*Biometrika* (1999), **86,** 1, *pp.* 153–167

*Printed in Great Britain*

# Multivariate Student-*t* regression models: Pitfalls and inference

BY CARMEN FERNANDEZ

*Department of Mathematics, University of Bristol, Bristol BS8 1TW, U.K.*
carmen.fernandez@bristol.ac.uk

AND MARK F. J. STEEL

*Department of Economics, University of Edinburgh, Edinburgh EH8 9JY, U.K.*
mark.steel@ed.ac.uk

## SUMMARY

We consider likelihood-based inference from multivariate regression models with independent Student-*t* errors with unknown degrees of freedom. Some pitfalls are revealed of both Bayesian and maximum likelihood methods. Under a commonly used non-informative prior, Bayesian inference is precluded for certain samples, even though a well-defined conditional distribution exists of the parameters given the observables. We also find that adding new observations can destroy the possibility of conducting posterior inference. Global maximisation of the likelihood function is a vacuous exercise since the latter is unbounded as we tend to the boundary of the parameter space. The unboundedness of the likelihood function also implies that the problems mentioned above for Bayesian analysis can even occur under proper priors. These pitfalls arise as a consequence of the fact that the recorded data have zero probability under the assumed sampling model. Therefore, a Bayesian analysis on the basis of set observations, which takes into account the precision with which the data were originally recorded, i.e. the 'rounding', is proposed and illustrated by several examples.

*Some key words*: Bayesian inference; Continuous distribution; Maximum likelihood; Missing data; Scale mixture of normals.

## 1. INTRODUCTION

The multivariate regression model with unknown scatter matrix is widely used in many fields of science. Applications to real data often indicate that the analytically convenient assumption of normality is not quite tenable, and thicker tails are called for in order to adequately capture the main features of the data. Thus, we consider regression error vectors that are distributed as scale mixtures of normals. We shall mainly emphasise the practically relevant case of independent sampling from a multivariate Student-*t* distribution with unknown degrees of freedom.

The Bayesian model will be completed with a commonly used improper prior on the regression coefficients and scatter matrix, and some proper priors on any remaining parameters, e.g. the degrees of freedom under Student-*t* sampling. Bayesian inference is based on the conditional distribution of the parameters given the observed sample, *y*, say. In practice, $y \equiv (y_1, \ldots, y_n)'$ consists of $n$ observations, each of which is a point in $\Re^p$, where $p \geqslant 1$ is the dimension of the sampling distribution. We shall therefore call such a $y$ a sample of 'point' observations. It is important to realise that any sample of point observations has zero probability of being generated by a continuous distribution, such as a scale mixture of normals as considered in this paper. A continuous distribution assumes that the data come to us as sets of positive Lebesgue measure, rather than as single points. In other words, it assumes that measurements cannot be taken with perfect precision. This seems quite appropriate for real-life situations, where one inevitably faces precision constraints in the measuring device, which means that a point observation is just a 'label' for an entire set. Usual statistical practice is, however, to ignore the rounding mechanism and conduct inference on the basis of the recorded point observations, which have zero probability under the assumed sampling model. This can lead to pitfalls, and the main object of this paper is to examine these issues in the context of a multivariate linear regression model with independent Student-$t$ errors with unknown degrees of freedom. Most of our discussion will be framed within the Bayesian paradigm, although § 5 also mentions some problems with the maximum likelihood approach.

Section 3 examines the usual posterior inference on the basis of a recorded sample of point observations. As examples of the pitfalls encountered, certain proper priors do not allow for posterior inference for some samples $y$, that is $p(y)$, the denominator in Bayes' formula, becomes infinite, or adding more observations can actually destroy the possibility of conducting posterior inference! From a measure theory point of view, these problems are explained by the fact that a conditional distribution is only defined up to a set of measure zero on the conditioning variable, which implies that Bayesian inference could be precluded under certain samples of point observations. Note that this does not per se rule out Bayesian inference with a continuous sampling model on the basis of point observations, but the latter can certainly not be guaranteed in general. An obvious sufficient condition for posterior inference, i.e. for obtaining a finite value of $p(y)$, in such situations is the combination of a bounded likelihood with a proper prior. However, many interesting statistical models, such as mixtures of normals, typically lead to unbounded likelihoods, whilst improper priors often arise as convenient devices to represent prior ignorance. This takes us outside the sufficient condition stated above, and the feasibility of conducting posterior inference becomes a case-specific problem which depends on the behaviour of both the likelihood and the prior over regions of the parameter space with unbounded likelihood or infinite prior mass. Often, the behaviour of the likelihood function will become worse if we record the data more crudely, i.e. with coarser rounding, as this will typically increase the degree of fit that the parameters can provide to the data, thus leading to higher likelihood values over certain regions in the parameter space. It should, however, be kept in mind that, since the source of the problems described here lies in the use of observations that have zero probability under the sampling model, they can arise under any dataset consisting of point observations.

Section 4 presents a solution through the use of 'set' observations, which have positive probability under the sampling model. These sets will be constructed as neighbourhoods of the recorded point observations, on the basis of the precision with which the sample was generated. This takes us back to a fully coherent Bayesian framework, where posterior inference is always feasible under a proper prior, and where the arrival of new set observations can never destroy the possibility of conducting inference. A Gibbs sampling scheme, see e.g. Gelfand & Smith (1990) and Casella & George (1992), is seen to be a convenient way to implement this solution in practice. Some examples are presented: a univariate regression model for the well-known stackloss data (Brownlee, 1965, p. 454), and a bivariate location-scale model for the Iris setosa data of Fisher (1936). The analysis through set observations is naturally extended to the case where some components of the multivariate response are not observed. We illustrate this with the artificial data of Murray (1977), augmented with some extreme values in Liu & Rubin (1995).

Finally, in § 5 the Student-$t$ likelihood function for point observations is analysed in some detail: it is found that the likelihood is unbounded as we tend to the boundary of the parameter space in a certain direction. This casts some doubt on the meaning and validity of a maximum likelihood analysis of this model, as performed e.g. in Lange, Little & Taylor (1989), Lange & Sinsheimer (1993) and Liu & Rubin (1994, 1995). This behaviour is illustrated through the stackloss data example, and it also explains the source of the problems encountered by Lange et al. (1989) and Lange & Sinsheimer (1993) when applying the EM algorithm for joint estimation of regression coefficients, scale and degrees of freedom to the radioimmunoassay dataset of Tiede & Pagano (1979).

Sketched proofs are grouped in the Appendix. With some abuse of notation, we do not explicitly distinguish between random variables and their realisations, and $p(.)$, a density function with respect to Lebesgue measure, or $P(.)$, a measure, can correspond to either a probability measure or a general $\sigma$-finite measure.

