Robust Bayesian Analysis and Optimal Experimental Designs
In Normal Linear Models with Many Parameters – I

by

Anirban DasGupta* and W. J. Studden**
Purdue University

Technical Report #88-14

Department of Statistics
Purdue University

July 1988

* Research supported by NSF Grant DMS-8702620.
** Research supported by NSF Grant DMS-8802535.

◇

Robust Bayesian Analysis and Optimal Experimental Designs
In Normal Linear Models with Many Parameters – I

Abstract

In a treatment of Bayesian robustness issues in regression problems, we work out the effects of changing the prior and/or the loss function in the canonical normal linear model where $Y \sim N(X\underset{\sim}{\theta}, \sigma^2 I)$ and $(\underset{\sim}{\theta}, \sigma^2)$ is assigned a prior $\pi(\underset{\sim}{\theta}, \sigma^2)$ belonging to a suitable class Γ. The decision problems include estimation of $\underset{\sim}{\theta}$ under quadratic loss, estimation of the mean of the response variable under a quadratic or a piecewise linear loss, testing a hypothesis about $\underset{\sim}{\theta}$ etc. Two different classes of priors are considered: conjugate priors and their mixtures, and priors π such that $L \le \pi \le U$ where $L \le U$ are two fixed lower and upper envelopes. While conjugate priors and their mixtures are attractive from a mathematical viewpoint, priors between a lower and an upper envelope demonstrate highly appealing stability properties over repeated sampling, as has been considered desirable by leading researchers in the area (Berger (1984), L. Brown's discussion on Berger (1984)). A wide variety of results are proved borrowing techniques from several branches of mathematics, such as the theory of moments, analytic geometry, large sample theory, and probability inequalities for convex sets such as generalizations of the Brunn-Minkowski theorem. Some of the obtained results include:

$\underline{\mathbf{a}}$ closed form characterization (Theorems 2.1, 4.2, 6.1) of the set of Bayes estimates of $\underset{\sim}{\theta}_{p\times 1}$ under the loss $(\underset{\sim}{\theta} - \underset{\sim}{a})' Q (\underset{\sim}{\theta} - \underset{\sim}{a})$ where Q is an arbitrary known p.d. matrix where the family of priors is

$$\Gamma_1 = \Big\{\pi(\underset{\sim}{\theta}, \sigma^2):\ \ \underset{\sim}{\theta} \sim N(\underset{\sim}{\mu}, \sigma^2\Sigma),\ \Sigma_1 \le \Sigma \le \Sigma_2, \underset{\sim}{\mu} \in C \text{ for a suitable convex set } C,\ \sigma^2 > 0 \text{ is known}\Big\},$$

$$\text{or } \Gamma_2 = \Big\{\pi(\underset{\sim}{\theta}, \sigma^2):\ \ L(\underset{\sim}{\theta}, \sigma^2) \le \pi \le kL(\underset{\sim}{\theta}, \sigma^2), k > 1, L = L_1(\underset{\sim}{\theta}|\sigma^2) \cdot L_2(\sigma^2), \text{ where } L_1 \text{ is } N(\underset{\sim}{\mu}, \sigma^2\Sigma) \text{ and } L_2 \text{ is an inverse gamma density}\Big\}.$$

Under Γ_1, the set of Bayes estimates is, in general, *not* a convex set unless C contains only one point, for example, if $p = 2$ and C is a circle, then it is proved that the set

of Bayes estimates is a limacon of Pascal. Under Γ_2, the Bayes estimates always form an ellipsoid;

b explicit formulae (Theorems 2.1, 4.1, Corollary 6.2) for ranges of posterior measures, such as the diameter in L_2 norm of the set of Bayes estimators of $\underset{\sim}{\theta}$;

c complete characterization (Theorem 2.3, and results leading to Example 15) of the joint set of posterior means and posterior variances of a linear combination $\underset{\sim}{c}'\underset{\sim}{\theta}$; under Γ_2, such a characterization is intimately connected with the solution of the Markov-Krein-Stieltjes moment problem;

d ranges of posterior probabilities of suitable subsets of the parameter space (Theorems 2.5, 2.6, 6.3);

e asymptotic behavior of the diameter of the set of Bayes estimates as the sample size $n \to \infty$ (Examples 2, 18);

f general results (Theorems 3.1, 3.2, and results leading to Examples 9 and 10) (normal linear model setup *not* assumed) showing the effect of enlarging a family Γ of priors by considering mixtures of priors in Γ.

Only very simple examples are given to illustrate the theory but the results derived apply to trigonometric and polynomial regressions, one and two way ANOVA, and other linear models.

1. Introduction

In a statistical decision problem in the Bayesian framework, formally one has three components or inputs: a model or the likelihood function ℓ, a prior distribution π for the parameters θ, and a loss function L. Under fixed ℓ, π, and L, one computes the Bayes action by minimizing, over the action space, the (posterior) expected loss of the different possible actions. In a formal Bayes framework, this is the action that will be taken after observing the data generated by the experiment.

It is quite apparent, however, that the optimal Bayes action is guaranteed to be Bayes or optimal only under the assumed ℓ, π, and L, while evidently the assumed model, prior, and the loss are each at best only approximately valid. Therefore, given the inherent approximate nature of these inputs, it is important that we attempt to find out the effect of "reasonable" deviations from the assumed inputs in the framework of a clearly stated theory, a framework that is conceptually applicable to a large number of statistical decision problems. Huber (1964, 1967, 1973, 1977) and Hampel (1971, 1974, 1975) were among the foremost statisticians to have formulated a precisely stated and well understood general theory of robust statistics, *mainly* in terms of deviations from the assumed likelihood function ℓ. Since their monumental works, attention has focussed on robustness with respect to the other inputs of a statistical decision problem; Berger (1984), in a pioneering article, revived general interest in the very important question of robustness with respect to the prior. Robustness with respect to the loss was addressed in the excellent articles of Hwang (1985), and Brown and Hwang (1988). While a comprehensive study of decision-theoretic robustness under simultaneous variation in ℓ, π, and L is the ideal goal from the decision theory viewpoint, in this article we address the problem of robustness with respect to the prior, in the spirit of Berger (1984), Berger and Berliner (1986), Berger and Sivaganesan (1986), DeRobertis and Hartigan (1981), Leamer (1978, 1982), and Polachek (1984), etc. Other very important references in the field of robustness with respect to the prior include Berger (1987), Berger and Berliner (1984), Berger and Delampady (1988), O'Hagan and Berger (1988), Edwards, Lindman, and Savage (1963), Goldstein (1980), Good and Crook (1987), Hartigan (1969), Hill (1980), Kadane and Chuang (1987), Kudō

(1967), Lindley and Smith (1972), Potzelberger (1988), Wolfenson and Fine (1982) etc. We hope that the statistical ideas and the mathematical techniques of this article would provide insight into the more comprehensive problem of decision theoretic robustness. Some of the ideas in this article and in the companion article DasGupta and Studden (1988a) are also being pursued by Jameson Burt and Leon Gleser.

Consider the usual regression setup where $\underset{\sim}{Y}_{n\times 1} \sim N(X\underset{\sim}{\theta}, \sigma^2 I)$, where $\underset{\sim}{\theta}_{p\times 1}$ and σ^2 are unknown parameters and $X_{n\times p}$ is the design matrix of the independent variables. To keep notations simple, assume for the moment that σ^2 is known. The canonical normal problem where $\underset{\sim}{Y} \sim N(\underset{\sim}{\theta}, \sigma^2\Sigma_0)$ where Σ_0 is a known matrix is covered by our setup. Typically, in regression problems, interest lies in prediction and in testing a hypothesis about or estimating the vector of regression coefficients $\underset{\sim}{\theta}$ or a known linear combination $\underset{\sim}{c}'\underset{\sim}{\theta}$ of the regression coefficients. One reason for interest in linear combinations is that the expected value of the response variable corresponding to a particular combination of the independent variables is a linear combination of the regression coefficients. In the robust Bayesian framework, one has a class Γ of prior distributions for the unknown parameter $\underset{\sim}{\theta}$; the typical questions of interest would then be: 'how different can Bayesian measures be as the prior π ranges over Γ? For example, if one wants to estimate the vector of regression coefficients $\underset{\sim}{\theta}$, and assumes a quadratic loss $L(\underset{\sim}{\theta}, \underset{\sim}{a}) = (\underset{\sim}{\theta} - \underset{\sim}{a})'Q(\underset{\sim}{\theta} - \underset{\sim}{a})$ where Q is a fixed positive definite matrix, then the Bayes estimate for any prior π in Γ is the posterior mean

$$\hat{\theta}_\pi = E(\underset{\sim}{\theta}|\underset{\sim}{Y}). \tag{1.1}$$

As π ranges over Γ, the estimates $\hat{\theta}_\pi$ form a set S in the p-dimensional euclidean space $\mathbb{R}^p$ (if Γ is a convex class of priors, then S is automatically a convex set in $\mathbb{R}^p$; even otherwise, S is often convex). Intuitively, one would say that robustness in the Bayes estimate is present if S is a "small" set in $\mathbb{R}^p$; evidently, there are many possible ways of defining the size of a set. Among others, one can consider such intuitive measures as the euclidean diameter of S, namely,

$$D = \sup_{\underset{\sim}{u},\underset{\sim}{v}\in S} ||\underset{\sim}{u} - \underset{\sim}{v}||_2, \tag{1.2}$$

or its Lebesgue measure λ

$$\lambda = \int_S d\underset{\sim}{u}. \tag{1.3}$$

(1.2) seems to be more acceptable because it says how different two Bayes estimates can be. Note that in problems with a single parameter, (1.2) and (1.3) would usually be equivalent measures of variation with respect to the prior. Therefore, for estimating a one-dimensional parametric function $\psi(\underset{\sim}{\theta})$, such as a linear combination $\underset{\sim}{c}'\underset{\sim}{\theta}$, the choice of an index measuring the variation in the estimates is more clear; typically, one considers the range of the estimates namely,

$$R = \sup_{\pi \in \Gamma} E(\psi(\underset{\sim}{\theta})|\underset{\sim}{Y}) - \inf_{\pi \in \Gamma} E(\psi(\underset{\sim}{\theta})|\underset{\sim}{Y}). \tag{1.4}$$

The range can also be used when the primary interest is in finding the posterior probability of a set. Such calculations when the class Γ is the ϵ-contaminated class of Huber (1973) were previously done in Berger and Sivaganesan (1986).

In the context of estimating a parametric function $\psi(\underset{\sim}{\theta})$, another very important quantity of interest is the posterior expected loss of the suggested Bayes action. For ordinary squared error loss, this is just the posterior variance of $\psi(\underset{\sim}{\theta})$. The idea here is that one would like to have an assurance of small losses for all plausible priors, because if the posterior expected losses change in a big range then one doesn't feel confident taking that action. Of special interest is the two-dimensional set of posterior variances versus posterior means; such a set immediately shows the range of the posterior expected loss for each possible value of the Bayes estimate and also helps identifying the priors in Γ for which the extremal values of the posterior expected losses are attained when the posterior mean is fixed. Later in the article, we show examples illustrating this fact.

At this stage, we give a brief exposition of the type of classes Γ considered in the present article. Conceptually, one starts with a subjectively elicited prior π_0, and would like to be robust for all priors π in a 'neighborhood' of π_0; neighborhoods could be specified in any of several possible ways, for example by a metric d on equivalence classes of the class of nonnegative measures on $\mathbb{R}^p$. Thus Γ could be described as

$$\Gamma = \{\pi: \ d(\pi_0, \pi) \le \epsilon\}, \tag{1.5}$$

where ϵ is a (possibly small) specified number. A metric d that has been proposed in the literature (see DeRobertis (1978)) is

$$d(\pi, \pi') = \operatorname*{ess\ sup}_{\theta,\phi} \left[\log \left(\frac{\pi(\theta)}{\pi(\phi)} \middle| \frac{\pi'(\theta)}{\pi'(\phi)} \right) \right], \tag{1.6}$$

where π, π' are two nonnegative densities on the parameter space. Perhaps a more natural appeal of the class Γ induced by this metric is that Γ is the convex cone of nonnegative functions given by

$$\Gamma = \{\alpha\pi\colon\ \alpha > 0,\ \pi_0(\underset{\sim}{\theta}) \le \pi(\underset{\sim}{\theta}) \le k\pi_0(\underset{\sim}{\theta})\}, \tag{1.7}$$

for some fixed number $k > 0$. Γ can be restricted to only infinitely differentiable priors between π_0 and $k\pi_0$ without changing any of the results in this article. The class Γ described above models the prior as lying between two bands; the band is wider and more priors are allowed for larger values of k. Another very attractive feature of this class is that when the likelihood function ℓ is combined with the prior π, the set of resulting posteriors is again a class of the form (1.7). This stability in the set of posteriors over repeated sampling allows one to quickly judge the effect of additional samples on the collection of possible opinions about the parameter $\underset{\sim}{\theta}$ (see L. Brown's discussion on Berger (1984) for an account of this). Another very reassuring stability property enjoyed by this class is that as π ranges over Γ, the set of prior means is often an ellipse and gets translated into a new ellipse with every additional observation. This is one of the classes of priors considered in this article.

Conjugate priors, on the other hand, are very attractive because of their mathematical convenience, and indeed in many problems give a rich enough class of priors for an honest robustness check. Conjugate priors, by definition, have the first stability property described above. A second class of priors considered in this article is

$$\Gamma = \{\pi\colon\ \ \pi \text{ is a } N(\underset{\sim}{\mu}, \sigma^2\Sigma) \text{ density},\ \underset{\sim}{\mu} \in C, \quad \Sigma_1 \le \Sigma \le \Sigma_2\}, \tag{1.8}$$

where Σ_1 and Σ_2 are arbitrary nnd matrices such that $\Sigma_1 \le \Sigma_2$ and C is a suitable convex set in $\mathbb{R}^p$. We will often let C be a singleton set, implying that we feel sure about the location of the prior, but let Σ vary, implying that we do not feel confident about the spread of the prior. At other times, we will let C be a nonempty convex neighborhood of a fixed $\underset{\sim}{\mu}_0$ in $\mathbb{R}^p$, like a circle or an ellipse or a rectangle. In such cases, characterization of the set S of all Bayes estimates of $\underset{\sim}{\theta}$ and evaluation of its diameter D in L_2-norm give rise to interesting geometric problems. A surprising result contained in this article is that if C is an ellipse, the set S often is a trisectrix or a cardioid, and therefore is *not convex*. Classes similar to (1.8) have earlier been studied by Leamer (1978, 1982), and Polachek (1984).

A common feature of both of the classes (1.7) and (1.8) is that they only allow priors with similar tail behavior. For example, if in (1.7) one takes π_0 to be a normal prior, then a t-type tail cannot be accommodated by staying within the class (1.7). Similarly, a t-type prior cannot be accommodated by the class (1.8). Workers in the area, however, emphasize the need for taking priors with different tails, because frequently one feels unsure about the rate at which the prior density tends to zero. Keeping this in mind, we have also considered mixtures of the priors described in (1.8). For example, arbitrary mixtures of all normal priors with a fixed mean $\underset{\sim}{\mu}$ include *all* completely monotone densities with mean $\underset{\sim}{\mu}$; in particular, any elliptically symmetric t-prior with a mean $\underset{\sim}{\mu}$ can be generated by taking such normal mixtures. Not surprisingly, completely arbitrary mixtures of normal priors often lead to an overly conservative big class of priors and we find here that robustness may be unattainable with respect to such a big class of priors.

In this context, optimal design problems arise very naturally. It is quite possible that the diameter D of the set S of Bayes estimates (or other similar measures of variation) will tend to be big for poor choices of the design matrix X, whereas satisfactory robustness obtains for an optimal choice. One should thus make every effort to use the optimal design that gives the best robustness with respect to the prior; clearly, however, designing merely to get the most robust results can potentially lead to a collection of statistical procedures which give mostly similar answers (i.e., are robust), but have other undesirable properties. A more sensible formulation of the optimal design problem would be to impose robustness as a secondary constraint, with the primary goal being near Bayesness with respect to a fixed elicited prior. Such constrained optimal design problems have been addressed in DasGupta and Studden (1988a).

In Section 2, we work out the diameter D of the p-dimensional set S of Bayes estimates and the range of Bayes estimates of an arbitrary linear combination $\underset{\sim}{c}'\underset{\sim}{\theta}$, when Γ is the class (1.8). We also completely characterize the joint set of posterior means and posterior variances for estimating any linear combination $\underset{\sim}{c}'\underset{\sim}{\theta}$. In some interesting special cases, we work out the ranges of the posterior probabilities of sets of general interest. To give the reader an easy grasp of the main results, we start with the case when σ^2 is known. In section 3, we take the mixtures of priors in (1.8) discussed before in order to accommodate priors with thicker tails. We explicitly show the effect of this enlargement in the class of priors on the variations of Bayesian measures. For example, we give a result showing in which cases the joint set of posterior means and variances of $\underset{\sim}{c}'\underset{\sim}{\theta}$ *does not* increase in size even if arbitrary mixtures of the original normal priors are taken. Section 4 contains results for the case when the prior mean $\underset{\sim}{\mu}$ in (1.8) is unknown and varies in a convex set in $\mathbb{R}^p$. The extremal problems here lead to very interesting geometric problems and a highly unanticipated finding is that even if $\underset{\sim}{\mu}$ changes in a convex set, the posterior means may not change in a convex set. In section 5, we have briefly pointed out without proof which of the results for the known σ^2 case generalize to the case when σ^2 is unknown and a suitable prior for σ^2 is used, and which of the results do not or may not. In section 6, we consider the "density-band" class defined in (1.7); here too we have been able to completely characterize the set S and find an expression for its diameter. Ranges of Bayes estimates of $\underset{\sim}{c}'\underset{\sim}{\theta}$ are also worked out. The problem of characterizing the joint set of posterior means and variances gets especially interesting here, because points on the upper and lower boundary of this set are related to the extremal values of the first two moments in the Markov-Krein moment problem. It also turns that a common standardized set of means and variances generates the prior as well as the posterior set in a very simple way, thereby making it necessary to find only this standardized set of means and variances. We also work out the ranges of posterior probabilities of arbitrary sets. Section 7 contains some concluding remarks and discussion.

2. Normal priors, known σ^2.

In this section we work out the variations in different Bayesian measures under the assumptions of a known error variance σ^2 and a class of normal priors, as described in (1.8). The priors considered in this section were first proposed by Leamer (1978, 1982), and Polachek (1984). The very interesting idea of explicitly describing sets of posterior means originated with these articles.

Theorem 2.1. Let $\underset{\sim}{Y}_{n\times 1} \sim N(X\underset{\sim}{\theta}, \sigma^2 I)$, where $\underset{\sim}{\theta} \in \mathbb{R}^p$ is unknown and $\sigma^2 > 0$ is known. Let $\underset{\sim}{\theta}$ have a $N(\underset{\sim}{\mu}, \sigma^2\Sigma)$ distribution where $\underset{\sim}{\mu} \in \mathbb{R}^p$ is fixed and $\Sigma_1 \leq \Sigma \leq \Sigma_2$ in the sense that $\Sigma - \Sigma_1$ and $\Sigma_2 - \Sigma$ are nnd; here Σ_1, Σ_2 are arbitrary nnd matrices with $\Sigma_1 \leq \Sigma_2$. Suppose it is desired to estimate $\underset{\sim}{\theta}$ under the loss

$$L(\underset{\sim}{\theta}, \underset{\sim}{a}) = (\underset{\sim}{\theta} - \underset{\sim}{a})'Q(\underset{\sim}{\theta} - \underset{\sim}{a}),$$

where Q is a known p.d. matrix. Then, the set of all Bayes estimates of $\underset{\sim}{\theta}$ constitute a p-dimensional ellipsoid

$$S = \{\underset{\sim}{\theta}\colon\ (\underset{\sim}{\theta} - (\overline{\Lambda}\underset{\sim}{v} + \underset{\sim}{\mu}))'(\Lambda_2 - \Lambda_1)^{-1}(\underset{\sim}{\theta} - (\overline{\Lambda}\underset{\sim}{v} + \underset{\sim}{\mu})) \leq \frac{\underset{\sim}{v}'(\Lambda_2 - \Lambda_1)\underset{\sim}{v}}{4}\} \tag{2.1}$$

where $\Lambda_i = (X'X + \Sigma_i^{-1})^{-1}$, $\overline{\Lambda} = \frac{\Lambda_1 + \Lambda_2}{2}$, and $\underset{\sim}{v} = X'(\underset{\sim}{Y} - X\underset{\sim}{\mu})$. Moreover, the Euclidean diameter of S, defined as

$$D = \sup_{\underset{\sim}{u}, \underset{\sim}{v} \in S} ||\underset{\sim}{u} - \underset{\sim}{v}||_2 \tag{2.2}$$

is equal to $\sqrt{\underset{\sim}{v}'(\Lambda_2 - \Lambda_1)\underset{\sim}{v} \cdot \lambda_{\max}}$, where $\lambda_{\max}$ is the maximum eigenvalue of $(\Lambda_2 - \Lambda_1)$.

Proof: The ellipsoid S has previously been derived in Leamer (1978, 1982) and Polachek (1984). We sketch the proof here for the sake of completeness and also because our proof is different from the earlier ones given in the literature and makes a very interesting use of the Householder transformation of numerical analysis.

First observe that irrespective of Q, the Bayes estimate under the prior $N(\underset{\sim}{\mu}, \sigma^2\Sigma)$ of $\underset{\sim}{\theta}$ is its posterior mean

$$\hat{\underset{\sim}{\theta}} = E(\underset{\sim}{\theta}|\underset{\sim}{Y} = \underset{\sim}{y}) = (X'X + \Sigma^{-1})^{-1}(X'y + \Sigma^{-1}\underset{\sim}{\mu})$$
$$= \Lambda\underset{\sim}{v} + \underset{\sim}{\mu}, \tag{2.3}$$

where $\Lambda = (X'X + \Sigma^{-1})^{-1}$ and $\underset{\sim}{v}$ is as in the statement of the theorem. Standard results in matrix theory imply that $\Lambda_1 \leq \Lambda \leq \Lambda_2$. We will prove that the set of all vectors $\Lambda\underset{\sim}{v}$ (where $\Lambda_1 \leq \Lambda \leq \Lambda_2$) form an ellipsoid from which it will follow directly that the Bayes estimates form the ellipsoid stated in the theorem.

