) or (4) now depends on $v_C^{ML}(\gamma)$ or $v_C^{B}(\gamma)$, the correct asymptotic variance. Consequently, constructing the family of correct parametric families, with one family for each value of γ, is now required. It is not clear how best to carry out this construction. One possibility is to elaborate on the probability integral transform approach. In particular, we can take the correct model under γ to be $\mathcal{F}_\gamma = \{f_\gamma(x|\theta)\colon \theta \in \Theta\}$, where $f_\gamma(x|\theta)$ is the density of $X = F^{-1}(U|\theta)$, with $U \sim \tilde{g}(u|\gamma)$. It is possible to carry through calculation of R_{MAX} in this context, though the calculations are involved, and it is not clear how much meaning to ascribe to this class of 'correct' parametric models.

Appendix A

The asymptotic behaviour of parametric estimators under model misspecifications has been well studied in the literature. Maximum likelihood estimators have been considered by Huber (1967), Akaike (1973), Kent (1982) and White (1982). White (1982) gave a set of seven assumptions which comprise standard regularity conditions for asymptotic normality under model misspecifications. The fundamental extra assumption that is needed to extend from the correct model case is that there be a unique minimizer in equation (1). The case of Bayesian estimators has been investigated by Berk (1966, 1970), Diaconis and Freedman (1986a, b) and Bunke and Milhaud (1998), though some of this work deals only with the convergence of the posterior to a point mass, and some of the asymptotic normality results are for the sampling distribution of a Bayes estimator, rather than for the posterior distribution itself. It seems straightforward to extend asymptotic normality results for the sampling distribution of say the maximum likelihood estimator to results for posterior quantities and distributions by using Laplace's method, though as far as this author is aware the Laplace method has only been applied in 'right model' situations (see, for instance, Tierney and Kadane (1986) and Tierney *et al.* (1989)).

Although it was not needed in our investigation, it is interesting to note that under misspecification the asymptotic variance for the sampling distribution of the maximum likelihood estimator of θ is given, for instance by White (1982), as

$$V^{ML}(\gamma) = E_\gamma[-s'\{\tilde{\theta}(\gamma);\, X\}]^{-1}\, E_\gamma[s\{\tilde{\theta}(\gamma);\, X\}^{T}\, s\{\tilde{\theta}(\gamma);\, X\}]\, E_\gamma[-s'\{\tilde{\theta}(\gamma);\, X\}]^{-1}. \tag{15}$$

In contrast, via Laplace's method we find that the asymptotic variance of the posterior distribution on θ is

$$V^B(\gamma) = E_\gamma[-s'\{\tilde{\theta}(\gamma);\, X\}]^{-1}, \tag{16}$$

where the expectations in both equation (15) and equation (16) are with respect to $G(\cdot|\gamma)$. Thus for the estimand of interest we have the asymptotic variances

$$\nu_R^{ML}(\gamma) = h_R'\{\tilde{\theta}(\gamma)\}^T\, V^{ML}(\gamma)\, h_R'\{\tilde{\theta}(\gamma)\}$$

and

$$\nu_R^{B}(\gamma) = h_R'\{\tilde{\theta}(\gamma)\}^T\, V^{B}(\gamma)\, h_R'\{\tilde{\theta}(\gamma)\}.$$

These are not equal in general, except when $\gamma = 0$ in which case equations (15) and (16) coincide under the right model scenario.

Much has been made of strange asymptotic behaviour in situations where the minimizer of equation (1) is not unique, especially in the Bayesian context (e.g. Diaconis and Freedman (1986b) and Bunke and Milhaud (1998)). Under weak regularity conditions our local analysis is not subject to non-uniqueness. In particular, let

$$w(\theta, \gamma) = \int \log\{f(x|\theta)\}\, dG(x|\gamma).$$

In addition to standard regularity for $\mathcal{F}$ and $\mathcal{G}$ separately, we assume the 'joint' condition that $w(\cdot)$ is twice continuously differentiable in a neighbourhood of $(\theta^*, 0)$, with derivatives that can be computed by differentiating under the integral sign. Then equation (10) corresponds to $\partial w(\theta, \gamma)/\partial\theta = 0$. But since

$$\frac{\partial^2}{\partial\theta^2} w(\theta, \gamma)\bigg|_{(\theta,\gamma)=(\theta^*,0)} = -I_F(\theta^*)$$

is negative definite the implicit function theorem implies that a unique differentiable solution of equation (10) exists in a neighbourhood of $\gamma = 0$.