## 2. The model

Observations for the $p$-variate response variable $y_i$ are assumed to be generated through the linear regression model

$$y_i = \beta' x_i + \varepsilon_i \quad (i = 1, \ldots, n), \tag{2·1}$$

where $\beta$ is a $k \times p$ matrix of regression coefficients, $x_i$ is a $k$-dimensional vector of explanatory variables and the entire design matrix, $X = (x_1, \ldots, x_n)'$, is taken to be of full column rank $k$, written as $r(X) = k$. The error vectors $\varepsilon_i$ are independent and identically distributed as $p$-variate scale mixtures of normals with mean zero and positive definite symmetric covariance matrix $\Omega$. The mixing variables, denoted by $\lambda_i$ $(i = 1, \ldots, n)$, follow a probability distribution $P_{\lambda_i|\nu}$ on $\Re_+$, which can depend on a parameter $\nu \in \mathcal{N}$, possibly of infinite dimension. Thus, we have $n$ independent replications from the sampling density

$$p(y_i|\beta, \Omega, \nu) = \int_0^\infty \frac{\lambda_i^{p/2}}{(2\pi)^{p/2}|\Omega|^{1/2}} \exp\left\{-\frac{\lambda_i}{2}(y_i - \beta' x_i)'\Omega^{-1}(y_i - \beta' x_i)\right\} dP_{\lambda_i|\nu}. \tag{2·2}$$

By changing $P_{\lambda_i|\nu}$ we cover the class of $p$-variate scale mixtures of normals. The latter is a subset of the elliptical class, with ellipsoids in $\Re^p$ as isodensity sets, while allowing for a wide variety of tail behaviour. Leading examples are finite mixtures of normals, corresponding to a discrete distribution on $\lambda_i$, and the multivariate Student-$t$ distribution with $\nu > 0$ degrees of freedom, where $P_{\lambda_i|\nu}$ is a Ga$(\nu/2, \nu/2)$ distribution with unitary mean. Most of the subsequent discussion will focus on the practically relevant case of Student-$t$ sampling.

Special cases of the model in (2·1) are the multivariate location-scale model, where $k = 1$ and $x_i = 1$, and the univariate regression model for $p = 1$.

The Bayesian model needs to be completed with a prior distribution for $(\beta, \Omega, \nu)$. In most of the paper, we assume a product structure between the three parameters with

$$p(\beta, \Omega) \propto |\Omega|^{-(p+1)/2}, \tag{2·3}$$

$$P_{\nu} \text{ any probability measure on } \mathcal{N}. \tag{2·4}$$

The prior in (2·3) is the 'usual' default prior in the absence of compelling prior information on $(\beta, \Omega)$. Under Student-$t$ sampling with fixed $\nu$, it corresponds to Jeffreys' prior under 'independence', i.e. the product of the Jeffreys' priors for $\beta$, treating $\Omega$ as fixed, and $\Omega$, treating $\beta$ as fixed. It is thus invariant under separate reparameterisations of $\beta$ and of $\Omega$. Although our main Bayesian results will refer to this prior, other choices of priors and their implication for posterior inference will be briefly discussed at the end of § 3.

## 3. BAYESIAN INFERENCE USING POINT OBSERVATIONS

We now consider the feasibility of a Bayesian analysis of the model in (2·2)–(2·4) on the basis of the recorded point observations, as is the usual practice. Since the prior in (2·3) is improper, we first verify the existence of the conditional distribution of the parameters given the observables.

THEOREM 1. *Consider n independent replications from* (2·2) *with any mixing distribution* $P_{\lambda_i|\nu}$ *and the prior in* (2·3)–(2·4) *with any proper* $P_\nu$. *Then the conditional distribution of* $(\beta, \Omega, \nu)$ *given* $y \equiv (y_1, \ldots, y_n)'$ *exists if and only if* $n \geqslant k + p$.

Somewhat surprisingly, neither the mixing distribution nor the prior on $\nu$ affects the existence of the conditional distribution of the parameters given the observables, i.e. the posterior distribution. Thus, whenever $n \geqslant k + p$, the fact that the prior is improper is of no consequence for the existence of this conditional distribution, which puts us on equal footing with the case of proper priors. However, as explained in the Introduction, a conditional distribution is only defined up to a set of measure zero in the conditioning variables. In other words, Theorem 1 assures us that $p(y) < \infty$ except possibly on a set of Lebesgue measure zero in $\Re^{n \times p}$. Since any sample of point observations has Lebesgue measure zero, it follows that there is no guarantee that $p(y) < \infty$ for our particular observed sample, and the latter has to be verified explicitly.

We therefore complement Theorem 1 by considering any possible point $y \in \Re^{n \times p}$. Lemma 1 in the Appendix shows that both $P_{\lambda_i|\nu}$ and $P_\nu$ can now intervene. It is immediate from Lemma 1 that, for finite mixtures of normals, $p(y) < \infty$ if and only if $r(X:y) = k + p$, and that the latter is a requirement for any scale mixture of normals. Let us now analyse the more challenging case of Student-$t$ sampling.

DEFINITION 1. *For a design matrix X and a sample* $y \in \Re^{n \times p}$ *such that* $r(X:y) = k + p$, $s_j$ $(j = 1, \ldots, p)$ *is the largest number of observations such that the rank of the corresponding submatrix of X is k while the rank of the corresponding submatrix of* $(X:y)$ *is* $k + p - j$.

Since $r(X:y) = k + p$, we obtain that $k \leqslant s_p < s_{p-1} < \ldots < s_1 < n$.

THEOREM 2. *Let* $y = (y_1, \ldots, y_n)'$ *be a sample of n independent replications from a p-variate Student-t distribution in* (2·2), *and consider the prior in* (2·3)–(2·4). *Assuming that* $r(X:y) = k + p$, *defining*

$$m = \max_{j=1,\ldots,p} \left\{ j \frac{n-k}{n-s_j} - p \right\},$$

*and recalling Definition* 1 *for* $s_1, \ldots, s_p$, *we obtain that*

(i) *if* $m=0$, *then* $p(y)<\infty$;

(ii) *if* $m>0$, *then* $p(y)<\infty$ *if and only if* $P_v(0,m]=0$ *and*

$$\int_m^{m+\rho}(v-m)^{-q}\,dP_v<\infty,$$

*for all* $\rho>0$, *where* $q$ *denotes the number of indices* $j\in\{1,\ldots,p\}$ *for which*

$$m=j\frac{n-k}{n-s_j}-p.$$

From Definition 1 we note that $s_j=k+p-j$, which implies $m=0$, for all $y\in\Re^{n\times p}$ excluding a set of Lebesgue measure zero. Thus, Theorem 2(i) will apply and inference is feasible with almost all samples, as was already clear from Theorem 1. However, as will be illustrated in the examples in § 4, observed samples often lead to values of $m>0$, in which case Theorem 2(ii) indicates that $P_v$ cannot put any mass on values of $v\leqslant m$. As an immediate consequence, inference based on samples for which $m>0$ is precluded under any prior $P_v$ with support including $(0,K)$ for some $K>0$. This negative result extends to improper priors for $v$. Thus, the Bayesian model considered in Theorem 2 with popular choices for $P_v$ such as the improper uniform on $\Re_+$, Jeffreys' prior (Liu, 1995) or distributions in the gamma family (Geweke, 1993) can never lead to a posterior distribution whenever $m>0$ for the particular sample of point observations under consideration. Bounding $v$ away from zero by some fixed constant (Rellers & Rogers, 1977; Liu, 1995, 1996) provides no general solution either, since $m$ is typically updated as sample size grows and there exists no upper bound for $m$ independent of sample size. This continual updating of $m$ has the rather shocking consequence that adding new observations can actually destroy the propriety of a posterior which was proper with the previous sample!