Define $S_0 = \{\Lambda\underset{\sim}{v},\ \Lambda_1 \leq \Lambda \leq \Lambda_2\}$, and

$$S_1 = \{\underset{\sim}{\theta}\colon\ (\underset{\sim}{\theta} - \overline{\Lambda}\underset{\sim}{v})'(\Lambda_2 - \Lambda_1)^{-1}(\underset{\sim}{\theta} - \overline{\Lambda}\underset{\sim}{v}) \leq \frac{\underset{\sim}{v}(\Lambda_2 - \Lambda_1)\underset{\sim}{v}}{4}\}. \tag{2.4}$$

First note that there exist an orthogonal matrix P such that $P\frac{(\Lambda_2 - \Lambda_1)}{2}P'$ is a diagonal matrix D and an orthogonal matrix Q such that $QD^{\frac{1}{2}}P\underset{\sim}{v} = (\alpha\ 0\ 0 \ldots 0)' = \underset{\sim}{w}$ (say), where α is non-zero if $\underset{\sim}{v}$ is non-zero.

Clearly

$$\begin{aligned}
\Lambda\underset{\sim}{v} &= (\Lambda - \overline{\Lambda})\underset{\sim}{v} + \overline{\Lambda}\underset{\sim}{v} \\
&= P'P(\Lambda - \overline{\Lambda})P'P\underset{\sim}{v} + \overline{\Lambda}\underset{\sim}{v} \\
&= P'D^{\frac{1}{2}}D^{-\frac{1}{2}}P(\Lambda - \overline{\Lambda})P'D^{-\frac{1}{2}}D^{\frac{1}{2}}P\underset{\sim}{v} + \overline{\Lambda}\underset{\sim}{v} \\
&= P'D^{\frac{1}{2}}Q'QD^{-\frac{1}{2}}P(\Lambda - \overline{\Lambda})P'D^{-\frac{1}{2}}Q'QD^{\frac{1}{2}}P\underset{\sim}{v} + \overline{\Lambda}\underset{\sim}{v} \\
&= P'D^{\frac{1}{2}}Q'C\underset{\sim}{w} + \overline{\Lambda}\underset{\sim}{v},
\end{aligned} \tag{2.5}$$

where $C = QD^{-\frac{1}{2}}P(\Lambda - \overline{\Lambda})P'D^{-\frac{1}{2}}Q'$ and $\underset{\sim}{w}$ is as defined before.

Now observe that

$$\begin{aligned}
\Lambda_1 \leq \Lambda \leq \Lambda_2 &\Leftrightarrow -\frac{\Lambda_2 - \Lambda_1}{2} \leq \Lambda - \overline{\Lambda} \leq \frac{\Lambda_2 - \Lambda_1}{2} \\
&\Leftrightarrow -D \leq P(\Lambda - \overline{\Lambda})P' \leq D \\
&\Leftrightarrow -I \leq D^{-\frac{1}{2}}P(\Lambda - \overline{\Lambda})P'D^{-\frac{1}{2}} \leq I \\
&\Leftrightarrow -I \leq QD^{-\frac{1}{2}}P(\Lambda - \overline{\Lambda})P'D^{-\frac{1}{2}}Q' \leq I \\
&\Leftrightarrow -I \leq C \leq I.
\end{aligned} \tag{2.6}$$

Hence, if we can show that the set

$$S_2 = \{C\underset{\sim}{w}\colon\ \ -I \le C \le I\}$$

is a sphere, then it will follow from (2.5) that S_0 is an ellipsoid. Without loss of any generality, we assume that $\alpha = 1$ which makes $\underset{\sim}{w}$ the first unit vector $(1\ 0\ 0 \ldots 0)'$. Since S_2 is convex, it will suffice to show that any vector $\underset{\sim}{e}$ of euclidean norm 1 can be written in the form

$$\underset{\sim}{e} = C\underset{\sim}{w},$$

for some $-I \le C \le I$.

This can be done by using the well known Householder transformation

$$C = I - 2\underset{\sim}{a}\underset{\sim}{a}',$$

where $a_1 = \sqrt{\frac{1-e_1}{2}}$ and $a_j = \frac{-e_j}{2a_1}$, where $e_1, \ldots, e_p$ are the coordinates of $\underset{\sim}{e}$; since $\underset{\sim}{a}'\underset{\sim}{a} = 1$, it follows that $-I \le C \le I$. Note e_1 can be assumed to be smaller than 1, for if $e_1 = 1$ then C can be chosen to be I. This proves that the set of Bayes estimates S is the ellipsoid stated in the theorem.

That the euclidean diameter of S is equal to

$$\sqrt{\underset{\sim}{v}'(\Lambda_2 - \Lambda_1)\underset{\sim}{v}\lambda_{\max}}$$

follows from the above ellipsoid representation of S.

Corollary 2.2. (a) For any $L_{k\times p}$ where rank $(L) = k \le p$, the set of Bayes estimates of $L\underset{\sim}{\theta}$ form the ellipsoid

$$\begin{aligned} S_L^* = \{\underset{\sim}{u}\colon\ \ &(\underset{\sim}{u} - L(\overline{\Lambda}\underset{\sim}{v} + \underset{\sim}{\mu}))'(L(\Lambda_2 - \Lambda_1)L')^{-1}(\underset{\sim}{u} - L(\overline{\Lambda}\underset{\sim}{v} + \underset{\sim}{\mu})) \\ &\le \frac{\underset{\sim}{v}'(\Lambda_2 - \Lambda_1)\underset{\sim}{v}}{4}\}; \end{aligned} \tag{2.7}$$

(b) For any p-dimensional vector $\underset{\sim}{c}$, the Bayes estimates of $\underset{\sim}{c}'\underset{\sim}{\theta}$ form the interval

$$\begin{aligned} &\underset{\sim}{c}'(\Lambda\underset{\sim}{v} + \underset{\sim}{\mu}) - \frac{1}{2}\sqrt{v'(\Lambda_2 - \Lambda_1)\underset{\sim}{v} \cdot \underset{\sim}{c}'(\Lambda_2 - \Lambda_1)\underset{\sim}{c}} \\ &\le u \\ &\le \underset{\sim}{c}'(\Lambda\underset{\sim}{v} + \underset{\sim}{\mu}) + \frac{1}{2}\sqrt{\underset{\sim}{v}'(\Lambda_2 - \Lambda_1)\underset{\sim}{v} \cdot \underset{\sim}{c}'(\Lambda_2 - \Lambda_1)\underset{\sim}{c}}. \end{aligned} \tag{2.8}$$

Proof: (a) Follows from the facts that $E(L\underset{\sim}{\theta}|\underset{\sim}{Y} = \underset{\sim}{y}) = LE(\underset{\sim}{\theta}|\underset{\sim}{Y} = \underset{\sim}{y})$, that the Bayes estimates of $\underset{\sim}{\theta}$ form the ellipsoid (2.1) and that the image of an ellipsoid under a linear transformation is also an ellipsoid (see Johnson and Wichern (1982), pp. 219–220). (b) Follows from (a) on taking $L = \underset{\sim}{c}'$; if $\underset{\sim}{c} = \underset{\sim}{0}$, the result is trivial.

In regression problems, probably of more interest is the problem of predicting k future values of the response variable $\underset{\sim}{Y}$ corresponding to the vectors of predictor variables $\underset{\sim}{x}_{01}, \underset{\sim}{x}_{02}, \ldots, \underset{\sim}{x}_{0k}$. Since the problem of predicting $(Y_1, \ldots, Y_k)$ under squared error loss is equivalent to estimating $(EY_1, \ldots, EY_k)$ under squared error loss, and since

$$(EY_1 \ldots EY_k)' = L\underset{\sim}{\theta},$$

where

$$L_{k\times p} = (\underset{\sim}{x}_{01} \ldots \underset{\sim}{x}_{0k})',$$

we have the result that the *set of Bayes predictors* of $(Y_1, \ldots, Y_k)$ under squared error loss, under the prior modelling of Theorem 2.1, is the ellipsoid (2.7).

Example 1. Consider a simple linear regression model

$$y_i = \theta_0 + \theta_1 x_i + \epsilon_i, \quad 1 \le i \le n,$$

where the independent variable x assumes values in the interval $[-1, 1]$ and ϵ_i are iid $N(0, 1)$. Suppose the joint distribution of $\underset{\sim}{\theta} = (\theta_0 \ \theta_1)'$ is $N(\underset{\sim}{0}, \Sigma)$, where $I \le \Sigma \le 5I$; this implies that the variances of θ_0 and θ_1 are between 1 and 5 and the correlation is between $\pm\frac{2}{3}$. Thus a substantial amount of variation is allowed in the variances and the correlation. Suppose now $n = 20$ observations are taken and 10 of these are taken at $X = 1$ and another 10 at $X = -1$ (this design is used just as an artifact, but note that this design has many standard optimum properties). Let $\hat{\underset{\sim}{\theta}}_L = (X'X)^{-1}X'\underset{\sim}{y}$ denote the usual least squares estimate of $\underset{\sim}{\theta}$. Typically, for $\hat{\underset{\sim}{\theta}}_L$ near zero good (Bayesian) robustness is anticipated. Let us take $\hat{\underset{\sim}{\theta}}_L = (3.15 \ \ 3.15)'$; this choice roughly corresponds to the value of $\hat{\underset{\sim}{\theta}}'_L\hat{\underset{\sim}{\theta}}_L$ equal to its marginal mean plus one marginal standard deviation under the marginal distribution of $\underset{\sim}{Y}$ induced by the *worst* prior covariance matrix (which will be $5I$ under our modelling). Our choice of $\hat{\underset{\sim}{\theta}}_L$ thus is not a value chosen conveniently near zero so good robustness will automatically be obtained. Now, using Theorem (2.1), one has that the set of Bayes estimates of $\underset{\sim}{\theta}$ is the circle

$$(\theta_0 - 3.06)^2 + (\theta_1 - 3.06)^2 \leq .0072,$$

with center at (3.06, 3.06) and radius .085. Thus, in particular, the estimates of each regression coefficient vary in the range 3.06 $\pm$ .085. Note that on the other hand the standard error of the least squares estimates is .2236, more than 2.6 times the half-width of the interval of Bayes estimates. This indicates that encouraging Bayesian robustness can be obtained in very reasonable situations with just a moderate sample size. The circle is plotted in Figure 1.

Example 2. Ordinarily, it will be desirable that the diameter of the set of Bayes estimates converges to zero (for "all" values of $\underset{\sim}{y}$) as $n \to \infty$. We work out a simple example below where this is the case (the same result is true in much more generality).

Let again $\Sigma_1 = I$ and $\Sigma_2 = kI$ and consider the simple linear regression model with the same design (i.e. 50% of the observations taken at each of $X = 1$ and $X = -1$); thus $X'X = nI$. Then straightforward computation gives

$$D^2 = \frac{(1-\frac{1}{k})^2 \underset{\sim}{y}' X X' \underset{\sim}{y}}{(n+1)^2 (n+\frac{1}{k})^2}. \tag{2.9}$$

Under the prior $N(\underset{\sim}{0}, kI)$, $\underset{\sim}{u}_n = X'\underset{\sim}{y}$ is marginally distributed as $N(\underset{\sim}{0}, (n + n^2 k)I)$. Thus

$$\begin{aligned} D^2 &= \frac{(1-\frac{1}{k})^2 \underset{\sim}{u}_n' \underset{\sim}{u}_n}{(n+1)^2 (n+\frac{1}{k})^2} \\ &= \frac{(1-\frac{1}{k})^2 nk W_n}{(n+1)^2 (n+\frac{1}{k})}, \end{aligned} \tag{2.10}$$

where $W_n \sim \chi^2(p)$.

$$\therefore D \leq \sqrt{k}(1-\frac{1}{k})\frac{\sqrt{W_n}}{n}. \tag{2.11}$$

Now for any fixed $\epsilon > 0$,

$$\infty > E(\frac{\sqrt{W_1})}{\epsilon}$$

$$\geq \sum_{n=1}^{\infty} P(\frac{\sqrt{W_1}}{\epsilon} \geq n)$$
$$= \sum_{n=1}^{\infty} P(\frac{\sqrt{W_n}}{\epsilon} \geq n)$$
$$= \sum_{n=1}^{\infty} P(\frac{\sqrt{W_n}}{n} \geq \epsilon), \tag{2.12}$$

which implies by the Borel-Cantelli lemma that $P(\limsup \frac{\sqrt{W_n}}{n} \geq \epsilon) = 0$, and therefore on a set of probability 1, $\frac{\sqrt{W_n}}{n} < \epsilon$ after a finite stage (depending on ϵ), implying that $\frac{\sqrt{W_n}}{n}$ converges to zero almost surely under the marginal distribution of $\underset{\sim}{Y}$ induced by the prior $N(\underset{\sim}{0}, kI)$. It follows from (2.11) that the diameter of the set of Bayes estimates goes to zero. As mentioned before the result is true in more generality.

As mentioned in section 1, of special interest in regression problems are parametric functions $\underset{\sim}{c}'\underset{\sim}{\theta}$ where $\underset{\sim}{c}$ is an arbitrary non random vector. The range of the posterior means of $\underset{\sim}{c}'\underset{\sim}{\theta}$ was given in Corollary 2.2. We now work out, for a given linear combination $\underset{\sim}{c}'\underset{\sim}{\theta}$, the joint set of posterior mean and posterior variance as the prior ranges in the class (1.8). A complete description of this set would enable the reader to immediately figure out the range of posterior variances for any fixed value of the posterior mean and also to quickly find answers to questions such as for which value of the mean the posterior variance (or the range of the posterior variance) is maximized. The following theorem describes this two dimensional set.

Theorem 2.3. Let $\underset{\sim}{c}_{p\times 1}$ be an arbitrary but fixed vector. Under the setup of Theorem 2.1, the set

$$S(c) = \{(E(\underset{\sim}{c}'\underset{\sim}{\theta}|\underset{\sim}{Y} = \underset{\sim}{y}),\ \text{Var}\ (\underset{\sim}{c}'\underset{\sim}{\theta}|\underset{\sim}{Y} = \underset{\sim}{y}))\}$$

is a two-dimensional ellipse

$$\{\underset{\sim}{u}:\ (\underset{\sim}{u} - \underset{\sim}{u}_0)'D^{-1}(\underset{\sim}{u} - \underset{\sim}{u}_0) \leq A^2\} \tag{2.13}$$

where

$$\underset{\sim}{u}_0 = \begin{pmatrix} \underset{\sim}{v}'\overline{\Lambda}\underset{\sim}{c} + \underset{\sim}{c}'\underset{\sim}{\mu} \\ \sigma^2\underset{\sim}{c}'\overline{\Lambda}\underset{\sim}{c} \end{pmatrix}, \tag{2.14}$$

$$D = \begin{pmatrix} \underset{\sim}{v}'(\Lambda_2 - \Lambda_1)\underset{\sim}{v} & \sigma^2 \underset{\sim}{c}'(\Lambda_2 - \Lambda_1)\underset{\sim}{v} \\ & \sigma^4 \underset{\sim}{c}'(\Lambda_2 - \Lambda_1)\underset{\sim}{c} \end{pmatrix} \tag{2.15}$$

and

$$A^2 = \frac{\underset{\sim}{c}'(\Lambda_2 - \Lambda_1)\underset{\sim}{c}}{4}. \tag{2.16}$$

Note D^{-1} exists unless $\underset{\sim}{v}$ and $\underset{\sim}{c}$ are linearly dependent.

Proof: Since the posterior distribution of $\underset{\sim}{\theta}$ under the prior $N(\underset{\sim}{\mu}, \sigma^2\Sigma)$ is $N(\Lambda\underset{\sim}{v} + \underset{\sim}{\mu}, \sigma^2\Lambda)$, clearly,

$$\begin{pmatrix} E(\underset{\sim}{c}'\underset{\sim}{\theta}|\underset{\sim}{Y} = \underset{\sim}{y}) \\ \text{Var}\,(\underset{\sim}{c}'\underset{\sim}{\theta}|\underset{\sim}{Y} = \underset{\sim}{y}) \end{pmatrix} = \begin{pmatrix} \underset{\sim}{c}'(\Lambda\underset{\sim}{v} + \underset{\sim}{\mu}) \\ \sigma^2\underset{\sim}{c}'\Lambda\underset{\sim}{c} \end{pmatrix} = \begin{pmatrix} \underset{\sim}{v}' \\ \sigma^2\underset{\sim}{c}' \end{pmatrix}\Lambda\underset{\sim}{c} + \begin{pmatrix} \underset{\sim}{c}'\underset{\sim}{\mu} \\ 0 \end{pmatrix}. \tag{2.17}$$

It was effectively proved in Theorem 2.1 that for any vector $\underset{\sim}{c}$, the set of points $\{\Lambda\underset{\sim}{c}, \Lambda_1 \leq \Lambda \leq \Lambda_2\}$ form the ellipsoid

$$\{\underset{\sim}{\theta}\!: \; (\underset{\sim}{\theta} - \overline{\Lambda}\underset{\sim}{c})'(\Lambda_2 - \Lambda_1)^{-1}(\underset{\sim}{\theta} - \overline{\Lambda}\underset{\sim}{c}) \leq \frac{\underset{\sim}{c}'(\Lambda_2 - \Lambda_1)\underset{\sim}{c}}{4}\} \tag{2.18}$$

Consequently, on letting $L_{2\times p} = \binom{\underset{\sim}{v}'}{\sigma^2\underset{\sim}{c}'}$, it follows that the set of points $\{L\Lambda\underset{\sim}{c}, \; \Lambda_1 \leq \Lambda \leq \Lambda_2\}$ form the two-dimensional ellipse

$$\{\underset{\sim}{\theta}\!: (\underset{\sim}{\theta} - L\overline{\Lambda}\underset{\sim}{c})'(L(\Lambda_2 - \Lambda_1)L')^{-1}(\underset{\sim}{\theta} - L\overline{\Lambda}\underset{\sim}{c}) \leq \frac{\underset{\sim}{c}'(\Lambda_2 - \Lambda_1)\underset{\sim}{c}}{4}\}, \tag{2.19}$$

provided $\underset{\sim}{v}$ and $\underset{\sim}{c}$ are linearly independent (which will be true with a marginal probability of 1 for any prior in the class (1.8)). The result now follows from (2.17) and (2.19) on noting that $\underset{\sim}{u}_0 = L\overline{\Lambda}\underset{\sim}{c} + \binom{\underset{\sim}{c}'\underset{\sim}{\mu}}{0}$ and $D = L(\Lambda_2 - \Lambda_1)L'$.

Example 3. In the setup of example 1, with $\underset{\sim}{c} = (0\ 1)'$ (i.e., for estimating the slope of the regression line), the ellipse $S(c)$ is

$$0.1326(u_1 - 3.06)^2 + 1052.6316(u_2 - .0485)^2 - 16.7084(u_1 - 3.06)(u_2 - .0485) \leq .0005 \tag{2.20}$$

Below we give the minimum and the maximum posterior variance for a few selected values of the posterior mean. The whole ellipse is plotted in Figure 2.

Mean	Variance	
	min.	max.
2.98	.0476	.0481
3	.0476	.0485
3.02	.0476	.0488
3.04	.0477	.0490
3.06	.0478	.0492
3.08	.0480	.0493
3.10	.0482	.0494
3.12	.0485	.0494
3.14	.0489	.0493

The range of the posterior variance is maximized for a posterior mean of 3.06 (the midpoint of the range of means), this maximum range being approximately $.0492 - .0478 = .0014$. Another very important reason for the interest in the two dimensional set $S(c)$ is that a complete characterization of the joint set of posterior means and posterior variances enables us to find the ranges of other quantities of posterior interest, for example, the posterior quantiles of $\underset{\sim}{c}'\underset{\sim}{\theta}$. The posterior quantiles, apart from giving one an idea of the spread of the posterior distribution, are important also because they are in fact the Bayes estimates of $\underset{\sim}{c}'\underset{\sim}{\theta}$ with respect to the family of losses

$$\begin{aligned} L(\theta, a) &= k_0(a-\theta) \text{ if } \theta \le a \\ &= k_1(\theta - a) \text{ if } \theta > a; \end{aligned} \tag{2.21}$$

(the Bayes estimate under this loss is the posterior $\frac{k_1}{k_0+k_1}$th quantile of $\underset{\sim}{c}'\underset{\sim}{\theta}$). In our problem, under the family of priors (1.8), the posterior distribution of $\underset{\sim}{c}'\underset{\sim}{\theta}$ is normal, and hence the αth posterior quantile of $\underset{\sim}{c}'\underset{\sim}{\theta}$ as the prior ranges over the class (1.8) forms the set of numbers

$$\begin{aligned} Q_\alpha &= \{E(\underset{\sim}{c}'\underset{\sim}{\theta}|\underset{\sim}{Y} = \underset{\sim}{y}) + z_\alpha \cdot \sqrt{\text{Var } \underset{\sim}{c}'\underset{\sim}{\theta}|\underset{\sim}{Y} = \underset{\sim}{y}}\} \\ &= \{u_1 + z_\alpha\sqrt{u_2}\colon \ \underset{\sim}{u} = (u_1 \ u_2)' \in S(c)\}, \end{aligned} \tag{2.22}$$

where z_α is the αth quantile of a $N(0,1)$ distribution. Since a characterization of the set $S(c)$ is available, finding the infimum and the supremum of the set of real numbers Q_α is an easy one variable optimization problem. The following theorem makes this formal.

Theorem 2.4. Let $\underset{\sim}{u} = (u_1\ u_2)'$ belong to the ellipse $(\underset{\sim}{u} - \underset{\sim}{u}_0)'D^{-1}(\underset{\sim}{u} - \underset{\sim}{u}_0) \leq A^2$; suppose $\underset{\sim}{u}_0 = (\underset{\sim}{u}_{01}\ \underset{\sim}{u}_{02})'$, and

$$D = \begin{pmatrix} \sigma_1^2 & \rho_{\sigma_1\sigma_2} \\ \rho_{\sigma_1\sigma_2} & \sigma_2^2 \end{pmatrix},$$

where $\rho^2 \leq 1$. Then, for $z > 0 (z < 0)$,

$$\max_{(\underset{\sim}{u}-\underset{\sim}{u}_0)'D^{-1}(\underset{\sim}{u}-\underset{\sim}{u}_0) \leq A^2} u_1 + z\sqrt{u_2} \tag{2.23}$$

$$= \max_{|u_1 - u_{01}| \leq A\sigma_1} \left[u_1 + z\sqrt{\{u_{02} + \rho\frac{\sigma_2}{\sigma_1}(u_1 - u_{01}) \underset{(-)}{+} \sigma_2\sqrt{(1-\rho^2)(A^2 - \frac{(u_1 - u_{01})^2}{\sigma_1^2})}\}} \right]$$

and

$$\min_{(\underset{\sim}{u}-\underset{\sim}{u}_0)'D^{-1}(\underset{\sim}{u}-\underset{\sim}{u}_0) \leq A^2} u_1 + z\sqrt{u_2} \tag{2.24}$$

$$= \min_{|u_1 - u_{01}| \leq A\sigma_1} \left[u_1 + z\sqrt{\{u_{02} + \rho\frac{\sigma_2}{\sigma_1}(u_1 - u_{01}) \underset{(+)}{-} \sigma_2\sqrt{(1-\rho^2)(A^2 - \frac{(u_1 - u_{01})^2}{\sigma_1^2})}\}} \right]$$

Proof: First note that obviously the minimum and the maximum of $u_1 + z\sqrt{u_2}$ are attained by points on the boundary of the ellipse. The proof of (2.23) follows on noting that if $(u_1, u_2)'$ is on the boundary of the ellipse $(\underset{\sim}{u} - \underset{\sim}{u}_0)'D^{-1}(\underset{\sim}{u} - \underset{\sim}{u}_0) \leq A^2$, then

$$u_2 = u_{02} + \rho\frac{\sigma_2}{\sigma_1}(u_1 - u_{01}) \pm \sigma_2\sqrt{(1-\rho^2)(A^2 - \frac{(u_1 - u_{01})^2}{\sigma_1^2})},$$

and that u_1 varies in the interval $|u_1 - u_{01}| \leq A\sigma_1$. The proof of (2.24) is similar.