Appendix B

Here we give some further details regarding the implementation of the family $\mathcal{G}$ proposed in Section 2.5. For $i = 2, \ldots, q-2$, the basis functions are given by $b_i(u) = b\{(u - i/q)/\Delta\}$ where

$$b(z) = \begin{cases} 0 & \text{if } z < -2, \\ (2+z)^3 & \text{if } z \in [-2, 1), \\ -3(1+z)^3 + 3(1+z)^2 + 3(1+z) + 1 & \text{if } z \in [-1, 0), \\ -3(1-z)^3 + 3(1-z)^2 + 3(1-z) + 1 & \text{if } z \in [0, 1), \\ (2-z)^3 & \text{if } z \in [1, 2), \\ 0 & \text{if } z \geqslant 2, \end{cases}$$

and $\Delta = 1/q$ is the interknot spacing on [0, 1]. Suitable modifications are made to define the basis functions $b_1(\cdot)$, $b_{q-1}(\cdot)$ and $b_q(\cdot)$ at the ends of the intervals (also, recall that $b_0(\cdot)$ has been omitted).

The quantities based on $\mathcal{G}$ that are needed to determine R_{MAX} are the score vector $t(\gamma;\, X)$ and the information matrix $I_G(\gamma)$, both evaluated at $\gamma = 0$. Straightforward calculation gives

$$t_i(0;\, x) = b_i\{F(x|\theta^*)\} - \int_0^1 b_i(u)\, du \tag{17}$$

and

$$[I_G(0)]_{ij} = \int_0^1 b_i(u)\, b_j(u)\, du - \int_0^1 b_i(u)\, du \int_0^1 b_j(u)\, du.$$

The integrands in both these expressions are piecewise polynomial and so can be calculated exactly in a straightforward but tedious manner. Armed with equation (17) and the score vector from model $\mathcal{F}$, the

expectations in equation (11) can be computed via numerical integration. This yields values for equations (12) and (13), which in turn permit the calculation of R_{MAX}.

References

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In *Proc. 2nd Int. Symp. Information Theory* (eds B. N. Petrov and F. Csàki), pp. 267–281. Budapest: Akadémiai Kiadó.

Berk, R. H. (1966) Limiting behaviour of posterior distributions when the model is incorrect. *Ann. Math. Statist.*, **37**, 51–58.

——— (1970) Consistency a posteriori. *Ann. Math. Statist.*, **41**, 894–906.

Bunke, O. and Milhaud, X. (1998) Asymptotic behaviour of Bayes estimates under possibly incorrect models. *Ann. Statist.*, **26**, 617–644.

Clarke, B. and Gustafson, P. (1998) On the overall sensitivity of the posterior distribution to its inputs. *J. Statist. Planng Inf.*, **71**, 137–150.

Cook, R. D. (1986) Assessment of local influence (with discussion). *J. R. Statist. Soc.* B, **48**, 133–169.

Diaconis, P. and Freedman, D. (1986a) On the consistency of Bayes estimates. *Ann. Statist.*, **14**, 1–26.

——— (1986b) On inconsistent Bayes estimates of location. *Ann. Statist.*, **14**, 68–87.

Huber, P. J. (1967) The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. 5th Berkeley Symp. Mathematical Statistics and Probability*. Berkeley: University of California Press.

Kent, J. T. (1982) Robust properties of likelihood ratio tests. *Biometrika*, **69**, 19–27.

McCulloch, R. E. (1989) Local model influence. *J. Am. Statist. Ass.*, **84**, 473–478.

Tierney, L. and Kadane, J. B. (1986) Accurate approximations for posterior moments and marginal densities. *J. Am. Statist. Ass.*, **81**, 82–86.

Tierney, L., Kass, R. E. and Kadane, J. B. (1989) Approximate marginal densities of nonlinear functions. *Biometrika*, **76**, 425–433.

White, H. A. (1982) Maximum likelihood estimation of misspecified models. *Econometrica*, **50**, 1–25.