An intuitive interpretation of Theorem 2 is most easily provided for the case of univariate regression, i.e. when $p=1$. The quantity $m$ then simplifies to $m=(s_1-k)/(n-s_1)$, where $s_1\geqslant k$ is the largest possible number of observations for which $y_i$ can be fitted exactly by $\beta'x_i$ for some fixed value of $\beta$. The value of $m$ thus increases with the proportion of observations in the sample that allow for a perfect fit for some parameter values. As this proportion increases, the likelihood can take higher values in neighbourhoods of the parameter values leading to perfect fit, thus making integrability over such regions more difficult. To compensate for this behaviour, prior assumptions need to become more stringent as $m$ increases, as stated in Theorem 2(ii). Note that the value of $m$ is updated as we get new observations and can become arbitrarily large as $n$ increases, thus preventing any prior $P_v$ from ensuring a finite $p(y)$ for each possible sample $y$.

If we specialise further to $k=1$ and take $x_i=1$, we are in the univariate location-scale model analysed in Fernández & Steel (1999). In this case, $m>0$ as soon as the sample contains any repeated observations.

We stress that the problems described in this section stem from the fact that point observations have zero probability under the assumed sampling model, and thus the existence of the conditional distribution stated in Theorem 1 does not guarantee that $p(y)<\infty$ for a particular observed sample. This is a general problem, that can arise under many different combinations of sampling models and priors, and even under proper priors if the likelihood is unbounded. Since the Student-$t$ distribution with unknown degrees of freedom leads to an unbounded likelihood function, as shall be shown in § 5, we cannot exclude that $p(y)=\infty$ for certain samples even under proper priors. As an example of this, consider a Student-$t$ sampling model in (2·2) with the prior in (2·3)–(2·4), where $P_\nu$ has mass on $(M, \infty)$ for some $M > 0$. Let $y_0$ be an initial sample for which $m < M$. Then, by Theorem 2, $p(y_0) < \infty$, and the corresponding posterior distribution can be used as a proper prior for the next sample. However, despite the propriety of this latter prior, Bayesian inference will be precluded under any new sample $y_1$ such that $m$ corresponding to $(y_0', y_1')'$ is bigger than $M$; see Theorem 2(ii).

An interesting alternative to the prior in (2·3)–(2·4) consists of choosing a proper inverted Wishart distribution for $\Omega$, while retaining a flat prior for $\beta$ and any proper prior for $\nu$. In this case, we can show that any sample of size $n = k$ from (2·2), with any choice of $P_{\lambda_i|\nu}$, leads to $p(y) < \infty$ and thus allows for posterior inference. However, as we obtain more observations, different things can happen depending on the particular sampling density in (2·2) and the prior $P_\nu$. Under normal sampling, any new sample will allow for posterior inference, and the same holds under Student-$t$ sampling if $p = 1$ or 2, or if the support of $P_\nu$ is bounded away from zero. However, under Student-$t$ sampling with $p > 2$ and certain priors $P_\nu$ with mass arbitrarily close to zero, the arrival of new observations can destroy the propriety of the previously proper posterior distribution.

## 4. Bayesian inference using set observations

A main conclusion from § 3 is that the appealing coherency properties that Bayesian inference inherits from the use of probability or measure theory cannot be guaranteed to hold if we condition on a zero probability event. A formal solution to these problems is to consider 'set' observations, which have positive probability under the continuous sampling model. In practice, it seems natural to consider a neighbourhood $S_i$ of the recorded point observation $y_i$ on the basis of the precision of the measuring device. With set observations, posterior inference is always guaranteed under any proper prior. For the improper prior in (2·3)–(2·4) a formal examination leads to the following theorem.

Theorem 3. *Consider the Bayesian model* (2·2)–(2·4) *and n compact sets* $S_i$ $(i = 1, \ldots, n)$ *of positive Lebesgue measure in* $\Re^p$, *homeomorphic to hypercubes. Then*

$$P(y_1 \in S_1, \ldots, y_n \in S_n) < \infty$$

*if and only if* $r(X : y) = k + p$ *for all* $y_1 \in S_1, \ldots, y_n \in S_n$.

Note that neither the mixing distribution $P_{\lambda_i|\nu}$ nor the prior of $\nu$, $P_\nu$, intervenes in Theorem 3, which thus holds for any scale mixture of normals. In contrast to the analysis with point observations, Bayesian inference is now fully coherent in that adding extra observations can never destroy the possibility of conducting inference.

The condition $r(X : y) = k + p$, which was always necessary under point observations, see Lemma 1, becomes both necessary and sufficient when extended to sets as in Theorem 3. In the case of a location-scale model, with $k = 1$ and $x_i = 1$, the latter condition is equivalent to the absence of a $(p - 1)$-dimensional affine space that intersects with all of the sets $S_1, \ldots, S_n$. Figure 1 graphically illustrates this issue in the bivariate case, $p = 2$; while the set observations in Fig. 1(a) allow for posterior inference, the latter is precluded in Fig. 1(b).

In general, Bayesian inference using set observations can easily be implemented through a Gibbs sampler on the parameters augmented with $y = (y_1, \ldots, y_n)'$, always conditioning on the set observations $y_1 \in S_1, \ldots, y_n \in S_n$. This amounts to drawing $y_i^*$ $(i = 1, \ldots, n)$ independently from the sampling distribution in (2·2) truncated to the set observation

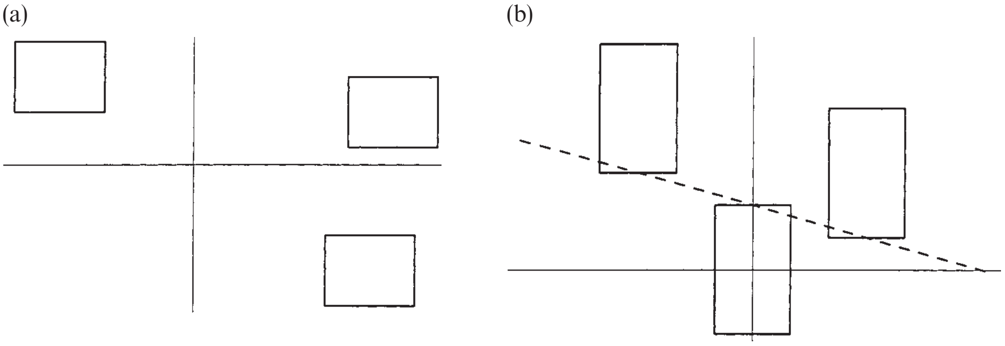

Fig. 1. (a) Posterior inference through set observations. (b) No posterior inference through set observations.

$S_i$, and the parameters from the usual posterior distribution given the drawings $(y_1^*, \ldots, y_n^*)'$. In many cases, it will prove convenient to augment further with the mixing variables $\lambda_i$ $(i = 1, \ldots, n)$ introduced in (2·2).