Theorem 2.4 thus shows that finding the maximum and the minimum of any posterior quantile is a single variable optimization problem for any arbitrary linear combination $\underset{\sim}{c}'\underset{\sim}{\theta}$. In general, closed form expressions for the maximum in (2.23) (and the minimum in (2.24)) are complicated, but in any particular case finding the numerical values is an easy computing exercise.

Example 4. Again consider the situation of example 2; also let $\underset{\sim}{c} = (0\ 1)'$. The ranges of several posterior quantiles for this case are listed below. The maximums and the minimums were calculated on a Casio fx–7000G graphics calculator.

		Posterior quantile	
α	z_α	min	max
.25	$-.67$	2.8301	2.9948
1/3	$-.44$	2.8805	3.0457
2/3	.44	3.0730	3.2408
.75	.67	3.1232	3.2918

Interestingly, for each α, the length of the interval of posterior quantiles is approximately .17, which is also the approximate length of the interval of posterior means (and hence posterior median); this suggests that roughly the same amount of robustness may be achieved for quite different loss functions, like the quadratic loss of Theorem 2.1 and the piecewise linear loss in (2.21). All of these observations in the robustness study are made convenient and possible by the explicit elliptical representation of the set $S(c)$.

Finally, we now show a result on the ranges of posterior probabilities of spheres centered at the prior mean $\underset{\sim}{\mu}$. Such spheres $S = \{\underset{\sim}{\theta}\colon \ ||\underset{\sim}{\theta} - \underset{\sim}{\mu}|| \leq k\}$ are of intrinsic interest because practitioners are very often interested in the probability (or the likelihood) that the unknown parameter $\underset{\sim}{\theta}$ lies in a "small" neighborhood of the apriori guess $\underset{\sim}{\mu}$. To the best of our knowledge, this is the first time that the classical Anderson-type theorems on probability contents of symmetric sets have been used in a Bayesian context. For the techniques to go through, we need to assume that the design matrix X and the prior variance covariance matrices Σ are such that for any Σ and Σ^* in the class of prior covariance matrices, $(X'X + \Sigma^{-1})$ and $(X'X + \Sigma^{*-1})$ commute. The condition, though apparently restrictive, is satisfied in some common cases. Examples are given after the following theorem.

Theorem 2.5. Let

$$\underset{\sim}{Y} \sim N(X\underset{\sim}{\theta}, \sigma^2 I)$$

$$\underset{\sim}{\theta} \sim N(\underset{\sim}{\mu}, \sigma^2 \Sigma),$$

where $\underset{\sim}{\mu}, \sigma^2$ are fixed, and Σ belongs to a class of matrices Γ. Suppose for all Σ, Σ^* belonging to Γ,

$$(X'X + \Sigma^{-1})(X'X + \Sigma^{*-1}) = (X'X + \Sigma^{*-1})(X'X + \Sigma^{-1}). \tag{2.25}$$

Let $S = \{\underset{\sim}{\theta}: \ ||\underset{\sim}{\theta} - \underset{\sim}{\mu}|| \leq k, \ k > 0\}$. Then, $\Sigma \leq \Sigma^*$ implies

$$P_\Sigma(\underset{\sim}{\theta} \in S|\underset{\sim}{Y} = \underset{\sim}{y}) \geq P_{\Sigma^*}(\underset{\sim}{\theta} \in S|\underset{\sim}{Y} = \underset{\sim}{y}) \tag{2.26}$$

Proof: Let $\Lambda = (X'X + \Sigma^{-1})^{-1}$, and $\Lambda^* = (X'X + \Sigma^{*-1})^{-1}$. Under the hypothesis (2.26), there exists an orthogonal matrix P such that $\Lambda = P'DP$ and $\Lambda^* = P'D^*P$, where D and D^* are diagonal matrices. Let $D = \text{diag}(d_1, \ldots, d_p)$ and $D^* = \text{diag}(d_1^*, \ldots, d_p^*)$. Clearly, $0 < d_i \leq d_i^*$ for every i, $1 \leq i \leq p$. Now,

$$\begin{aligned}
&P_\Sigma(\underset{\sim}{\theta} \in S|\underset{\sim}{Y} = \underset{\sim}{y}) \\
&\quad = P(||\underset{\sim}{\theta}|| \leq k|\underset{\sim}{\theta} \sim N(\Lambda\underset{\sim}{v}, \Lambda)) \\
&\quad = P(||\underset{\sim}{W}|| \leq k|\underset{\sim}{W} \sim N(D\underset{\sim}{u}, D)) \quad (\text{where } \underset{\sim}{u} = P\underset{\sim}{v}) \\
&\quad = P(\underset{\sim}{Z}'D\underset{\sim}{Z} \leq k^2|\underset{\sim}{Z} \sim N(D^{\frac{1}{2}}\underset{\sim}{u}, I)).
\end{aligned} \tag{2.27}$$

Now, since $Z_1, \ldots, Z_p$ are independently distributed, the $N(\underset{\sim}{\delta}, I)$ distributions form a location parameter family, the standard normal density is log concave, $|\sqrt{d_i}u_i| \leq |\sqrt{d_i^*}u_i|$ for every i, and the set $\{\underset{\sim}{Z}: \ \underset{\sim}{Z}'D\underset{\sim}{Z} \leq k^2\}$ is sign invariant and convex, it follows from Theorem 4.1.5 in Tong (1980) that

$$P(\underset{\sim}{Z}'D\underset{\sim}{Z} \leq k^2|\underset{\sim}{Z} \sim N(D^{\frac{1}{2}}\underset{\sim}{u}, I)) \geq P(\underset{\sim}{Z}'D\underset{\sim}{Z} \leq k^2|\underset{\sim}{Z} \sim N(D^{*\frac{1}{2}}\underset{\sim}{u}, I). \tag{2.28}$$

Since $\{\underset{\sim}{Z}'D\underset{\sim}{Z} \leq k^2\} \supseteq \{\underset{\sim}{Z}'D^*\underset{\sim}{Z} \leq k^2\}$ and

$$P(\underset{\sim}{Z}'D^*\underset{\sim}{Z} \leq k^2|\underset{\sim}{Z} \sim N(D^{*\frac{1}{2}}\underset{\sim}{u}, I)) = P_{\Sigma^*}(\underset{\sim}{\theta} \in S|\underset{\sim}{Y} = \underset{\sim}{y}), \tag{2.29}$$

(by an argument parallel to (2.27)), the proof is now complete.

Example 5. The condition (2.25) is always satisfied if $X'X$ and the prior covariance matrices are diagonal. Thus, in the usual normal problem where $\underset{\sim}{Y} \sim N(\underset{\sim}{\theta}, \sigma^2\Sigma_0)$ and $\underset{\sim}{\theta} \sim N(\underset{\sim}{\mu}, \sigma^2\Sigma)$, and Σ_0, Σ are diagonal, our Theorem 2.5 will always apply. Suppose we are in the situation of Example 1 with the prior covariance matrix $\Sigma = \begin{pmatrix} \sigma_{11} & 0 \\ 0 & \sigma_{22} \end{pmatrix}$

being such that $1 \le \sigma_{11},\ \sigma_{22} \le 5$. Let $\underset{\sim}{\mu} = \underset{\sim}{0}$ and let S be the sphere $\{\underset{\sim}{\theta}\colon \ \ ||\underset{\sim}{\theta}|| \le 4\}$. By Theorem 2.5,

$$\inf_{\Sigma} P(\underset{\sim}{\theta} \in S | \underset{\sim}{Y} = \underset{\sim}{y}) = P(||\underset{\sim}{\theta}|| \le 4 | \underset{\sim}{\theta} \sim N\left(\begin{pmatrix} 3.12 \\ 3.12 \end{pmatrix}, \frac{1}{20.2} I\right)$$
$$= .0301.$$

Similarly,

$$\sup_{\Sigma} P(\underset{\sim}{\theta} \in S | \underset{\sim}{Y} = \underset{\sim}{y})$$
$$= .1272.$$

In general, one needs to evaluate two noncentral chi square probabilities, which can be done either using the PROBCHI option in SAS (1985), or by using the tables in Fix (1949); for satisfactory approximations, also see Abramowitz and Stegun (1964).

Example 6. If the design matrix X and the prior covariance matrices Σ are such that $X'X + \Sigma^{-1}$ are of the form $\alpha I + \beta \underset{\sim}{1}\underset{\sim}{1}'$ (commonly known as the "equi-correlation" structure), then hypothesis (2.25) is again satisfied. In particular, (2.25) will hold if $X'X$ and Σ are each of the form $\alpha I + \beta \underset{\sim}{1}\underset{\sim}{1}'$. For example, if one lets the prior covariance matrix Σ to vary in the set

$$\Gamma = \{\Sigma\colon \ \ I \le \Sigma = \begin{pmatrix} \sigma^2 & \sigma^2\rho \\ \sigma^2\rho & \sigma^2 \end{pmatrix} \le 5I\},$$

then with $\underset{\sim}{\hat{\theta}}_L$, $X'X$, and S as in the previous example, by Theorem 2.5 the infimum and the supremum of the posterior probability of S are again attained at $\Sigma = 5I$ and $\Sigma = I$ respectively; consequently, they are .0301 and .1272 as before. Note that the set Γ described above is the subset of all matrices of the form $\alpha I + \beta \underset{\sim}{1}\underset{\sim}{1}'$ of the previously considered set $I \le \Sigma \le 5I$. As mentioned before, if Σ belongs to Γ, then $1 \le \sigma^2 \le 5$ and $|\rho| \le \frac{2}{3}$.

A hypothesis of general interest in regression problems is that a fixed linear combination $\underset{\sim}{c}'\underset{\sim}{\theta}$ of the regression coefficients is in a neighborhood of its prior mean $\underset{\sim}{c}'\underset{\sim}{\mu}$. Using an argument similar to that of Theorem 2.5, it is easy to prove the following result. We omit the proof.

Theorem 2.6. Suppose hypothesis (2.25) holds; let $\underset{\sim}{c}$ be any fixed p-dimensional vector;

then $\Sigma \leq \Sigma^*$ implies

$$P_\Sigma(|\underset{\sim}{c}'(\underset{\sim}{\theta} - \underset{\sim}{\mu})| \leq k|\underset{\sim}{Y} = \underset{\sim}{y}) \geq P_{\Sigma^*}(|\underset{\sim}{c}(\underset{\sim}{\theta} - \underset{\sim}{\mu})| \leq k|\underset{\sim}{Y} = \underset{\sim}{y}) \tag{2.30}$$

Example 7. Suppose in the setup of Example 5, we want to know if $E(Y|X = -.9)$ is between $\pm.5$; this will then correspond to finding $P(-.5 \leq c'\theta \leq 5)$ where $\underset{\sim}{c} = (1, -.9)$. By (2.30),

$$\begin{aligned}\inf_\Sigma P(-.5 \leq \underset{\sim}{c}'\underset{\sim}{\theta} \leq .5|\underset{\sim}{Y} = \underset{\sim}{y}) &= P(-.5 \leq X \leq .5|X \sim N(.312, .0896)) \\ &= .7323.\end{aligned}$$

Similarly,

$$\begin{aligned}\sup_\Sigma P(-.5 \leq \underset{\sim}{c}'\underset{\sim}{\theta} \leq .5|\underset{\sim}{Y} = \underset{\sim}{y}) &= P(-.5 \leq X \leq .5|X \sim N(.3, .0862)) \\ &= .7484.\end{aligned}$$

Using the class of priors of Example 5, one will perhaps accept the hypothesis that $|E(Y|X = .9)| \leq 0.5$ since the posterior probability of the hypothesis is large for all priors under consideration and the results are very robust (the probability changes between .7323 and .7484).

3. Mixture normal priors.

A natural way to enlarge the class of priors considered in section 2 is to consider their mixtures. If we want to use priors of the general normal shape with a mode at some point $\underset{\sim}{\mu}$ but are unwilling to use regular normal priors alone, a natural class to consider is the class of mixtures of normal priors. Taking the mixtures of the priors $N(\underset{\sim}{\mu}, \Sigma)$, $\Sigma_1 \leq \Sigma \leq \Sigma_2$, has different effect on the class of Bayes estimates of $\underset{\sim}{\theta}$ for different losses. We consider the quadratic loss $(\underset{\sim}{\theta} - \underset{\sim}{a})'Q(\underset{\sim}{\theta} - \underset{\sim}{a})$ and the piecewise linear loss of (2.21). Interestingly, for quadratic losses, enlarging the class of priors by taking mixtures does not change the class of Bayes estimates of $\underset{\sim}{\theta}$. This is the assertion of the following theorem.

Theorem 3.1. In an arbitrary decision problem, suppose that $\underset{\sim}{\theta}$ is to be estimated under the loss $L(\underset{\sim}{\theta}, \underset{\sim}{a}) = (\underset{\sim}{\theta} - \underset{\sim}{a})'Q(\underset{\sim}{\theta} - \underset{\sim}{a})$, where Q is a fixed p.d. matrix. Let

$$\Gamma = \{\pi_\alpha(\underset{\sim}{\theta}):\ \ \alpha \in I\} \tag{3.1}$$

be any class of priors. Let S be the set of Bayes estimates of $\underset{\sim}{\theta}$ as the prior varies in Γ. Let

$$\Gamma^* = \{\pi(\underset{\sim}{\theta})\colon \ \pi(\underset{\sim}{\theta}) = \int_I \pi_\alpha(\underset{\sim}{\theta}) d\tau(\alpha), \text{ where } \int_I d\tau(\alpha) = 1\}. \tag{3.2}$$

Let S^* be the set of all Bayes estimates of $\underset{\sim}{\theta}$ as the prior varies in Γ^*. If S is convex, then $S^* = S$.

Proof: Follows from the fact that Bayes estimates under the loss of this theorem are posterior means and hence S^* is the convex hull of S.

Theorem 2.1 and Theorem 3.1 together imply that the class of Bayes estimates of $\underset{\sim}{\theta}$ under arbitrary mixtures of the $N(\underset{\sim}{\mu}, \Sigma)$ priors (where $\Sigma_1 \leq \Sigma \leq \Sigma_2$) is the same ellipsoid (2.1) and hence (2.2), (2.7), and (2.8) are all valid. Although mixing the priors has *no* effect on the set of Bayes estimates for quadratic loss, it usually enlarges the set of Bayes estimates for the piecewise linear loss (2.21). This is because under the loss (2.21), the Bayes estimate of $\underset{\sim}{\theta}$ is the vector of $\frac{k_1}{k_0+k_1}$-th posterior quantiles of θ_i, and the p-dimensional set of posterior quantiles may not be invariant under mixing of the priors. We do not have at the moment an explicit representation of the set of Bayes estimates of $\underset{\sim}{\theta}$ under mixture normal priors for the loss (2.21). However, *for any single linear combination* $\underset{\sim}{c}'\underset{\sim}{\theta}$, the range of its Bayes estimates under the loss (2.21) remains unaffected by a mixing of the priors. This invariance result is proved in the next theorem.

Theorem 3.2. Let θ be a scalar parameter in an arbitrary statistical decision problem. Suppose θ is to be estimated under the loss (2.21). Consider the class of priors (3.1) and let $\hat{\theta}_\alpha$ denote the Bayes estimate of θ under the prior $\pi_\alpha(\theta)$, $\alpha \in I$. Let

$$\begin{aligned} \delta &= \frac{k_1}{k_0+k_1}, \\ \underline{\tau}_\delta &= \inf_{\alpha \in I} \hat{\theta}_\alpha \\ \overline{\tau}_\delta &= \sup_{\alpha \in I} \hat{\theta}_\alpha. \end{aligned} \tag{3.3}$$

Let $\underline{\tau}_\delta^*$ and $\overline{\tau}_\delta^*$ denote the infimum and the supremum of the Bayes estimates of θ under the class of priors Γ^*. Then $\underline{\tau}_\delta^* = \underline{\tau}_\delta$ and $\overline{\tau}_\delta^* = \overline{\tau}_\delta$.

Proof: Let $\pi^*(\theta)$ be an arbitrary prior in Γ^*; hence, $\pi^*(\theta) = \int \pi_\alpha(\theta)d\tau(\alpha)$ for some probability measure τ on I. Let $\hat{\theta}^*_\delta$ denote the Bayes estimate of θ under the prior π^*; thus $\hat{\theta}^*_\delta$ is the δth posterior quantile of θ when the prior is $\pi^*(\theta)$. The proof follows from the fact that the posterior of θ under the prior π^* is a mixture of the set of posteriors generated by the class of priors Γ, and hence

$$\underline{\tau}_\delta \leq \hat{\theta}^*_\delta \leq \overline{\tau}_\delta. \tag{3.4}$$

Since π^* is arbitrary, and Γ^* contains Γ, the result follows.

Theorem 3.2 implies that the range of the Bayes estimates of an arbitrary linear combination $\underset{\sim}{c}'\underset{\sim}{\theta}$ can be computed by simply appealing to (2.23) and (2.24) when the class of priors consists of arbitrary mixtures of the $N(\underset{\sim}{\mu}, \Sigma)$ distribution, where $\Sigma_1 \leq \Sigma \leq \Sigma_2$, and when the loss is the piecewise linear loss of (2.21).

As discussed before, in the problem of estimating a linear combination $\underset{\sim}{c}'\underset{\sim}{\theta}$, the set $S(c)$ defined in Theorem 2.3 is of independent interest by itself. Under mixture normal priors, even though the range of the means of $\underset{\sim}{c}'\underset{\sim}{\theta}$ is the same as that for the regular normal priors, the set $S(c)$ as such is usually a genuine superset of the ellipse (2.13) and is *no longer* an ellipse. In the following paragraphs we describe an easy method to obtain the set $S(c)$ for mixture normal priors; in particular, we will demonstrate for what values of the least squares estimate $\hat{\underset{\sim}{\theta}}_L$ (which is the minimal sufficient statistic for $\underset{\sim}{\theta}$), the set $S(c)$ *remains the same as before*, i.e., is again the ellipse defined in (2.13), (2.14), and (2.15).

Consider, instead of the set $S(c)$, the set $M(c)$ of the vectors of first two posterior moments of $\underset{\sim}{c}'\theta$, as the prior $\pi(\underset{\sim}{\theta})$ varies in the set Γ of $N(\underset{\sim}{\mu}, \Sigma)$ priors, $\Sigma \leq \Sigma_1 \leq \Sigma_2$; thus,

$$\begin{aligned} M(c) &= \left\{ \left(E\underset{\sim}{c}'\underset{\sim}{\theta}|\underset{\sim}{Y} = \underset{\sim}{y}, E(\underset{\sim}{c}'\underset{\sim}{\theta})^2|\underset{\sim}{Y} = \underset{\sim}{y}\right) : \pi \in \Gamma \right\} \\ &= \left\{ (u_1, u_1^2 + u_2) \colon \ (u_1, u_2) \in S(c) \right\} \end{aligned} \tag{3.5}$$

Similarly, let $M^*(c)$ denote the set of first two posterior moments of $\underset{\sim}{c}'\underset{\sim}{\theta}$ when the prior $\pi(\underset{\sim}{\theta})$ varies in the mixture class Γ^*. Clearly, $M^*(c)$ is a convex set in $\mathbb{R}^2$ because Γ^* is a convex class of priors, and in fact, $M^*(c)$ is precisely the convex hull of $M(c)$. If now

$S^*(c)$ denotes the set of two dimensional vectors of the posterior mean and the posterior variance of $\underset{\sim}{c}'\underset{\sim}{\theta}$ when the prior changes in the mixture class Γ^*, i.e.,

$$S^*(c) = \{(E\underset{\sim}{c}'\underset{\sim}{\theta}|\underset{\sim}{Y} = \underset{\sim}{y},\ \mathrm{Var}(\underset{\sim}{c}'\underset{\sim}{\theta})|\underset{\sim}{Y} = \underset{\sim}{y}):\ \ \pi \in \Gamma^*\}, \tag{3.6}$$

then clearly,

$$S^*(c) = \{(u, v - u^2):\ \ (u,v) \in M^*(c)\}. \tag{3.7}$$

Thus, $S^*(c)$ can be generated by following the sequence

$$S(c) \to M(c) \to \text{ Convex hull of } M(c) \to S^*(c).$$

Thus it is immediate that $S^*(c) = S(c)$ if $M(c)$ is itself convex. Usually, however, $M(c)$ will not be a convex set because Γ is not a convex class of priors. Sometimes, rather fortunately, $M(c)$ is convex. We will provide a condition under which this is the case. Without loss, let us assume that the error variance $\sigma^2 = 1$. It is easy to see that the lower boundary of $S^*(c)$ is the same as the lower boundary of $S(c)$. Thus we only need consider points on the upper boundary $\bar{\partial}S(c)$ of $S(c)$. Recall that $(u_1, u_2) \in \bar{\partial}S(c)$

$$\Rightarrow u_2 = u_{02} + \rho\frac{\sigma_2}{\sigma_1}(u_1 - u_{01}) + \frac{\sigma_2}{\sigma_1}\sqrt{1-\rho^2}\sqrt{(A\sigma_1)^2 - (u_1 - u_{01})^2}. \tag{3.8}$$

Define $W = u_1^2 + u_2$. From (3.8) it follows that

$$\frac{d^2W}{du_1^2} = 2 - \frac{\sigma_2}{\sigma_1}\sqrt{1-\rho^2}\cdot\frac{(A\sigma_1)^2}{((A\sigma_1)^2 - (u_1 - u_{01})^2)^{3/2}}. \tag{3.9}$$

Thus the upper boundary of $M(c)$ is concave (i.e., $M(c)$ is convex from below) if

$$\begin{aligned}
&2 \le \frac{\sigma_2}{\sigma_1}\sqrt{1-\rho^2}\frac{(A\sigma_1)^2}{((A\sigma_1)^2 - (u_1 - u_{01})^2)^{3/2}} \quad \forall\ \ u_1\\
&\Leftrightarrow 2 \le \frac{\sigma_2}{\sigma_1}\frac{\sqrt{1-\rho^2}}{A\sigma_1}\\
&\Leftrightarrow \sigma_1^2 \le \sqrt{1-\rho^2} \text{ since (2.16) implies } A = \frac{\sigma_2}{2}.
\end{aligned} \tag{3.10}$$

Thus, from (2.15), it follows that $S^*(c) = S(c)$ if

$$(\underset{\sim}{v}'(\Lambda_2 - \Lambda_1)\underset{\sim}{v})^2 \le 1 - \frac{(\underset{\sim}{c}'(\Lambda_2 - \Lambda_1)\underset{\sim}{v})^2}{\underset{\sim}{c}'(\Lambda_2 - \Lambda_1)\underset{\sim}{c}\cdot\underset{\sim}{v}'(\Lambda_2 - \Lambda_1)\underset{\sim}{v}} \tag{3.11}$$

Example 8. Let $p = 2$, $n = 20$, $\underset{\sim}{\mu} = 0$, $\Sigma_1 = I$, $\Sigma_2 = 5I$, $X'X = 20I$. Let $\underset{\sim}{\hat{\theta}}_L = (\hat{\theta}_1, \hat{\theta}_2)'$ denote the least squares estimate of $\underset{\sim}{\theta}$. Then (3.11) reduces to

$$.56906(\underset{\sim}{\hat{\theta}}_L' \underset{\sim}{\hat{\theta}}_L)^2 \leq 1 - \frac{(\underset{\sim}{c}' \underset{\sim}{\hat{\theta}}_L)^2}{\underset{\sim}{c}' \underset{\sim}{c} \cdot \underset{\sim}{\hat{\theta}}_L' \underset{\sim}{\hat{\theta}}_L} \tag{3.12}$$

In particular, if $\underset{\sim}{c} = (0\ 1)'$, then (3.12) further reduces to

$$.56906(\hat{\theta}_1^2 + \hat{\theta}_2^2)^3 \leq \hat{\theta}_1^2. \tag{3.13}$$

So, for example, if the least squares estimate of the intercept θ_1 is .5, then the set $S(c)$ would remain invariant under mixing if the least squares estimate of the slope is between $\pm.71428$. Typically, $S(c)$ would remain invariant under mixing if the least squares estimate $\underset{\sim}{\hat{\theta}}_L$ is close to zero; this is just saying that if the likelihood function is concentrated near the location of the priors, then good Bayesian robustness will obtain.

Let us now briefly consider the problem of obtaining the set $M^*(c)$ when (3.10) does not hold and therefore $M(c)$ is not convex. Observe that (3.10) certainly holds (i.e., $\frac{d^2W}{du_1^2} < 0$) if $u_1 = u_{01} \pm A\sigma_1$, and therefore by continuity near $u_1 = u_{01} \pm A\sigma_1$. Geometrically, it is clear that the upper boundary of the convex hull of $M(c)$ coincides with the upper boundary of $M(c)$ for u_1 near $u_{01} \pm A\sigma_1$; in between, it is a straightline L joining two appropriate points $\underset{\sim}{P} = (u_{11}^*, w_{11}^*)$ and $\underset{\sim}{Q} = (u_{21}^*, w_{21}^*)$ on the upper boundary of $M(c)$ (see Figure 3). These two points must be such that L is tangent to $M(c)$ at the points $\underset{\sim}{P}$ and $\underset{\sim}{Q}$. Consequently, $\frac{dw}{du_1}$ assumes the same value at $\underset{\sim}{P}$ and $\underset{\sim}{Q}$; using the fact that $W = u_1^2 + u_2$ and expression (3.8), one then has

$$2(u_{21}^* - u_{11}^*) = \frac{\sigma_2\sqrt{1-\rho^2}}{\sigma_1}\left[\frac{u_{21}^* - u_{01}}{\sqrt{(A\sigma_1)^2 - (u_{21}^* - u_{01})^2}} - \frac{u_{11}^* - u_{01}}{\sqrt{(A\sigma_1)^2 - (u_{11}^* - u_{01})^2}}\right]. \tag{3.14}$$

Also, since $\underset{\sim}{P}$ and $\underset{\sim}{Q}$ both lie on L, it is clear that

$$\left.\frac{dW}{du_1}\right|_{\underset{\sim}{P}} = \left.\frac{dW}{du_1}\right|_{\underset{\sim}{Q}} = \frac{w_{21}^* - w_{11}^*}{u_{21}^* - u_{11}^*},$$

which reduces to

$$\begin{aligned} &\frac{\sigma_2}{\sigma_1}\sqrt{1-\rho^2}\left(\sqrt{(A\sigma_1)^2 - (u_{11}^* - u_{01})^2} - \sqrt{(A\sigma_1)^2 - (u_{21}^* - u_{01})^2}\right) \\ &= \left(2u_{21}^* - \frac{\sigma_2}{\sigma_1}\sqrt{1-\rho^2} \cdot \frac{(u_{21}^* - u_{01})}{\sqrt{(A\sigma_1)^2 - (u_{21}^* - u_{01})^2}}\right) \cdot (u_{11}^* - u_{21}^*) \end{aligned} \tag{3.15}$$

Equations (3.14) and (3.15) give the required values u_{11}^* and u_{21}^*. In general, (3.14) and (3.15) have to be solved numerically; in some cases symmetry arguments can be given leading to an easy derivation of u_{11}^*, u_{21}^*. Here is an example.

Example 9. An interesting special case is when $u_{01} = \rho = 0$ (i.e., the two axes of the ellipse $S(c)$ are parallel to the coordinate axes and the posterior mean varies in an interval symmetric about zero). In this case, (3.11) immediately implies that $S^*(c) = S(c)$ if

$$\sigma_1^2 = \underset{\sim}{v}'(\Lambda_2 - \Lambda_1)\underset{\sim}{v} \leq 1.$$

Otherwise, $S^*(c)$ is a superset of $S(c)$; one question of specific interest then is how much does the maximum posterior variance of $\underset{\sim}{c}'\underset{\sim}{\theta}$ increase due to prior mixing. An easy symmetry argument implies that in this example $u_{11}^{*2} = u_{21}^{*2} = \frac{\sigma_2^2}{4}(\sigma_1^2 - \frac{1}{\sigma_1^2})$. Using this, it is not difficult to prove that the maximum posterior variance increases from $u_{02} + \frac{\sigma_2^2}{2}$ to $u_{02} + \frac{\sigma_2^2}{4}(\sigma_1^2 + \frac{1}{\sigma_1^2})$. Evidently, the increase in the maximum posterior variance is nominal if $\sigma_1^2 = \underset{\sim}{v}'(\Lambda_2 - \Lambda_1)\underset{\sim}{v}$ is only marginally bigger than 1, but the increase can be substantial if σ_1^2 is substantially bigger than 1. Since σ_1^2 is likely to be smaller for $\hat{\underset{\sim}{\theta}}_L$ close to $\underset{\sim}{\mu}$, once again we find that prior mixing will have a less pronounced effect when the likelihood function is concentrated near the common mean of the priors. Finally, we give one example where the set $S^*(c)$ is found numerically by using the $S(c) \to M(c) \to M^*(c) \to S^*(c)$ algorithm.

Example 10. Let $p = 2$, $n = 10$, $\underset{\sim}{\mu} = \underset{\sim}{0}$, $\Sigma_1 = I$, $\Sigma_2 = 5I$, $X'X = 10I$, $\hat{\underset{\sim}{\theta}}_L = (1.5, 1.5)'$ and $\underset{\sim}{c} = (-1, 1)'$. The ellipse $S(c)$ in this case is given by

$$.31167u_1^2 + 70.12622(u_2 - .18895)^2 \leq .00356. \tag{3.16}$$

The maximum posterior variance equals .19608. If prior mixtures are allowed, the maximum posterior variance increases to .20137, a marginal 2.70% increase. The set $S(c)$ and the enlarged set $S^*(c)$ are plotted in Figure 3.

<u>Remarks and discussion</u>. If in the definition of the family of priors Γ in (1.8), Σ is allowed to be an arbitrary matrix, then the set of Bayes estimates of $\underset{\sim}{\theta}$ under a quadratic loss turns

out to be the ellipsoid S

$$(\underset{\sim}{\theta} - (\frac{M^{-1}\underset{\sim}{v}}{2} + \underset{\sim}{\mu}))'M(\underset{\sim}{\theta} - \frac{(M^{-1}\underset{\sim}{v} + \underset{\sim}{\mu})}{2}) \le \frac{\underset{\sim}{v}'M^{-1}\underset{\sim}{v}}{4}, \tag{3.17}$$

where, as usual, $M = X'X$, and $v = X'(\underset{\sim}{y} - X\underset{\sim}{\mu})$. In this case, the diameter D of S is equal to $\sqrt{\underset{\sim}{v}'M^{-1}\underset{\sim}{v} \cdot \lambda_{\max}}$, where $\lambda_{\max}$ is the maximum eigenvalue of M^{-1}. In this case, unlike in Example 2, D does not go to zero almost surely (or even in probability) as $n \to \infty$. This says that the family of $N(\underset{\sim}{\mu}, \Sigma)$ priors with an arbitrary Σ is too big and one cannot hope to achieve robustness in this case. One good theoretical reason for considering normal priors with arbitrary Σ, however, is that completely monotone priors can be generated by taking mixtures of $N(\underset{\sim}{\mu}, \Sigma)$ priors; for example, an elliptically symmetric t-prior with mean $\underset{\sim}{\mu}$ and scale matrix Σ_0 can be written as

$$\pi(\underset{\sim}{\theta}, \underset{\sim}{\mu}, \Sigma_0) = \int_0^\infty \frac{1}{(2\pi)^{\frac{p}{2}}|\sigma^2\Sigma_0|^{\frac{1}{2}}} e^{-\frac{1}{2\sigma^2}(\underset{\sim}{\theta} - \underset{\sim}{\mu})'\Sigma_0^{-1}(\underset{\sim}{\theta} - \underset{\sim}{\mu})} dG(\sigma^2), \tag{3.18}$$

where G is an inverse gamma distribution. The need for letting Σ to be arbitrary in the family of normal priors is that the integral in (3.18) ranges over the entire half line (a t-prior cannot be generated by taking mixtures of bounded variance normal priors). Our initial study shows, however, that mixtures of bounded variance normal priors give very good approximations to Cauchy and t-priors in a very safe neighborhood of $\underset{\sim}{\mu}$; so unless we have definite reasons to worry about the behavior of the prior in the extreme tails, our results in this section on mixture normal priors when Σ is between Σ_1 and Σ_2 will be useful.

4. Priors with unknown mean.

We now consider the problem of finding the variations in posterior measures when the prior mean $\underset{\sim}{\mu}$ is allowed to change in some convex set of $\mathbb{R}^p$. The pure Bayesian way of expressing uncertainty about the location of the prior would be to put a second stage prior on $\underset{\sim}{\mu}$. This has been considered in Polachek (1984). However, one then has to worry about the robustness of the analysis with respect to the hyperparameters of this second stage prior. Changing the mean $\underset{\sim}{\mu}$ in a convex set is an attractive alternative to putting a second stage prior; this also seems more natural because often we are not that uncertain about the location of the prior and it should be quite easy to write down a neighborhood of a prior guess where we think the prior mean lies. Mathematically, changing the mean $\underset{\sim}{\mu}$ in a convex set leads to novel geometric problems not encountered in Bayesian robustness studies before and brings into attention the surprising fact that the set of Bayes estimates of $\underset{\sim}{\theta}$ need not be convex even if $\underset{\sim}{\mu}$ changes in a nice convex set (like a sphere or an ellipsoid or a solid cube). For the purpose of the following results, we will consider the class of normal priors $N(\underset{\sim}{\mu}, \Sigma)$ with $\underset{\sim}{\mu}$ in an ellipsoid and $\Sigma_1 \leq \Sigma \leq \Sigma_2$. Again, from Theorem 3.1 it will follow that the set of posterior means under the class of priors $N(\underset{\sim}{\mu}, \Sigma)$, $\underset{\sim}{\mu} \in C, \Sigma > 0$ arbitrary, will include the posterior means under all priors which can be obtained by taking mixtures of these normal priors.

<u>**Theorem 4.1**</u>. Let $\underset{\sim}{Y} \sim N(X\underset{\sim}{\theta}, \sigma^2 I)$

$$\underset{\sim}{\theta} \sim N(\underset{\sim}{\mu}, \sigma^2\Sigma),$$

where $\Sigma_1 \leq \Sigma \leq \Sigma_2$, and $\underset{\sim}{\mu}$ belongs to the ellipsoid

$$I_1 = \{\underset{\sim}{\mu} : \ (\underset{\sim}{\mu} - \underset{\sim}{\mu}_0)' A (\underset{\sim}{\mu} - \underset{\sim}{\mu}_0) \leq 1\} \tag{4.1}$$

where $\underset{\sim}{\mu}_0$ is arbitrary but fixed.

Let $P(\Lambda_2 - \Lambda_1)^{-\frac{1}{2}} M^{-1} A M^{-1} (\Lambda_2 - \Lambda_1)^{-\frac{1}{2}} P' = L$ be the spectral decomposition of $(\Lambda_2 - \Lambda_1)^{-\frac{1}{2}} M^{-1} A M^{-1} (\Lambda_2 - \Lambda_1)^{-\frac{1}{2}}$. Define

$$\begin{aligned} \underset{\sim}{Z} &= P(\Lambda_2 - \Lambda_1)^{\frac{1}{2}} M \underset{\sim}{\hat{\theta}}_L, \\ \underset{\sim}{\nu}_0 &= P(\Lambda_2 - \Lambda_1)^{\frac{1}{2}} M \underset{\sim}{\mu}_0, \\ \text{and } C &= P(\Lambda_2 - \Lambda_1)^{-\frac{1}{2}} (I - \overline{\Lambda} M) M^{-1} (\Lambda_2 - \Lambda_1)^{-\frac{1}{2}} P' \end{aligned} \tag{4.2}$$

Then the euclidean diameter of the set S^* of all posterior means of $\underset{\sim}{\theta}$ under the family of priors (4.1) is given as

$$D = \sqrt{\lambda_{\max}(\Lambda_2 - \Lambda_1)} \cdot \sup_{\underset{\sim}{\nu}_1, \underset{\sim}{\nu}_2 \in I_2} \left\{ \frac{1}{2} \|\underset{\sim}{Z} - \underset{\sim}{\nu}_1\| + \frac{1}{2} \|\underset{\sim}{Z} - \underset{\sim}{\nu}_2\| + \|C(\underset{\sim}{\nu}_1 - \underset{\sim}{\nu}_2\| \right\}, \tag{4.3}$$

where $\lambda_{\max}(\Lambda_2 - \Lambda_1)$ is the maximum eigenvalue of $(\Lambda_2 - \Lambda_1)$ and I_2 is the ellipsoid

$$I_2 = \{\underset{\sim}{\nu} : (\underset{\sim}{\nu} - \underset{\sim}{\nu}_0)' L (\underset{\sim}{\nu} - \underset{\sim}{\nu}_0) \leq 1\} \qquad (4.4).$$

Before giving the proof of Theorem 4.1, we first give an example illustrating its application in finding the diameter of the set of Bayes estimates of $\underset{\sim}{\theta}$ under quadratic loss.

__Example 11__. Consider the setup of Example 1 where $p = 2$, $n = 20$, $\Sigma_1 = I, \Sigma_2 = 5I, \hat{\underset{\sim}{\theta}}_L = (3.15 \quad 3.15)'$, and $M = 20I$. Also assume that $\underset{\sim}{\mu}$ belongs to the unit circle $\underset{\sim}{\mu}'\underset{\sim}{\mu} \leq 1$ so that in the notation of the above theorem, $\underset{\sim}{\mu}_0 = \underset{\sim}{0}$ and $A = I$. It is easy to show that

$$\begin{aligned}
\underset{\sim}{Z} &= (2.7359 \quad 2.7359)', \\
\underset{\sim}{\nu}_0 &= \underset{\sim}{0}, \\
C &= .7625I, \\
L &= 1.3256I, \text{ and} \\
\Lambda_2 - \Lambda_1 &= .001886I.
\end{aligned}$$

Therefore, by Theorem 4.1,

$$D = \sqrt{.001886} \times \sup_{\underset{\sim}{\nu}_1, \underset{\sim}{\nu}_2} \left\{ \frac{1}{2} \| \underset{\sim}{Z} - \underset{\sim}{\nu}_1 \| + \frac{1}{2} \| \underset{\sim}{Z} - \underset{\sim}{\nu}_2 \| + .7625 \| \underset{\sim}{\nu}_1 - \underset{\sim}{\nu}_2 \| \right\}, \qquad (4.5)$$

where $\underset{\sim}{\nu}_1, \underset{\sim}{\nu}_2$ belong to the circle $\underset{\sim}{\nu}'\underset{\sim}{\nu} \leq .7544$. Since the circle $\underset{\sim}{\nu}'\underset{\sim}{\nu} \leq .7544$ is rotationally invariant and so are euclidean distances, the vector $\underset{\sim}{z}$ in (4.5) can be replaced by $(\sqrt{2.7359^2 + 2.7359^2},\ 0)' = (3.8691,\ 0)'$ without changing the problem. It is not difficult to prove that for such a $\underset{\sim}{z}$, the points $\underset{\sim}{\nu}_1, \underset{\sim}{\nu}_2$ which give the required maximum in (4.5) are such that one of them, say $\underset{\sim}{\nu}_1$, is in the first or the second quadrant and the other, i.e., $\underset{\sim}{\nu}_2$, is in the third quadrant. First consider the case when $\underset{\sim}{\nu}_1$ is in the second quadrant. We may therefore let $\underset{\sim}{\nu}_1 = (r\cos\theta, r\sin\theta), \underset{\sim}{\nu}_2 = (r\cos\phi, -r\sin\phi)$, with $r^2 = .7544$ and $\frac{\pi}{2} \leq \theta, \phi \leq \pi$. Thus, we need to maximize the function

$$\begin{aligned}
h(\theta, \phi) = &\frac{1}{2}\sqrt{(a - r\cos\theta)^2 + r^2\sin^2\theta} \\
&+ \frac{1}{2}\sqrt{(a - r\cos\phi)^2 + r^2\sin^2\phi} \\
&+ \delta\sqrt{r^2(\cos\theta - \cos\phi)^2 + r^2(\sin\theta + \sin\phi)^2}
\end{aligned}$$

(where $a = 3.8691$, and $\delta = .7625$)

$$= \frac{1}{2}\Bigg[\sqrt{a^2+r^2-2ar\cos\theta}+\sqrt{a^2+r^2-2ar\cos\phi}$$

$$+ 2\delta r\sqrt{2}\sqrt{1-\cos(\theta+\phi)}\Bigg],$$

$$\frac{\pi}{2} \le \theta, \phi \le \pi. \tag{4.6}$$

Since the maximum cannot be attained on the boundary of the rectangle $[\frac{\pi}{2}, \pi] \times [\frac{\pi}{2}, \pi]$, it is attained at a point in the interior of this rectangle; hence, at this point, $\frac{\partial h(\theta,\phi)}{\partial \theta}$ and $\frac{\partial h(\theta,\phi)}{\partial \phi}$ are both zero.

Now,

$$\frac{\partial h}{\partial \theta} = \frac{ar\sin\theta}{2\sqrt{a^2+r^2-2ar\cos\theta}} + \frac{\delta r\sin(\theta+\phi)}{\sqrt{2}\sqrt{1-\cos(\theta+\phi)}} = 0,$$

$$\text{and}\quad \frac{\partial h}{\partial \phi} = \frac{ar\sin\phi}{2\sqrt{a^2+r^2-2ar\cos\phi}} + \frac{\delta r\sin(\theta+\phi)}{\sqrt{2}\sqrt{1-\cos(\theta+\phi)}} = 0$$

imply that

$$\frac{ar\sin\theta}{2\sqrt{a^2+r^2-2ar\cos\theta}} = \frac{ar\sin\phi}{2\sqrt{a^2+r^2-2ar\cos\phi}}. \tag{4.7}$$

Since in the interval $\frac{\pi}{2} \le t \le \pi$, $\sin t$ and $\cos t$ are strictly decreasing and $\sin t$ is nonnegative, it follows that $\frac{\sin t}{\sqrt{a^2+r^2-2ar\cos t}}$ must be strictly decreasing on $\frac{\pi}{2} \le t \le \pi$. Consequently, (4.7) can hold only if $\theta = \phi$ (this is just saying that the optimum $\underset{\sim}{\nu}_1$ and $\underset{\sim}{\nu}_2$ have the same x coordinate, i.e., $\underset{\sim}{\nu}_2$ is right underneath $\underset{\sim}{\nu}_1$). Therefore, we have to simply maximize the one variable function

$$h(\theta) = \sqrt{a^2+r^2-2ar\cos\theta} + \delta r\sqrt{2}\sqrt{1-\cos 2\theta},$$

$$\frac{\pi}{2} \le \theta \le \pi. \tag{4.8}$$

Denoting $\cos\theta = x$, equivalently, we have to maximize

$$g(x) = \sqrt{a^2+r^2-2arx} + 2\delta r\sqrt{1-x^2},$$

$$-1 \le x \le 0. \tag{4.9}$$

The maximum is attained at $x = -.5$, and the maximum value is 5.5157.

Consider next the case when $\underset{\sim}{\nu}_1$ is in the first quadrant and $\underset{\sim}{\nu}_2$ is in the third quadrant. We may, therefore, let $\underset{\sim}{\nu}_1 = (r\cos\theta,\ r\sin\theta)$ and $\underset{\sim}{\nu}_2 = (-r\cos\phi,\ -r\sin\phi)$, with $0 \le \theta, \phi \le \frac{\pi}{2}$. In this case, we have to maximize the function

$$\begin{aligned} h(\theta,\phi) = &\frac{1}{2}\sqrt{(a - r\cos\theta)^2 + r^2\sin^2\theta} \\ &+ \frac{1}{2}\sqrt{(a + r\cos\phi)^2 + r^2\sin^2\phi} \\ &+ \delta\sqrt{r^2(\cos\theta + \cos\phi)^2 + r^2(\sin\theta + \sin\phi)^2}. \end{aligned}$$

Calculus gives that the maximum of h is attained on the boundary of the rectangle $\{(\theta,\phi) : 0 \le \theta, \phi \le \frac{\pi}{2}\}$. On the boundary of this rectangle, (at least) one of θ and ϕ equals 0 or $\frac{\pi}{2}$. In each of the four cases: $\theta = 0,\ \theta = \frac{\pi}{2},\ \phi = 0$, and $\phi = \frac{\pi}{2}$, the function h becomes a function of one variable. Then routine one variable calculus gives that the maximums in the four cases are 5.2974, 5.5509, 5.3347, and 5.3937 respectively. Therefore, the overall maximum is 5.5509 (attained at $\theta = \frac{\pi}{2}, \phi = .86466\pi$). Hence, from (4.5), the required diameter of the set of Bayes estimates is .2411. Recall from Example 1 that if $\underset{\sim}{\mu}$ is kept fixed at $\underset{\sim}{0}$, the diameter of the set of Bayes estimates if .1697. Thus, percentage wise, varying the prior mean results in a non–negligible increase in the diameter; but the encouraging news is that, *even so*, the diameter is quite small (the radius is still only about half of .2236, the standard error of the least squares estimate). See Figure 4 for a plot of the diameters of the set of Bayes estimates as a function of $||\hat{\underset{\sim}{\theta}}_L||$. As expected, the effect of varying $\underset{\sim}{\mu}$ gets more pronounced as $\hat{\underset{\sim}{\theta}}_L$ gets large.