Using a Gibbs sampler along these lines, we can easily analyse the following three examples under Student-$t$ sampling with the prior in (2·3)–(2·4). In all cases, we take

$$P_\nu = \text{exponential with mean } 10, \tag{4·1}$$

which spreads the prior mass over a wide variety of tail behaviour.

*Example* 1: *Stackloss data.* This classical dataset, originally presented in Brownlee (1965), was treated under Student-$t$ sampling by Lange et al. (1989) in a classical maximum likelihood framework. The data consist of $n = 21$ observations of a univariate response, stackloss, given an intercept and three other regressors. Thus, $p = 1$ and $k = 4$. Recalling Definition 1, we can derive that for this dataset $s_1 = 8$.

Thus, $m$ in Theorem 2 takes the value $\frac{4}{13}$, precluding Bayesian inference on the basis of these point observations under Student-$t$ sampling with the prior in (2·3) and (4·1), and we resort to set observations derived in accordance with the precision implicit in the number of digits recorded. In order to verify that these set observations fulfil the condition stated in Theorem 3, note that there exist three observations $y_i$, $y_j$ and $y_l$, say, with all the regressors equal except for one, $x_{iq} \neq x_{jq} \neq x_{lq}$ say, and such that

$$\max_{\substack{y_i \in S_i \\ y_l \in S_l}} \left( \frac{y_i - y_l}{x_{iq} - x_{lq}} \right) < \min_{\substack{y_i \in S_i \\ y_j \in S_j}} \left( \frac{y_i - y_j}{x_{iq} - x_{jq}} \right).$$

By way of illustration, Fig. 2 summarises posterior inference on the degrees of freedom $\nu$.

The mixing variables $\lambda_i$ in (2·2) can be seen as observation-specific precision factors, so that unusually small values of $\lambda_i$ correspond to ‘outlying’ observations. The EM algorithm used in Lange et al. (1989) takes the mean of the conditional distribution of $\lambda_i$ given all the other parameters and $y$ as the weight of observation $i$; see also Pettitt (1985) and West (1984). On the basis of these weights Lange et al. (1989) identify observations 21, 4, 3 and 1 as outliers, as in the robust analysis of Andrews (1974). In a Bayesian set-up, we naturally focus on the marginal posterior distribution of the $\lambda_i$’s and find indeed that the posterior means for these four observations are considerably lower than for the others. However, we also note that the posterior distributions of the $\lambda_i$’s display a substantial spread.

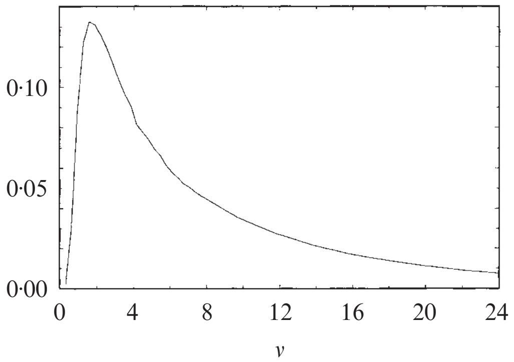

Fig. 2. Example 1: Stackloss data. Posterior density of $v$.

*Example* 2: *Fisher's Iris setosa data.* This dataset, consisting of $n = 50$ bivariate measurements of petal length and width for Iris setosa, was analysed in Fisher (1936) and Heitjan (1989). These data will simply be modelled as a bivariate location-scale model; thus, $p = 2$, $k = 1$ and $x_i = 1$. The original measurements were transformed to logarithms, as suggested in Gnanadesikan (1977, p. 219) and Heitjan (1989), and the set observations were transformed accordingly. Heitjan (1989) advocates the use of grouped likelihood for this example and maximises a normal likelihood integrated over the respective sets $S_i$ $(i = 1, \ldots, n)$. For these data, we can easily ascertain that $s_1 = 35$ and $s_2 = 29$, see Definition 1, which implies that $m = \frac{8}{3}$, and thus Theorem 2(ii) again indicates that point observations cannot form the basis of a Bayesian analysis under Student-$t$ sampling with the prior (2·3) and (4·1). The use of set observations leads for instance to the posterior densities plotted in Fig. 3, where we focus only on small values of $v$, and 'correlation' denotes the off-diagonal element of $\Omega$ divided by the square root of the product of the diagonal elements.

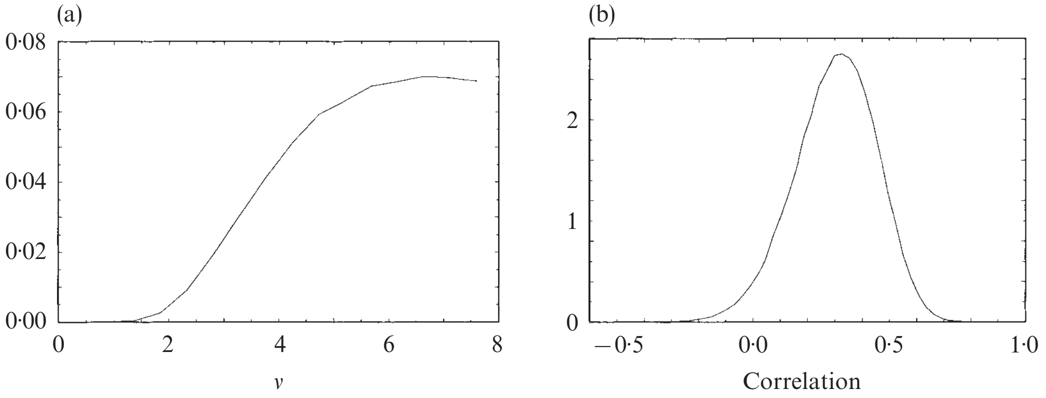

Fig. 3. Example 2: Iris data. (a) Posterior density of $v$. (b) Posterior density of correlation.

A natural extension of our context of set observations is that of missing observations. For a maximum likelihood analysis of this problem, the EM algorithm was used by Dempster, Laird & Rubin (1977), while Bayesian approaches rely on data augmentation (Tanner & Wong, 1987) or imputation methods (Rubin, 1987; Kong, Liu & Wong, 1994). Whereas so far we considered observations consisting of bounded sets $S_i$, the fact that some components of $y_i$ are missing implies that the corresponding set observation becomes unbounded, in the direction of each missing component. Suppose that $r < n$ observations lead to compact sets, while $n - r$ observations contain unobserved elements.