Before giving a proof of Theorem 4.1, we like to explicitly point out that as long as M, Σ_1, Σ_2 are proportional to the identity matrix, the technique used in this example will give the supremum in (4.3) ($\underset{\sim}{\mu}_0$ need not be zero). Also notice that in case $C = \frac{1}{2}I$, the maximization in (4.3) is equivalent to finding the triangle with the largest possible perimeter that can be constructed by joining any two points on the ellipsoid I_2 and the point $\underset{\sim}{Z}$. This is an interesting geometric problem by itself.

Proof of Theorem 4.1: From Theorem 2.1 it follows that

$$S^* = \bigcup_{\underset{\sim}{\mu} \in I_1} S_\mu,$$

where S_μ is the ellipsoid (2.1). Transform $\underset{\sim}{\theta}$ to $P(\Lambda_2-\Lambda_1)^{-\frac{1}{2}}\underset{\sim}{\theta}$ and $\underset{\sim}{\mu}$ to $\underset{\sim}{\nu} = P(\Lambda_2-\Lambda_1)^{\frac{1}{2}}M\underset{\sim}{\mu}$ where P is defined in the statement of Theorem 4.1. On doing these transformations, it follows that

$$D = \sqrt{\lambda_{\max}(\Lambda_2 - \Lambda_1)} \cdot \sup_{\underset{\sim}{\theta}_1, \underset{\sim}{\theta}_2} \|\underset{\sim}{\theta}_1 - \underset{\sim}{\theta}_2\|, \tag{4.10}$$

where $\underset{\sim}{\theta}_1, \underset{\sim}{\theta}_2 \in S^{**} = \bigcup_{\underset{\sim}{\nu} \in I_2} S_\nu^{**}$, where S_ν^{**} is the sphere

$$S_\nu^{**} = \left\{\underset{\sim}{\theta} : \ (\underset{\sim}{\theta} - \underset{\sim}{w})'(\underset{\sim}{\theta} - \underset{\sim}{w}) \leq \frac{1}{4}(\underset{\sim}{Z} - \underset{\sim}{\nu})'(\underset{\sim}{Z} - \underset{\sim}{\nu})\right\}, \tag{4.11}$$

and $\underset{\sim}{w}$, the center of the sphere S_ν^{**}, is given by

$$\underset{\sim}{w} = C^*\underset{\sim}{Z} + C\underset{\sim}{\nu}, \text{ with}$$

$C^* = P(\Lambda_2 - \Lambda_1)^{-\frac{1}{2}}\overline{\Lambda}(\Lambda_2 - \Lambda_1)^{-\frac{1}{2}}\overline{\Lambda}(\Lambda_2 - \Lambda_1)^{-\frac{1}{2}}P'$, and C is as defined in (4.2) (although it looks intimidating, derivation of (4.10) and (4.11) is really very straightforward). Let $\underset{\sim}{\theta}_1, \underset{\sim}{\theta}_2$ be any two points on the boundary of S^{**}. Suppose $\underset{\sim}{\theta}_1$ belongs to the boundary of $S_{\nu_1}^{**}$ and $\underset{\sim}{\theta}_2$ belongs to the boundary of $S_{\nu_2}^{**}$ (such $\underset{\sim}{\nu}_1$ and $\underset{\sim}{\nu}_2$ have to exist). Let

$$\underset{\sim}{w}_i = C^*\underset{\sim}{Z} + C\underset{\sim}{\nu}_i, \ i = 1, 2.$$

Then, $\|\underset{\sim}{\theta}_1 - \underset{\sim}{\theta}_2\|$

$$\begin{aligned} &\leq \|\underset{\sim}{\theta}_1 - \underset{\sim}{w}_1\| + \|\underset{\sim}{\theta}_2 - \underset{\sim}{w}_2\| + \|\underset{\sim}{w}_1 - \underset{\sim}{w}_2\| \\ &= \frac{1}{2}\|\underset{\sim}{Z} - \underset{\sim}{\nu}_1\| + \frac{1}{2}\|\underset{\sim}{Z} - \underset{\sim}{\nu}_2\| + \|C(\underset{\sim}{\nu}_1 - \underset{\sim}{\nu}_2)\|. \end{aligned} \tag{4.12}$$

$$\therefore \sup_{\theta_1, \theta_2 \in S^{**}} \|\underset{\sim}{\theta}_1 - \underset{\sim}{\theta}_2\| \leq \sup_{\underset{\sim}{\nu}_1, \underset{\sim}{\nu}_2 \in I_2} \{\frac{1}{2}\|\underset{\sim}{Z} - \underset{\sim}{\nu}_1\| + \frac{1}{2}\|\underset{\sim}{Z} - \underset{\sim}{\nu}_2\| + \|C(\underset{\sim}{\nu}_1 - \underset{\sim}{\nu}_2)\|\}. \tag{4.13}$$

To prove the opposite inequality, take any $\underset{\sim}{\nu}_1, \underset{\sim}{\nu}_2$ on the boundary of I_2. The spheres $S_{\nu_1}^{**}$ and $S_{\nu_2}^{**}$ are completely contained in S^{**}. On the other hand, the line L^* obtained by extending the line segment joining $\underset{\sim}{w}_1$ and $\underset{\sim}{w}_2$ to the boundaries of the spheres $S^{**}_{\nu_1}$ and $S^{**}_{\nu_2}$ is obviously contained in $S^{**}_{\nu_1} \cup S^{**}_{\nu_2}$ and hence contained in S^{**}.

$$\begin{aligned}
\therefore\ & \text{Diameter of } S^{**} \\
= & \sup_{\underset{\sim}{\theta}_1, \underset{\sim}{\theta}_2 \in S^{**}} \|\underset{\sim}{\theta}_1 - \underset{\sim}{\theta}_2\| \\
\geq & \text{ Length of } L^* \\
= & \frac{1}{2}\|\underset{\sim}{Z} - \underset{\sim}{\nu}_1\| + \frac{1}{2}\|\underset{\sim}{Z} - \underset{\sim}{\nu}_2\| + \|C(\underset{\sim}{\nu}_1 - \underset{\sim}{\nu}_2)\|.
\end{aligned}$$

Since $\underset{\sim}{\nu}_1, \underset{\sim}{\nu}_2$ are arbitrary, it follows that

$$\sup_{\underset{\sim}{\theta}_1, \underset{\sim}{\theta}_2 \in S^{**}} \|\underset{\sim}{\theta}_1 - \underset{\sim}{\theta}_2\| \geq \sup_{\underset{\sim}{\nu}_1, \underset{\sim}{\nu}_2 \in I_2} \left\{ \frac{1}{2}\|\underset{\sim}{Z} - \underset{\sim}{\nu}_1\| + \frac{1}{2}\|\underset{\sim}{Z} - \underset{\sim}{\nu}_2\| + \|C(\underset{\sim}{\nu}_1 - \underset{\sim}{\nu}_2)\| \right\} \tag{4.14}$$

The theorem now follows from (4.10), (4.13), and (4.14).

It clearly will be nice to have an explicit representation of the set S^* of all posterior means when the prior mean $\underset{\sim}{\mu}$ changes in a nice convex set such as an ellipsoid. We do have such a representation in some special cases. The following theorem is a result in this direction. In general, i.e., if Σ lies between two positive definite matrices Σ_1 and Σ_2, the problem of obtaining an explicit representation of the set of posterior means seems to be hard and remains an open problem.

Theorem 4.2. Consider the setup of Theorem 4.1. If $\Sigma \geq \Sigma_1$ where $\Sigma_1^{-1} = rI,\ A = \ell^{-2}I$, and $X'X = m_2 rI$, then the boundary of the set of posterior means of $\underset{\sim}{\theta}$ consists of the p-dimensional vectors $\underset{\sim}{z} + \hat{\underset{\sim}{\theta}}_L$ where $\underset{\sim}{z}$ satisfies

$$2(1+m_2)\underset{\sim}{z}'\underset{\sim}{z} - 2\ell\sqrt{\underset{\sim}{z}'\underset{\sim}{z}} + 2\underset{\sim}{z}'(\hat{\underset{\sim}{\theta}}_L - \underset{\sim}{\mu}_0) = 0. \tag{4.15}$$

In particular, if $p = 2$ and $\|\hat{\underset{\sim}{\theta}}_L - \underset{\sim}{\mu}_0\| = \ell$, then the posterior means form a translated cardioid; if $\|\hat{\underset{\sim}{\theta}}_L - \underset{\sim}{\mu}_0\| = 2\ell$, it is a translated trisectrix, and in general it is a translated limacon.

Discussion: Before proving the theorem, we present a short discussion of the assumptions and the conclusion of the theorem. The assumption that Σ_1 is a multiple of the identity matrix would not be very restrictive because if it was thought that $\Sigma \geq \Sigma_1$ where Σ_1 is not necessarily proportional to I, we could always enlarge the class of priors by observing that $\Sigma_1 \geq \lambda_{\min} I$ where $\lambda_{\min}$ is the minimum eigenvalue of Σ_1. This enlargement will not be very substantial unless the eigenvalues of Σ_1 are very widely scattered (some very large, others small). The assumption that M is proportional to the identity matrix, admittedly, is restrictive, but at least leads to an explicit representation of the posterior means in an important special case. Recall that in the theory of optimal regression designs, the optimum M matrix often turns out to be a multiple of the identity matrix. Finally, note that since cardioids, trisectrices, and Pascal's limacons are not convex, it follows that varying the prior mean in a very nice convex set would not necessarily result in a convex set of posterior means. Of course, as $m_2 \to \infty$ (which, roughly speaking, corresponds to the sample size tending to ∞), the set of posterior means approaches a circle in the setup of Theorem 4.2 and thus is asymptotically convex. More comprehensive problems of this nature are currently under investigation.

Proof of Theorem 4.2: Direct computation using (2.1) yields that S^*, the set of posterior means, consists of points $\underset{\sim}{\theta}$ where

$$\delta^2 \left(\underset{\sim}{\theta} - (\gamma \hat{\underset{\sim}{\theta}}_L + (1-\gamma)\underset{\sim}{\mu})\right)' \left(\underset{\sim}{\theta} - (\gamma \hat{\underset{\sim}{\theta}}_L + (1-\gamma)\underset{\sim}{\mu})\right) \leq (\hat{\underset{\sim}{\theta}}_L - \underset{\sim}{\mu})'(\hat{\underset{\sim}{\theta}}_L - \underset{\sim}{\mu}) \tag{4.16}$$

where $(\underset{\sim}{\mu} - \underset{\sim}{\mu}_0)'(\underset{\sim}{\mu} - \underset{\sim}{\mu}_0) \leq \ell^2$, and

$$\begin{aligned} \delta &= 2(1+m_2) \\ \gamma &= \frac{2m_2+1}{2(1+m_2)}. \end{aligned} \tag{4.17}$$

Consider the problem of characterizing the boundary of S^*. The set S^*, as such, is constructed by drawing circles of radius $\frac{1}{\delta^2}(\hat{\underset{\sim}{\theta}}_L - \underset{\sim}{\mu})'(\hat{\underset{\sim}{\theta}}_L - \underset{\sim}{\mu})$ around $\gamma\hat{\underset{\sim}{\theta}}_L + (1-\gamma)\underset{\sim}{\mu}$ where $\underset{\sim}{\mu}$ itself is any point on the boundary of the circle $(\underset{\sim}{\mu} - \underset{\sim}{\mu}_0)'(\underset{\sim}{\mu} - \underset{\sim}{\mu}_0) \leq \ell^2$. Corresponding to each such $\underset{\sim}{\mu}$, there exists a unique $\underset{\sim}{\theta}$ on the boundary of the circle (4.16) which is on the boundary of S^* (every other point in the circle (4.16) falls in the interior of S^*). The characterization of the boundary of S^* depends largely on identifying this unique $\underset{\sim}{\theta}$ for every fixed $\underset{\sim}{\mu}$.

Fix $\underset{\sim}{\mu} = \underset{\sim}{\mu}_1$; suppose the corresponding (unique) $\underset{\sim}{\theta}$ is $\underset{\sim}{\theta}_1$. Clearly, then,

$$\begin{aligned}\inf_{(\underset{\sim}{\mu}-\underset{\sim}{\mu}_0)'(\underset{\sim}{\mu}-\underset{\sim}{\mu}_0)\leq\ell^2}\Bigg[\delta^2\Big(\underset{\sim}{\theta}_1 - (\gamma\hat{\underset{\sim}{\theta}}_L + (1-\gamma)\underset{\sim}{\mu})\Big)'\Big(\underset{\sim}{\theta}_1 - (\gamma\hat{\underset{\sim}{\theta}}_L + (1-\gamma)\underset{\sim}{\mu})\Big)& \\ - (\hat{\underset{\sim}{\theta}}_L - \underset{\sim}{\mu})'(\hat{\underset{\sim}{\theta}}_L - \underset{\sim}{\mu})\Bigg]& \\ = 0& \end{aligned} \tag{4.18}$$

and this infimum must be attained at $\underset{\sim}{\mu} = \underset{\sim}{\mu}_1$. Now, minimizing the quantity in square brackets in (4.18), by straightforward algebra, is equivalent to minimizing $\underset{\sim}{\mu}'(\hat{\underset{\sim}{\theta}}_L - \underset{\sim}{\theta}_1)$, which, in turn, is equivalent to minimizing $(\underset{\sim}{\mu} - \underset{\sim}{\mu}_0)'(\hat{\underset{\sim}{\theta}}_L - \underset{\sim}{\theta}_1)$. An application of the Cauchy–Schwartz inequality gives that the minimizing $\underset{\sim}{\mu}$ satisfies

$$\underset{\sim}{\mu} = \underset{\sim}{\mu}_0 - \frac{\ell(\hat{\underset{\sim}{\theta}}_L - \underset{\sim}{\theta}_1)}{\sqrt{(\hat{\underset{\sim}{\theta}}_L - \underset{\sim}{\theta}_1)'(\hat{\underset{\sim}{\theta}}_L - \underset{\sim}{\theta}_1)}}. \tag{4.19}$$

Recall that the minimum is attained at $\underset{\sim}{\mu} = \underset{\sim}{\mu}_1$. Consequently, $\underset{\sim}{\mu}_1$ and $\underset{\sim}{\theta}_1$ share the relation

$$\underset{\sim}{\mu}_1 = \underset{\sim}{\mu}_0 - \frac{\ell(\hat{\underset{\sim}{\theta}}_L - \underset{\sim}{\theta}_1)}{\sqrt{(\hat{\underset{\sim}{\theta}}_L - \underset{\sim}{\theta}_1)'(\hat{\underset{\sim}{\theta}}_L - \underset{\sim}{\theta}_1)}}. \tag{4.20}$$

Substituting this expression for $\underset{\sim}{\mu}_1$ in (4.18) results in

$$\delta^2(\underset{\sim}{\theta}_1 - \hat{\underset{\sim}{\theta}}_L)'(\underset{\sim}{\theta}_1 - \hat{\underset{\sim}{\theta}}_L) + 2\delta(\underset{\sim}{\theta}_1 - \hat{\underset{\sim}{\theta}}_L)'(\hat{\underset{\sim}{\theta}}_L - \underset{\sim}{\mu}_0) - 2\delta\ell\sqrt{(\underset{\sim}{\theta}_1 - \hat{\underset{\sim}{\theta}}_L)'(\underset{\sim}{\theta}_1 - \hat{\underset{\sim}{\theta}}_L)} = 0. \tag{4.21}$$

If one now transforms to $\underset{\sim}{z} = \underset{\sim}{\theta}_1 - \hat{\underset{\sim}{\theta}}_L$, one immediately gets (4.15).

If $p = 2$ and $\|\hat{\underset{\sim}{\theta}}_L - \underset{\sim}{\mu}_0\| = \ell$, then letting s be the angle between $\underset{\sim}{z}$ and $(\hat{\underset{\sim}{\theta}}_L - \underset{\sim}{\mu}_0)$, (4.15) reduces to $\|\underset{\sim}{z}\| = \frac{\ell}{1+m_2}(1 - \cos s)$, which is a (rotated) cardioid; if $\|\hat{\underset{\sim}{\theta}}_L - \underset{\sim}{\mu}_0\| = 2\ell$, (4.15) reduces to $\|\underset{\sim}{z}\| = \frac{\ell}{1+m_2}(1 - 2\cos s)$, which is a (rotated) trisectrix. In general, (4.15) is

$$\|\underset{\sim}{z}\| = \frac{1}{1+m_2}(\ell - \|\hat{\underset{\sim}{\theta}}_L - \underset{\sim}{\mu}_0\|\cos s), \text{ which is a (rotated) limacon.}$$

This proves the Theorem.

For a plot of the set S^*, see Figure 5.

5. Unknown σ^2.

Many of the results for the case of a known σ^2 generalize quite easily to the case when σ^2 is unknown. To avoid repetitive argument, we merely point out which of the results generalize to the unknown case and which do not.

When σ^2 is unknown, and primary interest is in $\underset{\sim}{\theta}$, it seems reasonable to assign a fixed prior to σ^2, like a conjugate inverse gamma prior or a flat noninformative prior. So suppose that the prior structure is

$$\begin{aligned} \text{Given } \sigma^2, \quad \underset{\sim}{\theta} &\sim N(\underset{\sim}{0}, \sigma^2\Sigma), \\ \text{and } \tau = \frac{1}{\sigma^2} &\sim G(\alpha, \beta), \end{aligned}$$

where $G(\alpha, \beta)$ stands for a gamma distribution with density

$$g(\alpha, \beta) \propto \tau^{\alpha-1} e^{-\beta\tau} (\alpha, \beta \geq 0). \tag{5.1}$$

In the above, $\alpha = \beta = 0$ corresponds to the noninformative prior for a normal variance and formally, $\alpha = \beta = \infty$ corresponds to (known) $\sigma^2 = 1$ case. It is well known that in this case, the posterior distribution of $\underset{\sim}{\theta}$ is an elliptically symmetric t distribution with $n + 2\alpha$ degrees of freedom, location parameter $(M + \Sigma^{-1})^{-1} X' \underset{\sim}{y}$, and scale matrix $\frac{1}{n+2\alpha}(2\beta + \underset{\sim}{y}'\underset{\sim}{y} - \underset{\sim}{y}'X(M + \Sigma^{-1})^{-1}X'\underset{\sim}{y}) \cdot (M + \Sigma^{-1})^{-1}$ (the posterior dispersion matrix is $\frac{n+2\alpha}{n+2\alpha-2}$ times the scale matrix).

Since for a t distribution the location vector is also the mean, it is self–evident that under the quadratic loss $(\underset{\sim}{\theta} - \underset{\sim}{a})'Q(\underset{\sim}{\theta} - \underset{\sim}{a})$, the Bayes estimate of $\underset{\sim}{\theta}$ remains the same as that for the known σ^2 case (although the associated posterior expected loss *increases*, the increment being the penalty for not knowing σ^2). Consequently, the results stated in Theorem 2.1, Corollary 2.2, Theorem 4.1, and Theorem 4.2 carry over verbatim to the unknown σ^2 case. The result on mixture priors implied by Theorem 3.1 also carries over. We do not know if suitable versions of Theorems 2.3 and 2.5 are valid or not; we believe, however, that the set $S(c)$ of Theorem 2.3 is not an ellipse anymore, and Theorem 2.5 is true even though we have not been able to find a proof. Summarizing, then, the results for quadratic loss go through for the unknown σ^2 case, but the results for the piecewise linear loss of (2.21) do not go through verbatim, and we do not have explicit theorems for this loss in the unknown σ^2 case. From a practical point of view, the unknown σ^2 case is of interest, and it would be nice to have explicit results in this case for the losses in (2.21).

6. Another approach to prior modeling: Density bands.

In this section, we consider a different method of modelling prior indeterminacy. For fixed nonnegative functions $L(\underset{\sim}{\theta}, \sigma^2)$ and $U(\underset{\sim}{\theta}, \sigma^2)$ with $L \leq U$, consider the set of prior densities π which lie between the density bands L and U, i.e., $L \leq \pi \leq U$. Since Bayesian inference is invariant under multiplication of a prior by positive constants, consider formally the class of prior densities $\Gamma_{L,U}$ defined as

$$\Gamma_{L,U} = \{\alpha\pi : \; L \leq \pi \leq U, \alpha > 0\}. \tag{6.1}$$

$\Gamma_{L,U}$ is a convex class of priors and was first used in robust Bayesian inference by DeRobertis (1978) and DeRobertis and Hartigan (1981). Several very attractive features of the class $\Gamma_{L,U}$ were mentioned in section 1. We like to point out, in addition, that $\Gamma_{L,U}$ can be thought of as the family of acceptable bets for or against an event; i.e., for a measurable subset A of the parameter space, L and U correspond to the lower and upper probabilities (of the event A) and any probability in between is considered plausible. For detailed discussion, see DeRobertis (1978) and DeRobertis and Hartigan (1981). Another extremely attractive property of this class of priors is that if the likelihood function itself is considered uncertain, and it is thought that the likelihood function belongs to the convex class $\Gamma_{f,g}$ for some f and g, then the resulting set of posteriors is $\Gamma_{fL,gU}$; this is a highly reassuring stability property in the sense that theorems proved for a general convex class of distributions of the form (6.1) apply simultaneously to the set of priors, set of likelihoods, and set of posteriors. This is very very helpful for a study of decision theoretic robustness when the model and the prior are both considered indeterminate. We will have plenty of occasions to see how this stability property is useful when we relate the problems of finding extremal values of posterior measures to the Markov–Krein–Stieltjes moment problem (see Krein and Nudel'man (1977)).

For the purpose of the discussion in this section, we will let U be kL, where $k > 1$ is a fixed real number. The reason for this restriction is that for an arbitrary upper envelope U, it seems almost impossible to derive closed form robust Bayesian results when the parameter under consideration is a vector parameter. If $U = kL$ for some $k > 1$, all priors in $\Gamma_{L,U}$ have similar tail behavior. However, the band $L \leq \pi \leq kL$ contains multimodal and asymmetric priors even if L is symmetric and unimodal. Thus the closed form analytical results of this section will be useful in getting an overall understanding of the robust regression problem when there is little concern about the exact tail of the prior but we are not sure about the shape of the prior or suspect that the prior is not necessarily unimodal and we want to see the extent to which our analysis is robust by changing the prior in a nice convex class that allows a closed form theory. We start with a general theorem for the class of priors $\Gamma_{L,kL}$, and later specialize to standard choices of L.

<u>**Theorem 6.1**</u>. Let $\underset{\sim}{Y} \sim f(\underset{\sim}{y}, X, \underset{\sim}{\theta}, \sigma^2)$ and let $(\underset{\sim}{\theta}, \sigma^2)$ have a joint prior $\pi(\underset{\sim}{\theta}, \sigma^2)$ belonging to the convex band $\Gamma_{L,kL}$ where $k > 1$. If the likelihood function f and the lower envelope L are such that the marginal posterior of $\underset{\sim}{\theta}$ under the prior $\pi = L$ is elliptically symmetric, then the set of posterior means of $\underset{\sim}{\theta}$ as π changes in $\Gamma_{L,kL}$ is the p-dimensional ellipsoid (6.8).

Proof: Let $\underset{\sim}{c}'\underset{\sim}{\theta}$ be any fixed linear combination of the coordinates of $\underset{\sim}{\theta}$, and let $\pi_L(\underset{\sim}{\theta}|\underset{\sim}{y})$ denote the marginal posterior of $\underset{\sim}{\theta}$ under L. By hypothesis, for suitable $\underset{\sim}{\mu} = \underset{\sim}{\mu}(\underset{\sim}{y})$ and $D = D(\underset{\sim}{y})$,

$$\pi_L(\underset{\sim}{\theta}|\underset{\sim}{y}) = \text{ constant } \times \pi_L\left((\underset{\sim}{\theta} - \underset{\sim}{\mu})' D^{-1} (\underset{\sim}{\theta} - \underset{\sim}{\mu})\right). \tag{6.2}$$

Since the characteristic function of $(\underset{\sim}{c}'\underset{\sim}{\theta} - \underset{\sim}{c}'\underset{\sim}{\mu})(\underset{\sim}{c}'D\underset{\sim}{c})^{-\frac{1}{2}}$ is independent of $\underset{\sim}{c}$ (see Muirhead (1982, page 34)), it follows that for any $\underset{\sim}{c}$,

$$Z = (\underset{\sim}{c}'\underset{\sim}{\theta} - \underset{\sim}{c}'\underset{\sim}{\mu})(\underset{\sim}{c}'D\underset{\sim}{c})^{-\frac{1}{2}}$$

has the same density, say $g(z)$, which is symmetric about zero.

Let $\overline{\lambda} = \overline{\lambda}(\underset{\sim}{c})$ denote $\sup\limits_{\pi \in \Gamma_{L,kL}} E(\underset{\sim}{c}'\underset{\sim}{\theta}|\underset{\sim}{Y} = \underset{\sim}{y})$. From DeRobertis and Hartigan (1981), one then has

$$k\left[-\gamma \int_{z>\gamma} g(z)dz + \int_{z>\gamma} zg(z)dz\right] - \gamma \int_{z\leq\gamma} g(z)dz + \int_{z\leq\gamma} zg(z)dz = 0, \tag{6.3}$$

where $\gamma = \gamma(\underset{\sim}{y}) = (\overline{\lambda} - \underset{\sim}{c}'\underset{\sim}{\mu})(\underset{\sim}{c}'D\underset{\sim}{c})^{-\frac{1}{2}}$.

Using the facts that $\int g(z)dz = 1$ and $\int zg(z)dz = 0$, (6.4) reduces to

$$\begin{aligned}\gamma &= \frac{k-1}{k}\left[\gamma\int_{z\leq\gamma} g(z)dz - \int_{z\leq\gamma} zg(z)dz\right] \\ \Leftrightarrow h(\gamma) &= \frac{k-1}{k}\left[\int_{z\leq\gamma} g(z)dz - \frac{1}{\gamma}\int_{z\leq\gamma} zg(z)dz\right] - 1 = 0. \end{aligned}\tag{6.4}$$

Note that $\lim\limits_{\gamma\downarrow 0} h(\gamma) > 0$ and $\lim\limits_{\gamma\uparrow\infty} h(\gamma) < 0$ and $\frac{k-1}{k}[\gamma \int\limits_{z\leq\gamma} g(z)dz - \int\limits_{z\leq\gamma} zg(z)dz] - \gamma$ is a strictly decreasing function of γ if $\lim\limits_{z\to-\infty} zg(z) = 0$. Therefore, there exists a unique $\gamma = \gamma(\underset{\sim}{y})$ such that (6.4) holds. Hence, for any $\underset{\sim}{c}$,

$$\overline{\lambda}(\underset{\sim}{c}) = \underset{\sim}{c}'\underset{\sim}{\mu} + \gamma\sqrt{\underset{\sim}{c}'D\underset{\sim}{c}}. \tag{6.5}$$

Exactly a similar argument yields that

$$\begin{aligned}\underline{\lambda}(\underset{\sim}{c}) &= \inf_{\pi\in\Gamma_{L,kL}} E\left(\underset{\sim}{c}'\underset{\sim}{\theta}|\underset{\sim}{Y} = \underset{\sim}{y}\right) \\ &= \underset{\sim}{c}'\underset{\sim}{\mu} - \gamma\sqrt{\underset{\sim}{c}'D\underset{\sim}{c}}. \end{aligned}\tag{6.6}$$

$\therefore$ if $\hat{\underset{\sim}{\theta}}_\pi$ denotes the posterior mean of $\underset{\sim}{\theta}$ under any $\pi \in \Gamma_{L,kL}$, then for any $\underset{\sim}{c}$,

$$-\gamma \leq (\underset{\sim}{c}'\hat{\underset{\sim}{\theta}}_\pi - \underset{\sim}{c}'\underset{\sim}{\mu})(\underset{\sim}{c}'D\underset{\sim}{c})^{-\frac{1}{2}} \leq \gamma. \tag{6.7}$$

Since γ is independent of $\underset{\sim}{c}$, and for any $\underset{\sim}{c}$ the values $\pm\gamma$ are attained in (6.7), it now follows that $\{\hat{\underset{\sim}{\theta}}_\pi : \pi \in \Gamma_{L,kL}\}$ form the ellipsoid

$$S_L = \left\{\underset{\sim}{\theta} : (\underset{\sim}{\theta} - \underset{\sim}{\mu})'D^{-1}(\underset{\sim}{\theta} - \underset{\sim}{\mu}) \leq \gamma^2\right\}. \tag{6.8}$$

This proves the theorem.

Remark It is well known that under standard regularity conditions, the posterior distribution for general likelihood functions f and general priors L are approximately normal and hence elliptically symmetric as the sample size $n \to \infty$. The strength of Theorem 6.1 also lies in the interesting fact that the ellipsoidal representation is "approximately" valid under general conditions for large samples. See DasGupta and Studden (1988c) for details.

Corollary 6.2. (a) Under the hypotheses of Theorem 6.1, the diameter of the set of posterior means equals

$$D_L = 2\gamma \cdot \sqrt{\lambda_{\max}(D)}, \tag{6.9}$$

where γ is the unique solution to (6.4), D is as in (6.2), and $\lambda_{\max}(D)$ denotes the maximum eigenvalue of D.

(b) For any $\underset{\sim}{c}$, the posterior means of $\underset{\sim}{c}'\underset{\sim}{\theta}$ form the interval

$$\underset{\sim}{c}'\underset{\sim}{\mu} - \gamma\sqrt{\underset{\sim}{c}'D\underset{\sim}{c}} \le u \le \underset{\sim}{c}'\underset{\sim}{\mu} + \gamma\sqrt{\underset{\sim}{c}'D\underset{\sim}{c}}, \tag{6.10}$$

where $\underset{\sim}{\mu}$ is as in (6.2).

Proof: Part (a) follows from the ellipsoid representation S_L of the posterior means of $\underset{\sim}{\theta}$. Part (b) is already proved in (6.5) and (6.6).

We will now specialize to the case where the likelihood function f is normal. Two examples follow.

Example 12: Normal likelihood, priors with noninformative tail.

Let $\underset{\sim}{Y} \sim N(X\underset{\sim}{\theta}, \sigma^2 I)$ and let $L(\underset{\sim}{\theta}, \sigma^2) = \frac{1}{\sigma^2}$ (this amounts to putting independent noninformative priors on $\underset{\sim}{\theta}$ and σ^2). Then, in the notation of Theorem 6.1, π_L is an elliptically symmetric t with $n-p$ degrees of freedom, $\underset{\sim}{\mu} = \hat{\underset{\sim}{\theta}}_L$ (where $\hat{\underset{\sim}{\theta}}_L$ is the least squares estimate of $\underset{\sim}{\theta}$) and $D = \frac{1}{n-p}(\underset{\sim}{y}'\underset{\sim}{y} - \hat{\underset{\sim}{\theta}}_L' M \hat{\underset{\sim}{\theta}}_L)M^{-1}$. Thus, $D = \frac{SSE}{n-p}M^{-1}$, where SSE is the usual residual sum of squares. It follows that the density g is a univariate t with $n-p$ degrees of freedom, location parameter 0 and scale parameter 1. Equation (6.4), on manipulation, reduces to

$$c_m \cdot \frac{m}{m-1}\left(1 + \frac{\gamma^2}{m}\right)^{-\frac{m+1}{2}+1} + \gamma\left[T_m(\gamma) - \frac{k}{k-1}\right] = 0, \tag{6.11}$$

where $m = n-p, c_m = \Gamma(\frac{m+1}{2}) / \left(\sqrt{m\pi}\,\Gamma(\frac{m}{2})\right)$, and $T_m(\cdot)$ is the cdf of a standard t distribution with m degrees of freedom (see equation (3.16) in DeRobertis (1978)). For given k and n, (6.11) has to be solved numerically. Values of $\gamma = \gamma(k, n)$ are tabulated in Table 3.2 in DeRobertis (1978).

As a specific example, suppose $p = 2$, $n = 21$, $\hat{\sigma}^2 = \frac{SSE}{n-2} = 1$, $X'X = 21I$, and $\hat{\underset{\sim}{\theta}}_L = (3.15 \quad 3.15)'$ (this is roughly the setup of Example 1, except n is taken as 21 to make use of the table in DeRobertis (1978) possible); the diameter of the set of Bayes estimates of $\underset{\sim}{\theta}$ for different values of k are given below:

$\underline{k}$	2	3	4	5	6	7	8	9	10
$\underline{D_L}$	.1255	.1989	.2507	.2905	.3230	.3506	.3749	.3955	.4143

Compared with Example 1, there is much less robustness in this case. The reason for this is the flatness of the priors considered in $\Gamma_{L,kL}$ where L is noninformative. Even then, the radius of the set of posterior means is only about .21 for $k = 10$, roughly the same as the standard error of the least squares estimate (which was found to be .2236 in Example 1). The sets of posterior means for different values of k are plotted in Figure 6. Notice the marginally diminishing effect of increasing k.

Example 13. Normal likelihood, normal–gamma lower envelope. This example is like the preceding one, but here we take a proper prior as the lower envelope L. Specifically, let $L_1(\underset{\sim}{\theta}|\sigma^2)$ be $N(\underset{\sim}{0}, \sigma^2\Sigma)$ and let $L_2(\sigma^2)$ be an inverse gamma prior as in (5.1). Define the lower envelope L as $L(\underset{\sim}{\theta}, \sigma^2) = L_1(\underset{\sim}{\theta}|\sigma^2) \cdot L_2(\sigma^2)$. The marginal posterior distribution of $\underset{\sim}{\theta}$ is an elliptical multivariate t and has been described following (5.1). From Corollary 6.2 it follows that

$$D_L = 2\gamma \left\{ \frac{2\beta + SSE - \hat{\underset{\sim}{\theta}}'_L (M\Sigma + I)^{-1} M \hat{\underset{\sim}{\theta}}_L}{n + 2\alpha} \cdot \lambda_{\max}(M + \Sigma^{-1})^{-1} \right\}^{\frac{1}{2}}, \tag{6.12}$$

where λ satisfies equation (6.11) with $m = n + 2\alpha$. For specificity, let us assume an exponential prior for $\frac{1}{\sigma^2}$; this implies $\alpha = \beta = 1$. Also, let $p = 2$, $n = 22$, $\hat{\sigma}^2 = \frac{SSE}{n-2} = 1$, $X'X = 22I$, $\hat{\theta}_L = (3.15\ 3.15)'$, and $\Sigma = I$ (again, the choice of $n = 22$ enables use of Table 3.2 in DeRobertis (1978)). The diameters of the set of posterior means for different values of k are listed below. The actual sets are plotted in Figure 7. Again notice the diminishing effect of increasing k.

$\underline{k}$	2	3	4	5	6	7	8	9	10
$\underline{D_L}$	.0421	.0668	.0841	.0976	.1085	.1177	.1257	.1327	.1406

Corollary 6.2 also implies that as the prior $\pi(\underset{\sim}{\theta}, \sigma^2)$ changes in $\Gamma_{L,kL}$, the prior mean $\underset{\sim}{\mu}$ of $\underset{\sim}{\theta}$ changes in the circle

$$\underset{\sim}{\mu}'\underset{\sim}{\mu} \leq \gamma^2, \tag{6.13}$$

where γ is the solution to (6.11) with $m = 2$ ((6.13) follows on noting that under L, the marginal prior of $\underset{\sim}{\theta}$ is an elliptical t with 2 degrees of freedom, location parameter $\underset{\sim}{0}$, and scale matrix $\Sigma = I$). So, for example, if $k = 10$, then the prior mean $\underset{\sim}{\mu}$ varies in the circle $\underset{\sim}{\mu}'\underset{\sim}{\mu} \leq 4.0481$. Considering that this is indeed a substantial variation in the prior mean, the variation in the posterior mean is quite small (for $k = 10$, the diameter of the set of posterior means is only .1406).

Example 14. In this context, another question of intrinsic interest is in what way do k and n interact. Clearly, increasing k would enlarge the family of priors; so, for example, one could ask typically how many more observations would be needed to balance the effect of doubling the value of k. One way to understand how do k and n interact would be to simply look at the behavior of the diameters as k and n change. In this article we have not gone beyond considering a simple example; since the diameters involve the constants γ for which no analytical expressions seem possible, it may be worth fitting a curve to the (expected) diameters jointly in the variables k and n. A close approximation would give a very good overall idea of how these quantities interact with each other. Since $\hat{\underset{\sim}{\theta}}_L$ and SSE would obviously change with the sample size, comparison of posterior diameters may not be quite meaningful here. Instead, we list below the *expected squared diameters* of the set of Bayes estimates for different k and n, where the expectation has been taken under the marginal distribution of $\underset{\sim}{Y}$ induced by the prior L. The choice of L as the prior in this calculation is a natural choice. Also we calculate $E(D_L^2)$ rather than $E(D_L)$ because the latter calculation is more involved (although possible). $E(D_L^2)$ is not finite for any n if the prior on $\tau = \frac{1}{\sigma^2}$ is a simple exponential (i.e., if $\alpha = 1$ in (5.1)). For $\alpha > 1$, routine calculation using (6.9) gives,

$$E(D_L^2) = \frac{4\gamma^2}{(n+2\alpha)} \cdot (2\beta + p \cdot \frac{\beta}{\alpha - 1}) \cdot \lambda_{\max}(M + I)^{-1}, \tag{6.14}$$

where it has been assumed that in the prior L, Σ is equal to I. The following values are for $p = 2$, $\alpha = 2$, $\beta = 1$, and $M = nI$ (recall again that $M = nI$ is a classical design with a number of optimum properties).

n \ k	2	3	4	5	6	7	8	9	10
3	.0556	.1400	.2239	.3037	.3777	.4465	.5125	.5737	.6324
5	.0272	.0683	.1092	.1473	.1830	.2161	.2470	.2765	.3040
10	.0088	.0223	.0354	.0478	.0592	.0697	.0797	.0889	.0975
15	.0044	.0109	.0174	.0233	.0289	.0340	.0388	.0433	.0475
20	.0026	.0065	.0103	.0138	.0171	.0201	.0229	.0255	.0280
25	.0017	.0043	.0068	.0091	.0113	.0132	.0151	.0168	.0184
35	.0009	.0023	.0036	.0048	.0060	.0070	.0080	.0089	.0097

Initially, increasing the value of n has a remarkable effect on the expected diameters; for example, by simply increasing n from 3 to 5, the expected diameters get reduced by a factor of at least 2 for every k. Later on, increasing n seems to have a rather marginal effect and the diameters get reasonably stabilized when $n = 20$ or so. Also, for $n \geq 25$ or so, increasing k quite a bit does not seem to have any serious effect on the expected diameter. For example, if $n = 25$, then the diameter increases from .0113 to .0184 if k increases from 6 to 10. Also notice that the maximum standard error of the BLUE of normalized linear combinations $\underset{\sim}{c}'\underset{\sim}{\theta}$ is $\sqrt{\lambda_{\max}(M^{-1})}$, which is $\frac{1}{\sqrt{n}}$ if $M = nI$. Compared to this, the expected radius in the table above is small; even for $n = 10$, the expected radius is at most .156 ($(ED_L)^2 \leq ED_L^2 \leq .0975$), whereas $\max\limits_{\underset{\sim}{C}'\underset{\sim}{C}=1} \sqrt{\mathrm{Var}(\underset{\sim}{c}'\underset{\sim}{\hat{\theta}}_L)}$ is .316. Thus a substantial amount of variation in the priors has a small effect on the Bayes estimates compared to the uncertainty in the usual least squares estimates.

For reasons indicated in the paragraph following Corollary 2.2, the set of Bayes predictors of k future values of the response variable Y form an ellipsoid under the prior modelling of this section also. Below we state a corollary without proof.

Corollary 6.2. Consider the decision problem of predicting $Y_1, \ldots, Y_k$ corresponding to the vectors of predictor variables $\underset{\sim}{x}_{01}, \ldots, \underset{\sim}{x}_{0k}$ under a squared error loss $\sum\limits_{i=1}^{k}(y_i - \hat{y}_i)^2$. Let

$L_{k\times p} = (\underset{\sim}{x}_{01}, \ldots, \underset{\sim}{x}_{0k})'$. Then, under the hypotheses of Theorem 6.1, the set of Bayes predictors of $\underset{\sim}{Y} = (Y_1, \ldots, Y_k)'$ form the k dimensional ellipsoid

$$S_L^* = \{\underset{\sim}{u} - L\underset{\sim}{\mu})'(LDL')^{-1}(\underset{\sim}{u} - L\underset{\sim}{\mu}) \leq \gamma^2\}, \tag{6.15}$$

where $\underset{\sim}{\mu}$, D, and γ are as in (6.8).

Next, analogous to Theorem 2.3, we describe how to work out the boundary of the joint set of posterior means and posterior standard deviations of an arbitrary linear combination $\underset{\sim}{c}'\underset{\sim}{\theta}$ when the prior changes in $\Gamma_{L,kL}$. The techniques used in the Markov-Krein-Stieltjes moment problem are useful in describing the boundary of this set. We will do this in the case when the posterior distribution of $\underset{\sim}{\theta}$ under L is an elliptically symmetric t. This will then cover the cases when the likelihood function is normal and L is noninformative or conjugate normal-gamma as described in Example 13. To work out the upper and lower boundary of the mean-standard deviation sets it will suffice to calculate for each fixed value of the posterior mean, the maximum and the minimum posterior standard deviation, which in turn can obviously be found by finding the maximum and the minimum value of the second moment of $\underset{\sim}{c}'\underset{\sim}{\theta}$ corresponding to each fixed first moment. This is where the techniques of moment theory are useful. The most appealing part of the following analysis is that the set of means and standard deviations of an arbitrary $\underset{\sim}{c}'\underset{\sim}{\theta}$ is a plain location-scale translation of *a fixed set*. Each individual practitioner can thus generate his or her appropriate set very easily from this fixed set and separate computing will be unnecessary. For a given linear combination $\underset{\sim}{c}'\underset{\sim}{\theta}$, define a new scalar function $h(\underset{\sim}{\theta})$ as

$$h(\underset{\sim}{\theta}) = \frac{1}{\lambda}(\underset{\sim}{c}'\underset{\sim}{\theta} - \underset{\sim}{c}'\underset{\sim}{\nu})$$

where λ is a constant, $\underset{\sim}{\nu}$ is a fixed vector, *both are independent of* $\underset{\sim}{\theta}$, and both will be specified later. Define

$$S_h = \{(Eh(\underset{\sim}{\theta})|\underset{\sim}{y}, \sqrt{\mathrm{Var}\ h(\underset{\sim}{\theta})|\underset{\sim}{y}}\ ):\quad \pi(\underset{\sim}{\theta}, \sigma^2) \in \Gamma_{L,kL}\} \tag{6.16}$$

and

$$S_{\underset{\sim}{c}} = \{(E\underset{\sim}{c}'\underset{\sim}{\theta}|\underset{\sim}{y}, \sqrt{\mathrm{Var}\ \underset{\sim}{c}'\underset{\sim}{\theta}|\underset{\sim}{y}}\ ):\quad \pi(\underset{\sim}{\theta}, \sigma^2) \in \Gamma_{L,kL}\}.$$

Clearly, $S_{\underset{\sim}{c}} = \lambda S_h + (\underset{\sim}{c}'\underset{\sim}{\nu}, 0)$ where for any set A, and any vector γ,

$$\lambda A + \gamma = \{\lambda \underset{\sim}{Z} + \gamma:\quad \underset{\sim}{Z} \in A\}. \tag{6.17}$$

Recall now that as the prior π varies in $\Gamma_{L,kL}$, the posterior varies in Γ_{L^*,kL^*} where L^* is the posterior corresponding to the prior L. We remind the reader that the way Γ_{L^*,kL^*} is defined, a measure belonging to Γ_{L^*,kL^*} need not be a probability measure. Thus while computing posterior means and posterior standard deviations of a function $h(\underset{\sim}{\theta})$, we have to suitably normalize the posterior measure. Define then, for any fixed function h, and a fixed π^* belonging to Γ_{L^*,kL^*},

$$\begin{aligned} u &= \int \pi^* \\ v &= \int h\pi^* \\ w &= \int h^2\pi^*. \end{aligned} \tag{6.18}$$

Thus $E_{\pi^*}(h) = \frac{v}{u}$ and $E_{\pi^*}(h^2) = \frac{w}{u}$. Since Γ_{L^*,kL^*} is convex, it is quite easy to show that so is the set

$$M_h = \{(E_{\pi^*}h, E_{\pi^*}h^2): \ \pi^* \epsilon \Gamma_{L^*,kL^*}\}. \tag{6.19}$$