Theorem 4. *Consider the Bayesian model in* (2·2)–(2·4). *The observations consist of $r < n$ compact sets $S_1, \ldots, S_r$, whereas $S_{r+1}, \ldots, S_n$ are unbounded because of missing components. All n sets have positive Lebesgue measure in $\Re^p$, and are homeomorphic to hypercubes. Defining $X_{(r)} = (x_1, \ldots, x_r)'$ and $y_{(r)} = (y_1, \ldots, y_r)'$, we obtain*

(i) *if $r(X_{(r)} : y_{(r)}) = k + p$, for all $y_1 \in S_1, \ldots, y_r \in S_r$, then $P(y_1 \in S_1, \ldots, y_n \in S_n) < \infty$;*

(ii) *if we can find values $y_1 \in S_1, \ldots, y_n \in S_n$ for which $r(X : y) < k + p$, then $P(y_1 \in S_1, \ldots, y_n \in S_n) = \infty$.*

From Theorems 3 and 4(i) we immediately deduce that, whenever the compact set observations $S_1, \ldots, S_r$ lead to a proper posterior, the same holds if we add the unbounded set observations $S_{r+1}, \ldots, S_n$, corresponding to missing data. Clearly, adding observations that do not contradict the sampling model, i.e. of positive probability, can never destroy the existence of an already well-defined posterior distribution. On the other hand, Theorem 4(ii) says that the necessary condition stated in Theorem 3 for compact sets extends to any sample of sets $S_1, \ldots, S_n$, possibly unbounded.

As explained in the discussion of Theorem 3, the assumption of Theorem 4(ii) is most easily interpreted in the location-scale case, that is $k = 1$ and $x_i = 1$, where $r(X : y) < 1 + p$ means that there exists a $(p - 1)$-dimensional affine space that intersects all of the sets $S_1, \ldots, S_n$. Note that, if we can find one such space that, in addition, intersects with all $p$ coordinate axes, any new set corresponding to missing data will necessarily have an intersection with this $(p - 1)$-dimensional affine space. Thus, adding any number of sets corresponding to observations with missing components can never result in a posterior distribution. Figure 4(a) graphically illustrates this point in the bivariate case, $p = 2$.

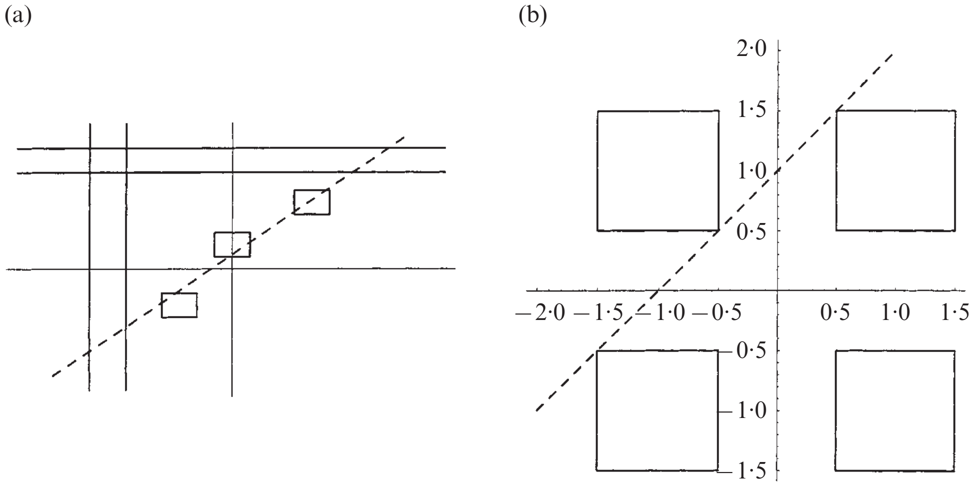

Fig. 4. (a) Set observations with missing components do not help. (b) Example 3. Compact set observations in artificial data.

*Example* 3: *Augmented Murray data.* In this example, we focus on the artificial data originally introduced by Murray (1977) and augmented with four extreme values by Liu

& Rubin (1995). This results in the following bivariate dataset:

$$\begin{array}{lcccccccccccccccc} y_{i1} & -1 & -1 & 1 & 1 & -2 & -2 & 2 & 2 & ? & ? & ? & ? & -12 & 12 & ? & ? \\ y_{i2} & -1 & 1 & -1 & 1 & ? & ? & ? & ? & -2 & -2 & 2 & 2 & ? & ? & -12 & 12 \end{array} \qquad (4{\cdot}2)$$

where $n=16$, $p=2$ and ? denotes a missing value. We use a location-scale model under Student-$t$ sampling, as in the maximum likelihood analysis of Liu & Rubin (1995) and the Bayesian analysis of Liu (1995). Here, Bayesian inference will be conducted under the prior (2·3) and (4·1), using set observations with unitary width for the observed components.

Figure 4(b) depicts the four compact sets, corresponding to the first four observations, and from Theorem 4(i) we can immediately deduce properness of the posterior as no single line can cross all four sets $S_1, \ldots, S_4$. Note that deleting any one of these compact set observations would result in an improper posterior, see Theorem 4(ii) and the discussion thereafter, as indicated by the dashed line in Fig. 4(b). Figure 5 plots posterior densities for the degrees of freedom $v$ and the correlation, as defined in Example 2. Compared to the same analysis on the basis of the original Murray data, i.e. the first 12 observations in (4·2), the correlation is more extreme and degrees of freedom tend to be substantially fewer. As expected, the analysis based on all sixteen set observations identifies the four extra observations as outliers through small values of the mixing variable $\lambda_i$ associated with these observations.

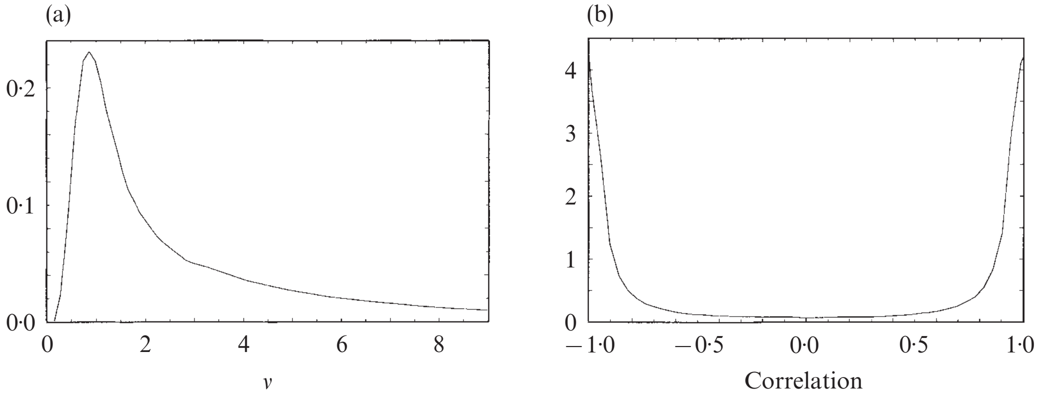

Fig. 5. Example 3: Artificial data. (a) Posterior density of $v$. (b) Posterior density of correlation.

## 5. The Student-$t$ likelihood function

In this section, we examine some peculiarities of the likelihood function corresponding to independent Student-$t$ sampling in a general regression context. We shall focus only on the use of point observations, and present some classical counterparts of the problems described in § 3 under a Bayesian treatment of this model.