Hence, for a point $(\frac{v_0}{u_0}, \frac{w_0}{u_0})$ on the upper boundary of M_h, there exist constants a and b such that

$$\begin{aligned} &\frac{w}{u} \leq a\frac{v}{u} + b \qquad \forall\, (\frac{v}{u}, \frac{w}{u}) \in M_h, \\ \text{and} \qquad &\frac{w_0}{u_0} = a\frac{v_0}{u_0} + b. \end{aligned} \tag{6.20}$$

If the point $(\frac{v_0}{u_0}, \frac{w_0}{u_0})$ corresponds to an extremal measure $\pi_0^* \in \Gamma_{L^*,kL^*}$, then (6.20) implies that $\forall\, \pi^* \in \Gamma_{L^*,kL^*}$,

$$\begin{aligned} &\int (h^2 - ah - b)(\pi^* - \pi_0^*) \leq 0 \\ &\Leftrightarrow \int (h - \lambda_1)(h - \lambda_2)(\pi^* - \pi_0^*) \leq 0 \qquad \forall\, \pi^* \in \Gamma_{L^*,kL^*} \end{aligned}$$

(where λ_i are $\frac{a \pm \sqrt{a^2+4b}}{2}$; note $a^2 + 4b \geq 0$ because $\frac{w_0}{u_0} \geq (\frac{v_0}{u_0})^2$)

$$\begin{aligned} \Rightarrow \pi_0^* &= L^* \qquad \text{if } \lambda_1 \leq h \leq \lambda_2 \\ &= kL^* \qquad \text{if } h < \lambda_1 \text{ or } h > \lambda_2. \end{aligned} \tag{6.21}$$

The representation (6.21) of measures maximizing the second moment is well known in the moment theory literature; we have briefly sketched the proof for the sake of completeness and ease in reading. A representation similar to (6.21) for measures minimizing the second moment is also available; if $(\frac{v_0}{u_0}, \frac{w_0}{u_0})$ belongs to the lower boundary of M_h, then the corresponding extremal measure is of the form

$$\begin{aligned} \pi_0^* &= L^* \text{ if } h < \lambda_1 \text{ or } h > \lambda_2 \\ &= kL^* \text{ if } \lambda_1 \leq h \leq \lambda_2. \end{aligned} \tag{6.22}$$

(6.20) now implies that for a point on the upper boundary of M_h, there exist λ_1, λ_2 such that

$$w_0 = (\lambda_1 + \lambda_2)v_0 - \lambda_1\lambda_2 u_0. \tag{6.23}$$

We now specialize to the case when L^* is an elliptical t with $m-1$ degrees of freedom, location parameter $\underset{\sim}{\nu}$, and scale matrix D. Recall from the paragraph following (5.1) that $\underset{\sim}{\nu}$ and D will depend on the data $(\underset{\sim}{y})$ on the dependent variable. Also we now specialize to

$$h(\underset{\sim}{\theta}) = \frac{1}{\sqrt{\underset{\sim}{c}' D \underset{\sim}{c}}}(\underset{\sim}{c}'\underset{\sim}{\theta} - \underset{\sim}{c}'\underset{\sim}{\nu}). \tag{6.24}$$

For a measure of the form (6.21),

$$\begin{aligned} u_0 &= \int_{\lambda_1 \leq h \leq \lambda_2} L^* + k\int_{h<\lambda_1 \cup h>\lambda_2} L^*, \\ v_0 &= \int_{\lambda_1 \leq h \leq \lambda_2} hL^* + k\int_{h<\lambda_1 \cup h>\lambda_2} hL^*, \\ \text{and } w_0 &= \int_{\lambda_1 \leq h \leq \lambda_2} h^2L^* + k\int_{h<\lambda_1 \cup h>\lambda_2} h^2L^*. \end{aligned} \tag{6.25}$$

Using the fact that if L^* is an elliptical t, then for any $\underset{\sim}{c}$, $h(\underset{\sim}{\theta})$ as defined in (6.24) is a *standard* univariate t with $m-1$ degrees of freedom, one gets from (6.25),

$$\begin{aligned} u_0 &= T_{m-1}(\lambda_2) - T_{m-1}(\lambda_1) + k[1 - \{T_{m-1}(\lambda_2) - T_{m-1}(\lambda_1)\}] \\ &= k - (k-1)[T_{m-1}(\lambda_2) - T_{m-1}(\lambda_1)], \end{aligned} \tag{6.26}$$

where $T_{m-1}(t)$ is the cdf of a standard t with $m-1$ degrees of freedom. Also,

$$v_0 = \int_{\lambda_1}^{\lambda_2} tp_{m-1}(t)dt + k[\int_{-\infty}^{\lambda_1} tp_{m-1}(t) + \int_{\lambda_2}^{\infty} tp_{m-1}(t)], \tag{6.27}$$

where $p_{m-1}(t)$ is the density of a standard t with $m-1$ degrees of freedom. Routine integration by parts simplifies (6.27) to

$$v_0 = \frac{(k-1)(m-1)}{(m-2)}[\{p_{m-1}(\lambda_2)-p_{m-1}(\lambda_1)\}+\frac{1}{m-1}\{\lambda_2^2 p_{m-1}(\lambda_2)-\lambda_1^2 p_{m-1}(\lambda_1)\}]. \quad (6.28)$$

Similarly, on integrating by parts twice,

$$w_0 = \frac{m-1}{m-3}[u_0+(k-1)\{\lambda_2 p_{m-1}(\lambda_2)-\lambda_1 p_{m-1}(\lambda_1)+\frac{\lambda_2^3 p_{m-1}(\lambda_2)-\lambda_1^3 p_{m-1}(\lambda_1)}{m-1}\}] \quad (6.29)$$

Observe now that since the standard t distribution is symmetric about zero, the set M_h is also symmetric about zero in the first coordinate, i.e., $(z_1, z_2) \in M_h$ implies $(-z_1, z_2) \in M_h$. Let then $\gamma = \sup_{\pi^*} E_{\pi^*}(h)$ and $-\gamma = \inf_{\pi^*} E_{\pi^*}(h)$; indeed, the constant γ is the solution to the equation (6.11) (with m replaced by $m-1$ because we have $(m-1)$ degrees of freedom here). The maximum value γ of the mean is attained for the measure

$$\begin{aligned} \pi_1^* &= L^* \text{ if } -\infty < h \leq \gamma \\ &= kL^* \text{ if } \gamma < h < \infty, \end{aligned} \quad (6.30)$$

and the minimum value $-\gamma$ is attained for the measure

$$\begin{aligned} \pi_2^* &= L^* \text{ if } -\gamma \leq h < \infty \\ &= kL^* \text{ if } h < -\gamma. \end{aligned} \quad (6.31)$$

Notice (6.30) and (6.31) are both of the form (6.21) with $\lambda_1 = -\infty$, $\lambda_2 = \gamma$ and $\lambda_1 = -\gamma$, $\lambda_2 = \infty$ respectively. Also observe that for measures of the form (6.21), the mean is zero if and only if $\lambda_1 = -\lambda_2$. Let $M > \gamma$ be such that the measure (6.21) with $-\lambda_1 = \lambda_2 = M$ gives the unique point on the upper boundary of M_h with a first coordinate equal to zero (i.e., the mean equal to zero). The value of M can be found from the equation (6.23) which, because $-\lambda_1 = \lambda_2 = M$, reduces to

$$w_0 = M^2 u_0, \quad (6.32)$$

where u_0 and w_0 are to be evaluated using $-\lambda_1 = \lambda_2 = M$ in (6.26) and (6.29). By varying λ_2 in the range $\gamma \leq \lambda_2 \leq M$, and then solving for the corresponding λ_1 by using (6.23), we generate the right half of the upper boundary of M_h; the left half is simply the symmetric image of the right half. The method described above has been carried out by us for various values of the degrees of freedom, resulting in Figures 8 through 12.

We now very briefly describe the theory and computation involved in working out the lower boundary of M_h. Recall that the measures corresponding to points on the lower boundary are of the form (6.22). The equations corresponding to (6.26), (6.28), and (6.29) are now

$$u_0 = 1 + (k-1)[T_{m-1}(\lambda_2) - T_{m-1}(\lambda_1)], \tag{6.34}$$

$$v_0 = -\frac{(k-1)(m-1)}{(m-2)}[\{p_{m-1}(\lambda_2) - p_{m-1}(\lambda_1)\} + \frac{1}{m-1}\{\lambda_2^2 p_{m-1}(\lambda_2) - \lambda_1^2 p_{m-1}(\lambda_1)\}],$$

and

$$w_0 = \frac{m-1}{m-3}[u_0 + (k-1)\{\lambda_1 p_{m-1}(\lambda_1) - \lambda_2 p_{m-2}(\lambda_2) - \frac{\lambda_2^3 p_{m-1}(\lambda_2) - \lambda_1^3 p_{m-1}(\lambda_1)}{m-1}\}]. \tag{6.35}$$

The choice $\lambda_1 = -\infty$, $\lambda_2 = -\gamma$ in (6.22) results in the mean $-\gamma$ and the choice $\lambda_1 = \gamma$, $\lambda_2 = \infty$ results in the mean γ. Again, there is an L, where $-\gamma < L$, such that $-\lambda_1 = \lambda_2 = L$ gives the point on the lower boundary of M_h with first coordinate equal to zero. L is found from equation (6.23) ((6.23) holds for points on the lower boundary also) which, because $-\lambda_1 = \lambda_2 = L$, reduces to

$$w_0 = L^2 u_0, \tag{6.36}$$

where u_0 and w_0 are as in (6.33) and (6.35) with $-\lambda_1 = \lambda_2 = L$. By varying λ_2 in the range $-\gamma \leq \lambda_2 \leq L$, we generate the left half of the lower boundary of M_h. The right half, again, is its symmetric image.

The set S_h of (6.16) is found from M_h by using the transformation

$$(z_1, z_2) \to (z_1, \sqrt{z_2 - z_1^2}),$$

which gives the set $S_{\underset{\sim}{c}}$ for an arbitrary $\underset{\sim}{c}$ on using the relation

$$S_{\underset{\sim}{c}} = \sqrt{\underset{\sim}{c}' D \underset{\sim}{c}}\, S_h + (\underset{\sim}{c}' \underset{\sim}{\nu}, 0).$$

The standardized S_h set which produces $S_{\underset{\sim}{c}}$ for any $\underset{\sim}{c}$ is shown in Figures 8 through 12 for various degrees of freedom. A glance at the picture gives a quick idea of the effect of increasing m (which is related to the sample size) on attainable robustness (the smaller the set S_h, smaller is the set $S_{\underset{\sim}{c}}$ and thus better the robustness).

Equations (6.26), (6.28), (6.29), (6.33), (6.34), and (6.35) are fairly complicated and we seriously doubt that there is a nice representation of $S_{\underset{\sim}{c}}$ like the elliptical representation of the mean-variance set in Theorem 2.3 for normal priors. Nevertheless, the practitioner's task is made very easy by making the standardized S_h set available. The other very useful fact is that since $S_{\underset{\sim}{c}} = \sqrt{\underset{\sim}{c}' D \underset{\sim}{c}} S_h + (\underset{\sim}{c}'\underset{\sim}{\nu}, 0)$, it immediately follows that

$$\sup_{\pi^*} \sqrt{\mathrm{Var}\ (\underset{\sim}{c}'\underset{\sim}{\theta})|\underset{\sim}{y}} = \sqrt{\underset{\sim}{c}' D \underset{\sim}{c}} \times \sup_{(z_1, z_2) \in S_h} z_2,$$

and

$$\inf_{\pi^*} \sqrt{\mathrm{Var}\ (\underset{\sim}{c}'\underset{\sim}{\theta})|\underset{\sim}{y}} = \sqrt{\underset{\sim}{c}' D \underset{\sim}{c}} \times \inf_{(z_1, z_2) \in S_h} z_2;$$

thus, in order to find the range of the posterior standard deviation of an arbitrary $\underset{\sim}{c}'\underset{\sim}{\theta}$, one merely needs to scale the tip and the pit of the standardized S_h set and separate computing will not be required. An example follows.

Example 15. Consider the situation of $\underset{\sim}{Y} \sim N(X\underset{\sim}{\theta}, \sigma^2 I)$, and $L(\underset{\sim}{\theta}, \sigma^2) = L_1(\underset{\sim}{\theta}|\sigma^2) L_2(\sigma^2)$ where L_1 is $N(\underset{\sim}{0}, \sigma^2 I)$ and L_2 is the prior implied by (5.1) with $\alpha = 2$, $\beta = 1$. Suppose $p = 2$ and let $X'X = nI$ and consider the problem of estimating the slope θ_1, which means that $\underset{\sim}{c} = (0\ 1)'$. The minimum and maximum posterior standard deviation of θ_1 are listed below. Assume that $n = 16$, $\hat{\underset{\sim}{\theta}}_L = (3.15\ 3.15)'$ and $\frac{SSE}{n-2} = 1$. Recall that the marginal posterior of $\underset{\sim}{\theta}$ under L is an elliptical t with $n + 2\alpha$ degrees of freedom, mean $\underset{\sim}{\nu} = (M + I)^{-1} X' \underset{\sim}{y}$, and scale matrix

$$D = \frac{1}{n + 2\alpha}(2\beta + \underset{\sim}{y}'\underset{\sim}{y} - \underset{\sim}{y}' X (M + I)^{-1} X' \underset{\sim}{y})(M + I)^{-1}.$$

Thus, in this case, $\sqrt{\underset{\sim}{c}' D \underset{\sim}{c}} = .31936$. On the other hand, the marginal prior of $\underset{\sim}{\theta}$ under L is also an elliptical t with 2α degrees of freedom, mean $\underset{\sim}{0}$ and scale matrix $\frac{\beta}{\alpha}\Sigma = \frac{I}{2}$. Consequently, the standardized S_h set for $2\alpha = 4$ degrees of freedom immediately provides the minimum and the maximum ***prior standard deviation*** on simply multiplying the high and the low of S_h by $\sqrt{\frac{\underset{\sim}{c}'\underset{\sim}{c}}{2}} = .70711$. This gives a truly convenient ground for comparing the effect of the data on the standard deviation of $\underset{\sim}{c}'\underset{\sim}{\theta}$. Some numbers follow as a simple illustration.

k	Min prior s.d.	Max prior s.d.	Min post s.d.	Max post s.d.
2	.8134	1.2212	.2823	.3987
4	.6580	1.4797	.2334	.4649
6	.5800	1.6490	.2078	.5055
8	.5299	1.7777	.1910	.5351
10	.4938	1.8825	.1788	.5585

As k increases, the minimum posterior standard deviation decreases and the maximum increases, implying that enlarging the family of priors enlarges the range of posterior expected losses: an anticipated effect. It is also clear that after a certain stage, increasing k seems to have a diminishing effect on the posterior robustness.

Finally, we now present a result on the ranges of posterior probabilities of sets as the prior changes in a band $\Gamma_{L,U}$. The result is stated in DeRobertis (1978) and therefore the proof is omitted (in any case, the result is very easy to prove).

Theorem 6.3. Consider an arbitrary statistical decision problem where the prior changes in the class $\Gamma_{L,U}$ where L, U are arbitrary subject to $L \leq U$. Let A be any (measurable) set in the parameter space. The supremum of the posterior probability of A is attained at the prior $\overline{\pi}$ that equals U on A and L on A^c, and the infimum is attained at the prior $\underline{\pi}$ that equals L on A and U on A^c.

The strength of Theorem 6.3 lies in the striking facts that the form of the extremal priors is the same for all sets A and that the extremal priors *do not* depend on the specific data at hand (normally, they do). Since A is arbitrary in Theorem 6.3, it is easy to find the ranges of posterior probabilities of a variety of sets under the prior modelling of this section. Two specific examples follow.

Example 16. Let $\underset{\sim}{Y} \sim N(X\underset{\sim}{\theta}, \sigma^2 I)$, let $L_1(\underset{\sim}{\theta}|\sigma^2)$ be the $N(\underset{\sim}{0}, \sigma^2 I)$ density, let $L_2(\sigma^2)$ be the inverse gamma density of (5.1) with $\alpha = 2$, $\beta = 1$; let also $n = 20$, $\hat{\underset{\sim}{\theta}}_L = (3.15,\ 3.15)'$, $\frac{SSE}{n-2} = 1$, $X'X = nI$, and suppose we want to find the maximum and the minimum posterior probability of the set $A = \{\underset{\sim}{\theta}\colon\ ||\underset{\sim}{\theta}|| \le 4\}$ for various values of k. Theorem 6.3 immediately implies that the maximum posterior probability equals $\frac{kp}{1+(k-1)p}$ where $p = P(||\underset{\sim}{\theta}|| \le 4)$ when $\underset{\sim}{\theta}$ has an elliptical t distribution with $n + 2\alpha = 24$ degrees of freedom, mean $(M + I)^{-1}X'\underset{\sim}{y} = \frac{20}{21}\hat{\underset{\sim}{\theta}}_L = (3\ 3)'$, and scale matrix $D = \frac{1}{24\times 21}(2 + SSE + \frac{20}{21}\hat{\underset{\sim}{\theta}}'_L\hat{\underset{\sim}{\theta}}_L) \cdot I = .0772I$ (again, see the paragraph following (5.1)). Similarly, the minimum posterior probability equals $\frac{p}{k-(k-1)p}$. The posterior density of $||\theta||^2$ is, in this case, an infinite mixture of type 2 Beta densities and a closed form expression for p, although possible, is complicated. A simple numerical integration gives $p = .186023$; the minimum and the maximum posterior probability of $||\underset{\sim}{\theta}|| \le 4$ is given below for various k.

$\underline{k}$	2	4	6	8	10
$\underline{\text{min}}$	.1025	.0540	.0367	.0278	.0223
$\underline{\text{max}}$	.3137	.4776	.5783	.6464	.6956

Example 17. Linear combinations of the coordinates of $\underset{\sim}{\theta}$ are of special interest in regression problems. The minimum and the maximum posterior probability that $\underset{\sim}{c}'\underset{\sim}{\theta}$ lies between a and b again equal $\frac{p}{k-(k-1)p}$ and $\frac{kp}{1+(k-1)p}$ where p is the posterior probability of this event under the prior L. Computation of p is considerably easier in this case because the posterior density of $\underset{\sim}{c}'\underset{\sim}{\theta}$ under L is itself a t; so, for example, in the setup of Example 16, with $\underset{\sim}{c} = (0\ 1)'$, the posterior density of $\underset{\sim}{c}'\underset{\sim}{\theta}$ is a t with 24 degrees of freedom, mean 3 and scale parameter .0772 (i.e., variance $= \frac{12}{11} \times .0772 = .0842$). Therefore, under L, the posterior probability that θ_1 is between ± 3.5 approximately equals .9573. Using this value, the following table immediately follows.

$\underline{k}$	2	3	4	5	8	10
$\inf P(\lvert\underset{\sim}{c}'\underset{\sim}{\theta}\rvert \le 3.5\lvert\underset{\sim}{y})$	.9181	.8820	.8486	.8176	.7370	.6915
$\sup P(\lvert\underset{\sim}{c}'\underset{\sim}{\theta}\rvert \le 3.5\lvert\underset{\sim}{y})$	.9782	.9853	.9890	.9911	.9944	.9955

It seems that the maximum posterior probability is more robust than the minimum, which is expected because the event under consideration was a 'favorable' event to start with. Also, again the effect of increasing k on the range of the posterior probability seems to diminish for large values of k.

Example 18. Using the techniques of Examples 2 and 16, it is again possible to prove that under very general conditions the Euclidean diameter of the set of Bayes estimates under the family of priors considered in this section converges to zero a.s. as $n \to \infty$ for *all* marginal distributions induced by the priors in $\Gamma_{L,U}$. We sketch the proof below in the case L is the normal-gamma prior of Example 13 with $\Sigma = I$, and $U = kL$, where $k > 1$.

From Corollary 6.2, the diameter equals

$$D_n = 2\gamma \cdot \sqrt{\lambda_{\max}(D)},$$

where γ depends on n and is to be found from equation (6.11). Note $\gamma = \gamma(n)$ decreases with n. As in Example 2, we will prove that for any $\epsilon > 0$, $\sum_{n=1}^{\infty} P_{m_\pi}(D_n > \epsilon) < \infty$, where m_π is the marginal density of $\underset{\sim}{y}$ induced by $\pi \in \Gamma_{L,kL}$.

Now note that

$$\Sigma P_{m_\pi}(D_n > \epsilon) \leq \Sigma P_{m_\pi}(h(\underset{\sim}{y}) \cdot \frac{\lambda_n}{\delta^2} > n) \tag{6.37}$$

where $h(\underset{\sim}{y}) = 2\beta + \underset{\sim}{y}'\underset{\sim}{y}$, $\lambda_n = \lambda_{\max}(M+I)^{-1}$, and $\delta^2 = \frac{\epsilon^2}{4\gamma^2(1)}$. (6.37) follows from the expression for D following (5.1), and the facts that $\gamma(n) \leq \gamma(1)$, and $\alpha > 0$. Now using the fact that if $\pi \in \Gamma_{L,kL}$, then $m_\pi \in \Gamma_{m_L,km_L}$, and hence that

$$P_{m_\pi}(h(\underset{\sim}{y}) \cdot \frac{\lambda_n}{\delta^2} > n) \leq \frac{kp_n}{1+(k-1)p_n}, \tag{6.38}$$

where $p_n = P_{m_L}(h(\underset{\sim}{y}) \cdot \lambda_n \delta^{-2} > n)$, clearly it is enough to prove that $\Sigma p_n < \infty$ (for inequality (6.38), refer back to Example 16). But, by Chebyshev's inequality,

$$\begin{aligned} \Sigma p_n &= \Sigma P_{m_L}(h^2(\underset{\sim}{y}) \cdot \frac{\lambda_n^2}{\delta^4} > n^2) \\ &\leq \Sigma E(h^2(\underset{\sim}{y}))\frac{\lambda_n^2}{n^2\delta^4}. \end{aligned} \tag{6.39}$$

Since under m_L, typically $Eh^2(\underset{\sim}{y}) = 0(n^2)$, the result will follow from (6.39) as long as $\Sigma\lambda_n^2 < \infty$. Typically, $\lambda_n = 0(\frac{1}{n})$ and hence the result of this example will typically be true. In (6.39), interestingly, the first order Chebyshev inequality that $P(h(\underset{\sim}{y}) \cdot \frac{\lambda_n}{\delta^2} > n) \leq \frac{Eh(\underset{\sim}{y}) \cdot \lambda_n}{n\delta^2}$ fails to establish that $\Sigma p_n < \infty$ because $Eh(\underset{\sim}{y})$ is $0(n)$ under m_L and $\Sigma\lambda_n$ will typically diverge.