In particular, we consider $n$ replications from the following sampling density function for $y_i \in \Re^p$:

$$p(y_i \mid \beta, \Omega, v) = \frac{\Gamma\{(v+p)/2\}}{\Gamma(v/2)(\pi v)^{p/2}|\Omega|^{1/2}} \left[ 1 + \frac{1}{v}\{y_i - g_i(\beta)\}'\Omega^{-1}\{y_i - g_i(\beta)\} \right]^{-(v+p)/2}, \qquad (5{\cdot}1)$$

where $\beta \in \mathscr{B}$ and $g_i(.)$ is a known continuous function from $\mathscr{B}$ to $\Re^p$, possibly depending on regressors $x_i$. Thus, we extend the linear regression context of the previous sections to more general regression functions. We shall reparameterise the matrix $\Omega$ as $(\sigma, V)$ through

$$\Omega = \sigma^2 V, \tag{5·2}$$

where $\sigma > 0$ and $V \in \mathscr{C}_1^p$, which denotes the set of $p \times p$ positive definite symmetric matrices with element $(1, 1)$ equal to one. This reparameterisation is useful for presenting the main result of this section.

THEOREM 5. *Let $l(\beta, \sigma, V, \nu)$ be the likelihood function corresponding to n independent replications from* (5·1) *with the reparameterisation in* (5·2). *Then we have the following.*

(i) *The function $l(\beta, \sigma, V, \nu)$ is finite and continuous in the entire parameter space $\mathscr{B} \times (0, \infty) \times \mathscr{C}_1^p \times (0, \infty)$.*

(ii) *For given values $\beta = \beta_0$, $V = V_0$ and $\nu = \nu_0$, let $0 \leqslant s(\beta_0) \leqslant n$ be the number of observations for which $y_i = g_i(\beta_0)$. We obtain*

(a) *if*

$$\nu_0 < \frac{s(\beta_0)p}{n - s(\beta_0)},$$

*then*

$$\lim_{\sigma \to 0} l(\beta_0, \sigma, V_0, \nu_0) = \infty;$$

(b) *if*

$$\nu_0 = \frac{s(\beta_0)p}{n - s(\beta_0)},$$

*then*

$$\lim_{\sigma \to 0} l(\beta_0, \sigma, V_0, \nu_0) \in (0, \infty);$$

(c) *if*

$$\nu_0 > \frac{s(\beta_0)p}{n - s(\beta_0)},$$

*then*

$$\lim_{\sigma \to 0} l(\beta_0, \sigma, V_0, \nu_0) = 0.$$

From Theorem 5, whenever we can find a value $\beta_0$ such that $y_i = g_i(\beta_0)$ holds for at least one observation, the likelihood function does not possess a global maximum. Indeed, for small enough values of $\nu$, see Theorem 5(ii)(a), we can make $l(\beta_0, \sigma, V_0, \nu_0)$ arbitrarily large by letting $\sigma$ tend to zero. Note that, in practice, we can typically find values $\beta_0$ such that $s(\beta_0) > 0$. For example, in the case of $p$-variate linear regression with $k$ regressors we can deduce from Definition 1 that

$$\max_{\beta \in \Re^{k \times p}} s(\beta) = s_p \geqslant k, \tag{5·3}$$

thus precluding global maximisation of the likelihood function, irrespective of the sample.

This finding casts some doubt on maximum likelihood estimation under Student-*t* regression models with unknown degrees of freedom $\nu$. In the existing literature, $\nu$ is typically allowed to vary in $\Re_+$ (Lange et al., 1989; Lange & Sinsheimer, 1993). Reported maximum likelihood estimates must, therefore, correspond to local and not to global maxima, although this is not stated in these papers. To our knowledge, the existence and uniqueness of such local maxima and the asymptotic properties of the corresponding estimators have not been formally established in the literature. In a pure location-scale context with fixed degrees of freedom, constrained to be sufficiently large, Maronna (1976) proves that the likelihood equations have a unique solution that leads to a consistent and asymptotically normal estimator of $(\beta, \Omega)$. However, we have not encountered similar results for unknown $\nu$.

We remind the reader that a Bayesian analysis of the linear regression model in Theorem 2 based on point observations breaks down if we assign prior probability to values of $\nu \leqslant m$. From Theorem 5(ii) with (5·3) the likelihood is unbounded if $\nu < s_p p/(n - s_p)$, which can generally be larger or smaller than $m$. In the case of univariate linear regression, $p = 1$, the latter quantity becomes $s_1/(n - s_1)$, which is always larger than $m = (s_1 - k)/(n - s_1)$. Thus, in this case, the likelihood is still integrable with the prior in (2·3)–(2·4) if $P_\nu$ bounds $\nu$ strictly away from $(s_1 - k)/(n - s_1)$, see Theorem 2, but is unbounded for values of $\nu$ smaller than $s_1/(n - s_1)$. Furthermore, there is a fundamental difference between classical and Bayesian results: as remarked in the discussion of Theorem 2, $m$ equals zero for all $y \in \Re^{n \times p}$ except for a set of Lebesgue measure zero, implying that a Bayesian analysis is feasible for almost all samples of size $n \geqslant k + p$; see also Theorem 1. From Theorem 5(ii)(a) and (5·3), on the other hand, it is immediately clear that global maximisation of the likelihood is precluded for any sample $y \in \Re^{n \times p}$.

*Example* 1 (*cont.*): *Stackloss data.* As explained in Example 1, $s_1 = 8$; the value $\beta_0 = (-36, 0{\cdot}5, 1, 0)'$ allows us to exactly fit eight of the 21 observations. Thus, from Theorem 5(ii)(a), taking $\nu_0 < \frac{8}{13}$ leads to $\lim_{\sigma \to 0} l(\beta_0, \sigma, 1, \nu_0) = \infty$; note that $p = 1$ implies $V_0 = 1$. Lange et al. (1989) estimate $\nu$ to be 1·1, which presumably corresponds to a local maximum of the likelihood.

In some cases, numerical optimisation procedures, such as the EM algorithm, may attempt to converge to an area with unbounded likelihood. A case in point is the analysis of the radioimmunoassay data in Lange et al. (1989) and Lange & Sinsheimer (1993). This concerns a nonlinear regression model with $p = 1$, that is univariate, and 4 regression parameters introduced in Tiede & Pagano (1979), where the $n = 14$ data points are listed. Whereas Lange et al. (1989, p. 883) already report that 'ML estimation of $\nu$ for this data is not very satisfactory' and report a maximum likelihood estimate of $\nu$ equal to 0·29, Lange & Sinsheimer (1993) report maximum likelihood estimates of $\nu$ equal to 0·05 and of $\sigma$ equal to 0. The latter also state that 10 of the 14 weights, i.e. the conditional means of the $\lambda_i$, are found to be zero and that the EM algorithm has not converged after 300 iterations. From their estimates of $\beta$ it is clear that they exactly fit four of the observations, while they consider values of $\nu$ smaller than $\frac{4}{10}$, which takes them to a region of unbounded likelihood. Thus, Theorem 5 provides an immediate explanation for the 'potential problems with the $t$' mentioned in Lange & Sinsheimer (1993, p. 195).

When some of the components of the $p$-variate observations $y_i$ $(i = 1, \ldots, n)$ are missing, the resulting likelihood still displays the same type of behaviour as explained in Theorem 5. In particular, the latter theorem will apply in this more general context if we replace the bound $\{s(\beta_0)p\}/\{n - s(\beta_0)\}$ for $\nu$ by the quantity $\sum_{i \in \mathscr{I}} p_i/(n - \sum_{i \in \mathscr{I}} 1)$, where $p_i \leqslant p$ is the number of observed components of $y_i$, and $\mathscr{I}$ is the set of indices for which the observed components of $y_i$ are exactly fitted by the corresponding components of $g_i(\beta_0)$. Thus, the stationary values reported in Liu & Rubin (1994, 1995) for Student-$t$ models with unknown $v$ and missing data do not correspond to global maxima of the likelihood function.

Therefore, we feel that the use of maximum likelihood methods for Student-$t$ models with unknown degrees of freedom cannot be advocated without further careful study of the existence and properties of local maxima. Alternatively, classical inference could be based on efficient likelihood estimation (Lehmann, 1983, Ch. 6), grouped likelihoods (Beckman & Johnson, 1987), modified likelihood (Cheng & Iles, 1987) or spacings methods. For a general discussion of nonregular likelihood problems, see Cheng & Traylor (1995).

## Acknowledgement

We gratefully acknowledge the extremely valuable help of F. Chamizo in the proof of Theorem 3 as well as useful comments from the referee, the associate editor, the editor, P. J. Green, B. Melenberg and W. J. Studden. Both authors were affiliated with the CentER for Economic Research, Tilburg University, the Netherlands, when most of this work took place, and benefited from a travel grant awarded by the Netherlands Organization for Scientific Research and the hospitality of the Department of Statistics at Purdue University. C. Fernández was also supported by a Training and Mobility of Researchers fellowship, financed by the European Commission.

## Appendix

### *A useful lemma and sketched proofs*

Lemma 1. *Under the assumptions of Theorem* 1, $p(y)<\infty$ *if and only if* $r(X:y)=k+p$ *and*

$$\int_{\mathcal{N}}\int_{(0,\infty)^n}\left(\prod_{i\neq m_1,\ldots,m_{k+p}}\lambda_i^{p/2}\right)\left(\prod_{i=k+1}^{k+p}\lambda_{m_i}^{-(n-k-p)/2}\right)dP_{\lambda_1|v}\ldots dP_{\lambda_n|v}\,dP_v<\infty, \tag{A·1}$$

*where*

$$\prod_{i=1}^{k}\lambda_{m_i}\equiv\max\left\{\prod_{i=1}^{k}\lambda_{l_i}:|x_{l_1}\ldots x_{l_k}|\neq 0\right\}, \tag{A·2}$$

$$\prod_{i=1}^{k+p}\lambda_{m_i}\equiv\max\left\{\prod_{i=1}^{k+p}\lambda_{l_i}:\begin{vmatrix}x_{l_1} & \cdots & x_{l_{k+p}}\\ y_{l_1} & \cdots & y_{l_{k+p}}\end{vmatrix}\neq 0\right\}. \tag{A·3}$$

*Proof.* Consider the joint distribution of $(y,\beta,\Omega,v,\lambda_1,\ldots,\lambda_n)$. We first integrate out $\beta$ as a matricvariate normal distribution. Next, integrability of $\Omega$ requires $r(X:y)=k+p$, in which case $\Omega$ has an inverted Wishart distribution. This leaves us with

$$|\Lambda|^{p/2}|X'\Lambda X|^{(n-k-p)/2}|L'\Lambda L|^{-(n-k)/2}, \tag{A·4}$$

where $\Lambda=\text{diag}(\lambda_1,\ldots,\lambda_n)$ and $L=(X:y)$, which we need to integrate with respect to $P_{\lambda_1|v}\ldots P_{\lambda_n|v}P_v$. Applying the Binet–Cauchy formula (Gantmacher, 1959, p. 9), we deduce that $|X'\Lambda X|$ has upper and lower bounds both proportional to $\prod_{i=1}^{k}\lambda_{m_i}$, defined in (A·2), whereas $|L'\Lambda L|$ has upper and lower bounds both proportional to $\prod_{i=1}^{k+p}\lambda_{m_i}$, defined in (A·3). This implies that integrability of the expression in (A·4) is equivalent to (A·1). □

*Proof of Theorem* 1. The conditional distribution of $(\beta,\Omega,v)$ exists if and only if $p(y)<\infty$ for all $y\in\Re^{n\times p}$, possibly excluding a set of Lebesgue measure zero. From Lemma 1, $p(y)<\infty$ requires $n\geqslant k+p$. If we assume $n\geqslant k+p$, then, for all $y\in\Re^{n\times p}$ excluding a set of Lebesgue measure zero,

$r(X:y)=k+p$ and

$$\max\{\lambda_i : i \neq m_1, \ldots, m_{k+p}\} \leqslant \min\{\lambda_{m_i} : i = k+1, \ldots, k+p\},$$

see (A·2) and (A·3), which implies a bounded integrand in (A·1). Sufficiency of $n \geqslant k+p$ is now immediate from Lemma 1. □

*Proof of Theorem* 2. We need to check whether or not (A·1) holds. Consider the following upper bound for the inside integral in (A·1), with $f_G(\lambda \mid v/2, v/2)$ denoting the density function of a $\mathrm{Ga}(v/2, v/2)$ distribution:

$$n! \int_0^\infty \lambda^{p(n-k-p)/2} f_G\left(\lambda \,\middle|\, \frac{v}{2}, \frac{v}{2}\right) d\lambda \int_0^\infty \lambda^{-p(n-k-p)/2} f_G\left(\lambda \,\middle|\, \frac{v}{2}, \frac{v}{2}\right) d\lambda,$$

which is a bounded function of $v$ for $v \geqslant C$, where $C$ is an arbitrary constant bigger than $p(n-k-p)$. Thus, integrability over this region follows immediately.

In order to examine integrability for $v < C$, we consider all $n!$ possible orderings of $\lambda_1, \ldots, \lambda_n$. It is enough to focus on those orderings for which $\lambda_{(n-s_j:n)} = \lambda_{m_{k+j}}$ for all $j = 1, \ldots, p$, where $s_j$ was given in Definition 1 and $\lambda_{(i:n)}$ denotes the $i$th order statistic, since they lead to the largest value for the integrand in (A·1). For any such ordering, we evaluate the integral in (A·1) iteratively, using the upper and lower bounds

$$\exp(-b\lambda_{i+1}) \frac{\lambda_{i+1}^a}{a} \leqslant \int_0^{\lambda_{i+1}} \lambda_i^{a-1} \exp(-b\lambda_i)\, d\lambda_i \leqslant \frac{\lambda_{i+1}^a}{a},$$

for any $a, b > 0$. This leads to Theorem 2. □

*Proof of Theorem* 3. After integrating out $\beta$ and $\Omega$ as in the proof of Lemma 1, we need to integrate the expression in (A·4) over $y_1 \in S_1, \ldots, y_n \in S_n$ and with respect to $P_{\lambda_1|v} \ldots P_{\lambda_n|v} P_v$. Since $r(X) = k$ we assume, without loss of generality, that $|x_1, \ldots, x_k| \neq 0$. Defining

$$\eta = y_{-k} - X_{-k} X_{(k)}^{-1} y_{(k)} \in \Re^{(n-k)\times p},$$

where $y_{-k} = (y_{k+1}, \ldots, y_n)'$, $y_{(k)} = (y_1, \ldots, y_k)'$ and similarly for $X$, allows us to rewrite (A·4) as

$$|\Lambda|^{p/2} |X'\Lambda X|^{-p/2} |\eta' Q(\Lambda) \eta|^{-(n-k)/2}, \qquad (\text{A·5})$$

where $Q(\Lambda) = \Lambda_{-k} - \Lambda_{-k} X_{-k} (X'\Lambda X)^{-1} X'_{-k} \Lambda_{-k}$, with $\Lambda_{-k} = \mathrm{diag}(\lambda_{k+1}, \ldots, \lambda_n)$. We now need to integrate (A·5) in $(y_{(k)}, \eta)$ over the appropriate sets, and with respect to $P_{\lambda_1|v} \ldots P_{\lambda_n|v} P_v$.

*Necessity*. By the definition of $\eta$, the assumption that there exist $y_1 \in S_1, \ldots, y_n \in S_n$ for which $r(X:y) < k+p$ is equivalent to a zero value for the minimum of $|\eta' Q(\Lambda) \eta|$. A change of variables from $\eta$ to an upper triangular matrix $T$ of rank $p$ such that $T'T = \eta' Q(\Lambda) \eta$, completed with the required extra variables, allows us to prove that this implies an infinite integral for the expression in (A·5).

*Sufficiency*. The condition $r(X:y) = k+p$ for all $y_1 \in S_1, \ldots, y_n \in S_n$, implies that $|\eta'\eta| \geqslant A > 0$ for some positive constant $A$, which, in turn, means that there always exists a submatrix of $\eta$ of order $p$ with determinant strictly bounded away from zero. Let us for example define $\eta = (\eta'_{(p)}, \eta'_{-p})'$ and consider the region where $|\eta_{(p)}|^2 \geqslant B > 0$ for some constant $B$. Then we can directly integrate out $\eta_{-p}$ from (A·5) as a matricvariate Student-$t$ distribution, which simply leaves us with $(|\eta_{(p)}|^2)^{-p/2}$. Since the latter expression is bounded, we directly obtain a finite integral under any $P_{\lambda_1|v} \ldots P_{\lambda_n|v} P_v$. □

*Proof of Theorem* 4. (i) We have $P(y_1 \in S_1, \ldots, y_n \in S_n) \leqslant P(y_1 \in S_1, \ldots, y_r \in S_r)$, which is finite by Theorem 3.

(ii) This is immediate, since the proof of the necessity in Theorem 3 never uses the fact that the sets are compact. □

The proof of Theorem 5 follows immediately from writing down the likelihood function.

## REFERENCES

ANDREWS, D. F. (1974). A robust method for multiple linear regression. *Technometrics* **16**, 523–31.

BECKMAN, R. J. & JOHNSON, M. E. (1987). Fitting the Student-*t* distribution to grouped data, with application to a particle scattering experiment. *Technometrics* **29**, 17–22.

Brownlee, K. A. (1965). *Statistical Theory and Methodology in Science and Engineering*, 2nd ed. New York: John Wiley.

Casella, G. & George, E. (1992). Explaining the Gibbs sampler. *Am. Statistician* **46**, 167–74.

Cheng, R. C. H. & Iles, T. C. (1987). Corrected maximum likelihood in non-regular problems. *J. R. Statist. Soc.* B **49**, 95–101.

Cheng, R. C. H. & Traylor, L. (1995). Non-regular maximum likelihood problems (with Discussion). *J. R. Statist. Soc.* B **57**, 3–44.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with Discussion). *J. R. Statist. Soc.* B **39**, 1–38.

Fernández, C. & Steel, M. F. J. (1999). On the dangers of modelling through continuous distributions: A Bayesian perspective (with Discussion). In *Bayesian Statistics 6*, Ed. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith. To appear. Oxford: Oxford University Press.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugen.* **8**, 179–88.

Gantmacher, F. R. (1959). *The Theory of Matrices*, **1**. New York: Chelsea.

Gelfand, A. E. & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Am. Statist. Assoc.* **85**, 398–409.

Geweke, J. (1993). Bayesian treatment of the independent Student-*t* linear model. *J. Appl. Economet.* **8**, S19–S40.

Gnanadesikan, R. (1977). *Methods for Statistical Data Analysis of Multivariate Observations.* New York: John Wiley.

Heitjan, D. F. (1989). Inference from grouped continuous data: A review (with Discussion). *Statist. Sci.* **4**, 164–83.

Kong, A., Liu, J. S. & Wong, W. H. (1994). Sequential imputations and Bayesian missing data problems. *J. Am. Statist. Assoc.* **89**, 278–88.

Lange, K. L., Little, R. J. A. & Taylor, J. M. G. (1989). Robust statistical modeling using the *t* distribution. *J. Am. Statist. Assoc.* **84**, 881–96.

Lange, K. L. & Sinsheimer, J. S. (1993). Normal/independent distributions and their applications in robust regression. *J. Comp. Graph. Statist.* **2**, 175–98.

Lehmann, E. L. (1983). *Theory of Point Estimation.* New York: John Wiley.

Liu, C. H. (1995). Missing data imputation using the multivariate *t* distribution. *J. Mult. Anal.* **53**, 139–58.

Liu, C. H. (1996). Bayesian robust multivariate linear regression with incomplete data. *J. Am. Statist. Assoc.* **91**, 1219–27.

Liu, C. H. & Rubin, D. B. (1994). The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence. *Biometrika* **81**, 633–48.

Liu, C. H. & Rubin, D. B. (1995). ML estimation of the multivariate *t* distribution with unknown degrees of freedom. *Statist. Sinica* **5**, 19–39.

Maronna, R. (1976). Robust M-estimators of multivariate location and scatter. *Ann. Statist.* **4**, 51–67.

Murray, G. D. (1977). Comment on 'Maximum likelihood from incomplete data via the EM algorithm' by A. P. Dempster, N. M. Laird and D. B. Rubin. *J. R. Statist. Soc.* B **39**, 27–8.

Pettitt, A. N. (1985). Re-weighted least squares estimation with censored and grouped data: An application of the EM algorithm. *J. R. Statist. Soc.* B **47**, 253–60.

Relles, D. A. & Rogers, W. H. (1977). Statisticians are fairly robust estimators of location. *J. Am. Statist. Assoc.* **72**, 107–11.

Rubin, D. B. (1987). A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing observations are modest: The SIR algorithm, Comment on Tanner and Wong (1987). *J. Am. Statist. Assoc.* **82**, 543–6.

Tanner, M. A. & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with Discussion). *J. Am. Statist. Assoc.* **82**, 528–50.

Tiede, J. J. & Pagano, M. (1979). The application of robust calibration to radioimmunoassay. *Biometrics* **35**, 567–74.

West, M. (1984). Outlier models and prior distributions in Bayesian linear regression. *J. R. Statist. Soc.* B **46**, 431–9.

[*Received June* 1997. *Revised February* 1998]