7. Closing remarks, future research, conclusions.

This article demonstrates that it may be possible to derive closed form results in multiparameter problems on robustness with respect to the prior for different classes of priors. In DasGupta and Studden (1988b), we build on the results of this article (especially, Theorems 2.1, 2.3, 6.1, and the results on the set S_h in section 6) to answer such questions: for a nonnegative measure λ (like Lebesgue measure), which set C minimizes $\lambda(C)$ subject to $\inf_{\pi} P(\underset{\sim}{\theta} \in C|\underset{\sim}{y}) \geq 1-\alpha$? That will be the smallest volume robust Bayes confidence set. See L. LeCam's discussion on Diaconis and Freedman (1986). The importance of our closed form results and formulae also rests on the fact that one can carry on with the statistically interesting problem of deriving the most or nearly the most robust inference. To give a specific example, Theorem 2.1 implies that the squared diameter of the set of Bayes estimates equals $\lambda_{\max}\{(M+\Sigma_2^{-1})^{-1} - (M+\Sigma_1^{-1})^{-1}\} \times (\hat{\underset{\sim}{\theta}}_L - \underset{\sim}{\mu})' M\{(M+\Sigma_2^{-1})^{-1} - (M+\Sigma_1^{-1})^{-1}\} M(\hat{\underset{\sim}{\theta}}_L - \underset{\sim}{\mu})$. One can then ask what kind of a design matrix should be used so that we are nearly Bayes under a fixed prior in the family of priors under consideration and most robust among all designs which are nearly Bayes for this fixed prior. Such Bayes optimal design questions arise very naturally from the present article and are treated in DasGupta and Studden (1988a).

Much work remains to be done. For the density bands of section 6, some more restrictions like a fixed mean or a fixed median seem quite natural. This is treated in DasGupta and Studden (1988c). The form of the extremal priors under such additional restrictions are worked out in that article. In DasGupta and Studden (1988d), we attempt to tie together the robust Bayesian and the classical perspectives; for example, it is proved that if a *conditional* Γ *minimax* procedure is developed by using either of the two family of priors considered in this article, then one can get a frequentist admissible and/or minimax procedure. Aesthetically, a more attractive way to consider robustness would be to vary the likelihood, the loss, and the prior simultaneously. Hopefully, this article gives some insight into such comprehensive robustness problems.

References

Abramowitz, M. and Stegun, I. A. (1964). Handbook of mathematical functions. National Bureau of Standards, *App. Math. Series.*

Berger, James (1984). The robust Bayesian viewpoint. Robustness in Bayesian Statistics. J. Kadane (ed.), North-Holland, Amsterdam.

Berger, James and Berliner, L. M. (1986). Robust Bayes and empirical Bayes analysis with ε-contaminated priors. *Ann. Statist.* **14**, 461–486.

Berger, James and Sivaganesan, S. (1986). Ranges of posterior measures for priors with unimodal contaminations. Tech. Report #86-41, Dept. of Statistics, Purdue University.

Berger, James and Delampady, M. (1988). Lower bounds on Bayes factors for multinomial and chi-squared tests of fit.

Berger, James (1987). Robust Bayesian analysis: sensitivity to the prior. Tech. report #87-10, Dept. of Statistics, Purdue University.

Brown, L. D. and Hwang, J. T. (1988). Universal domination and stochastic domination: U-admissibility and U-inadmissibility of the least squares estimator. To appear in *Ann. Statist.*

DasGupta, Anirban and Studden, W. J. (1988a). Robust Bayesian analysis and optimal experimental designs in normal linear models with many parameters–II. Technical report, Dept. of Statistics, Purdue University.

DasGupta, Anirban and Studden, W. J. (1988b). Frequentist behavior of smallest volume robust Bayes confidence sets. Tech. report, Dept. of Statistics, Purdue University.

DasGupta, Anirban and Studden, W. J. (1988c). Variations in posterior measures for priors in a band: effect of additional restrictions. Tech. Report, Dept. of Statistics, Purdue University.

DasGupta, Anirban and Studden, W. J. (1988d). Frequentist behavior of robust Bayes procedures: new applications of the Wald-Lehmann minimaxity theory. Tech. report, Dept. of Statistics, Purdue University.

DeRobertis, L. and Hartigan, J. A. (1981). Bayesian inference using intervals of measures. *Ann. Statist.* **9**, 235–244.

Diaconis, P. and Freedman, D. (1986). On the consistency of Bayes estimates. *Ann. Statist.* **14**, 1–67.

Edwards, W., Lindman, H. and Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, **70**, 193–242.

Fix, E. (1949). Tables of noncentral chi square, Univ. of Calif. Publ. in Stat. 1.

Goldstein, M. (1980). The linear Bayes regression estimator under weak prior assumptions. *Biometrika* **67**, 621–628.

Good, I. J. and Crook, J. F. (1987). The robustness and sensitivity of the mixed-Dirichlet Bayesian test for "independence" in contingency tables. *Ann. Statist.* **15**, 670–693.

Hampel, F. R. (1971). A general qualitative definition of robustness. *Ann. Math. Statist.* **42**, 1887–1896.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Jour. Amer. Stat. Assoc.* **69**, 383–393.

Hampel, F. R. (1975). Beyond location parameters: robust concepts and methods (with discussion). *Proc. 40th session ISI*, Vol. XLVI, Book 1, 375–391.

Hartigan, J. A. (1969). Linear Bayesian methods. *J. Roy. Statist. Soc. Ser. B* **31**, 446–454.

Hills, B. (1980). Robust analysis of the random model and weighted least squares regression. Evaluation of Econometric Models. Academic Press, New York.

Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35**, 73–101.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proc. 5th Berkeley Symp.* **1**, 221–233.

Huber, P. J. (1973). The use of Choquet capacities in Statistics. *Proc. 39th Session ISI*, Vol. 45, 181–188.

Huber, P. J. (1977). Robust methods of estimation of regression coefficients. *Math. Oper. Statist. Ser. Statist.* **8**, 41–53.

Hwang, J. T. (1985). Universal domination and stochastic domination-decision theory simultaneously under a broad class of loss functions. *Ann. Statist.* **13**, 295–314.

Kadane, J. and Chuang, D. T. (1978). Stable decision problems. *Ann. Statist.* **6**, 1095–1110.

Krein, M. G. and Nudel'man, A. A. (1977). The Markov moment problem and extremal problems. *Amer. Math. Soc.*, Providence Rhode Island.

Kudō, H. (1967). On partial prior information and the property of parametric sufficiency. *Proc. 5th Berkeley Symp.* **1**.

Leamer, E. E. (1978). Specification searches: ad hoc inference with nonexperimental data. John Wiley, New York.

Leamer, E. E. (1982). Sets of posterior means with bounded variance prior. *Econometrica* **50**, 725–736.

Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model. *J. Roy. Statist. Soc. (Ser. B)* **34**, 1–41.

O'Hagan, A. and Berger, J. (1988). Ranges of posterior probabilities for quasiunimodal priors with specified quantiles. *Jour. Amer. Statist. Assoc.* **83**, 503–508.

Polachek, W. (1985). Sensitivity analysis for general and hierarchical linear regression models. Bayesian Inference and Decision techniques with Applications, Eds. P. K.

Goel and A. Zellner. North-Holland, Amsterdam.

Potzelberger, Klaus (1988). HPD regions in linear regression. Probability and Bayesian Statistics, Ed. R. Viertls, Plenam Press, New York.

SAS User's Guide (1985). Version 5, SAS Institute, Inc., Cary, North Carolina.

Tong, Y. L. (1980). Probability inequalities in multivariate distributions. Academic Press, New York.

Wolfenson, M. and Fine, T. L. (1982). Bayes-like decision making with upper and lower probabilities. *Jour. Amer. Statist. Assoc.* **77**, 80–88.

Figure 1

Set of posterior means of θ : Normal priors, fixed prior mean
(Example 1)

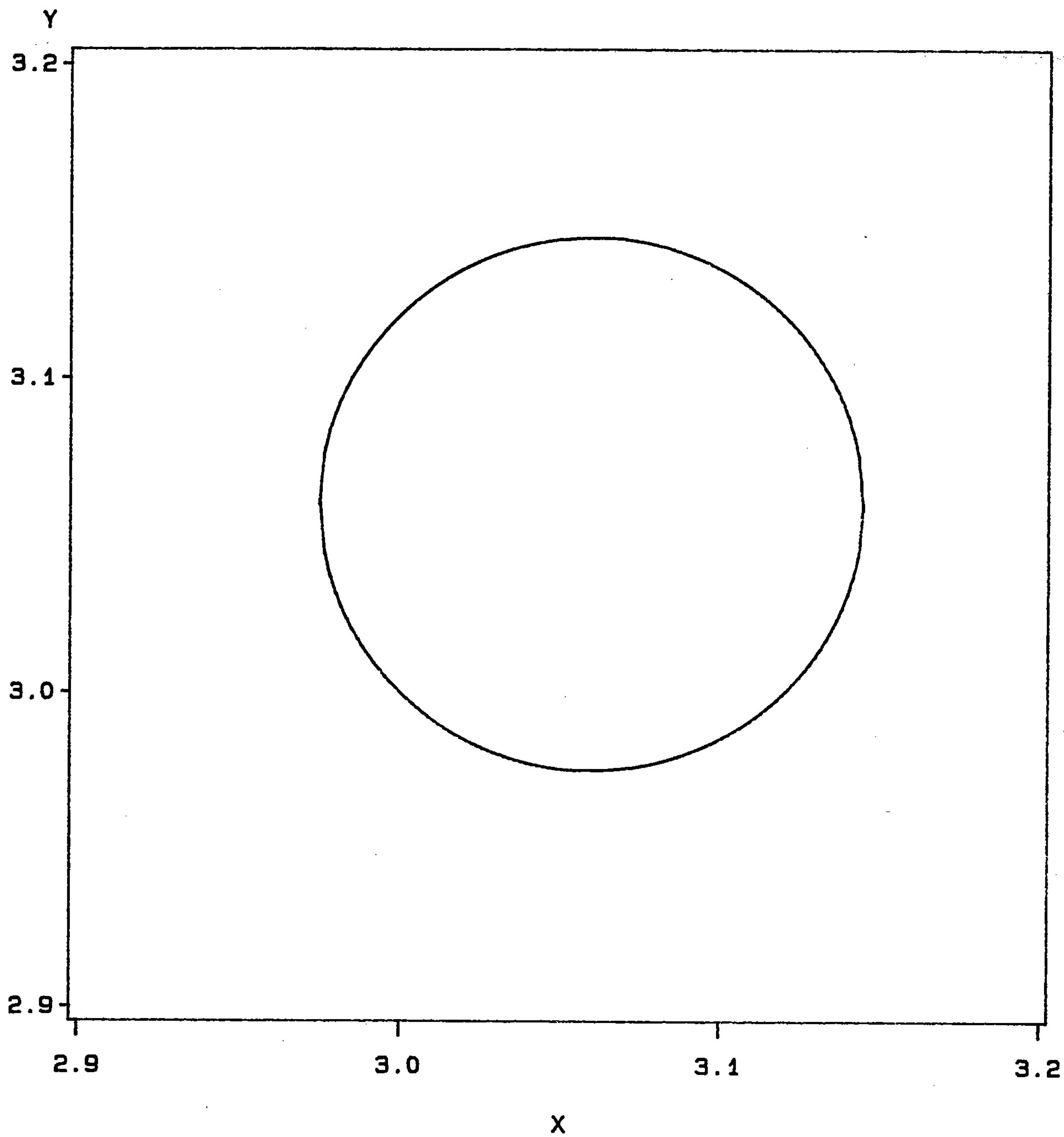


Figure 2

Set of posterior means and variances of the slope:
Normal Priors (Example 3)

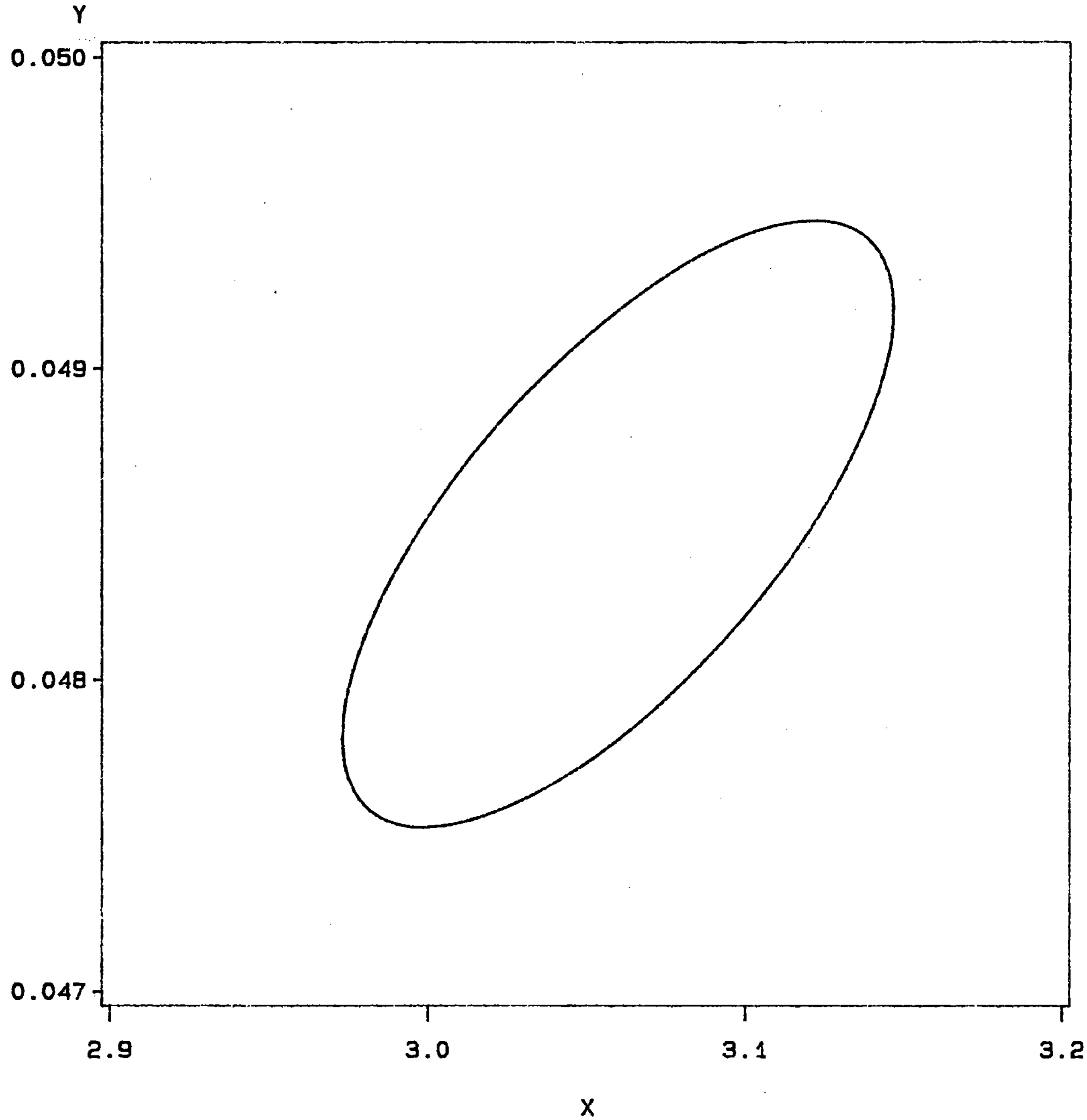


Figure 3

Set of posterior means and variances: normal and mixture normal priors (Example 10)

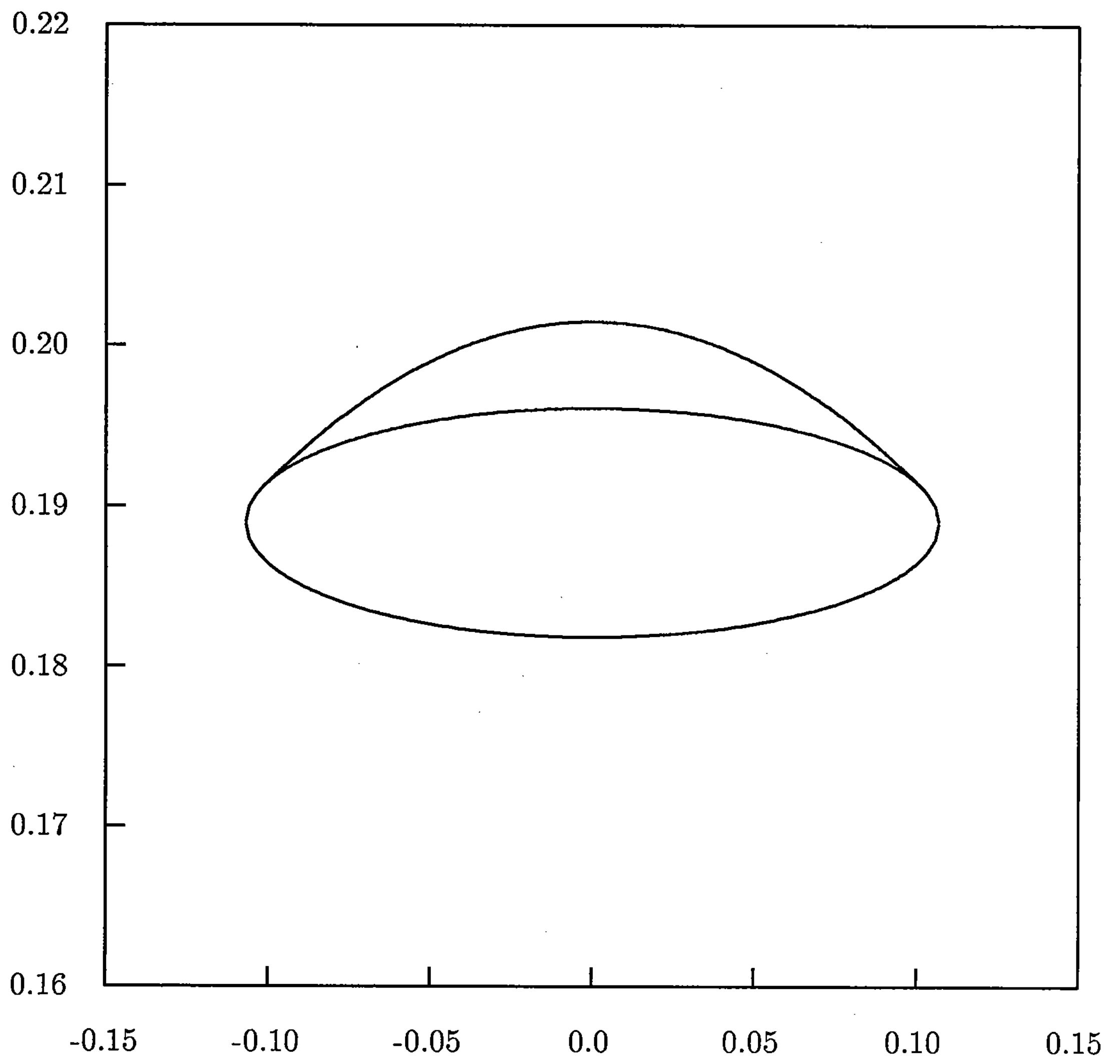


Figure 4

Plot of diameter of set of Bayes estimates vs. the least squares estimate:
normal priors. variable mean (Example 11)

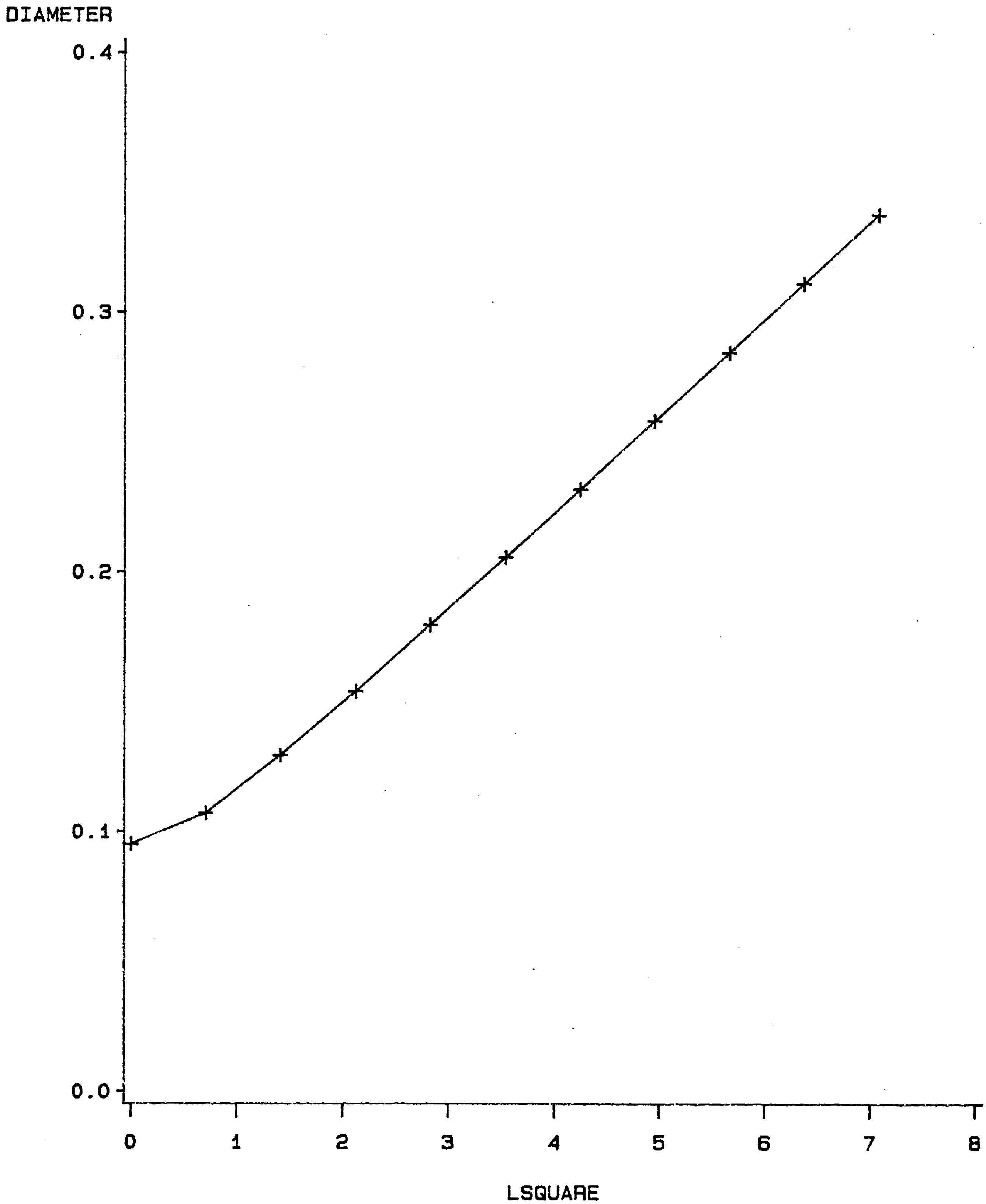


Figure 5

Set of posterior means: Normal priors, variable prior mean
(Example 11)

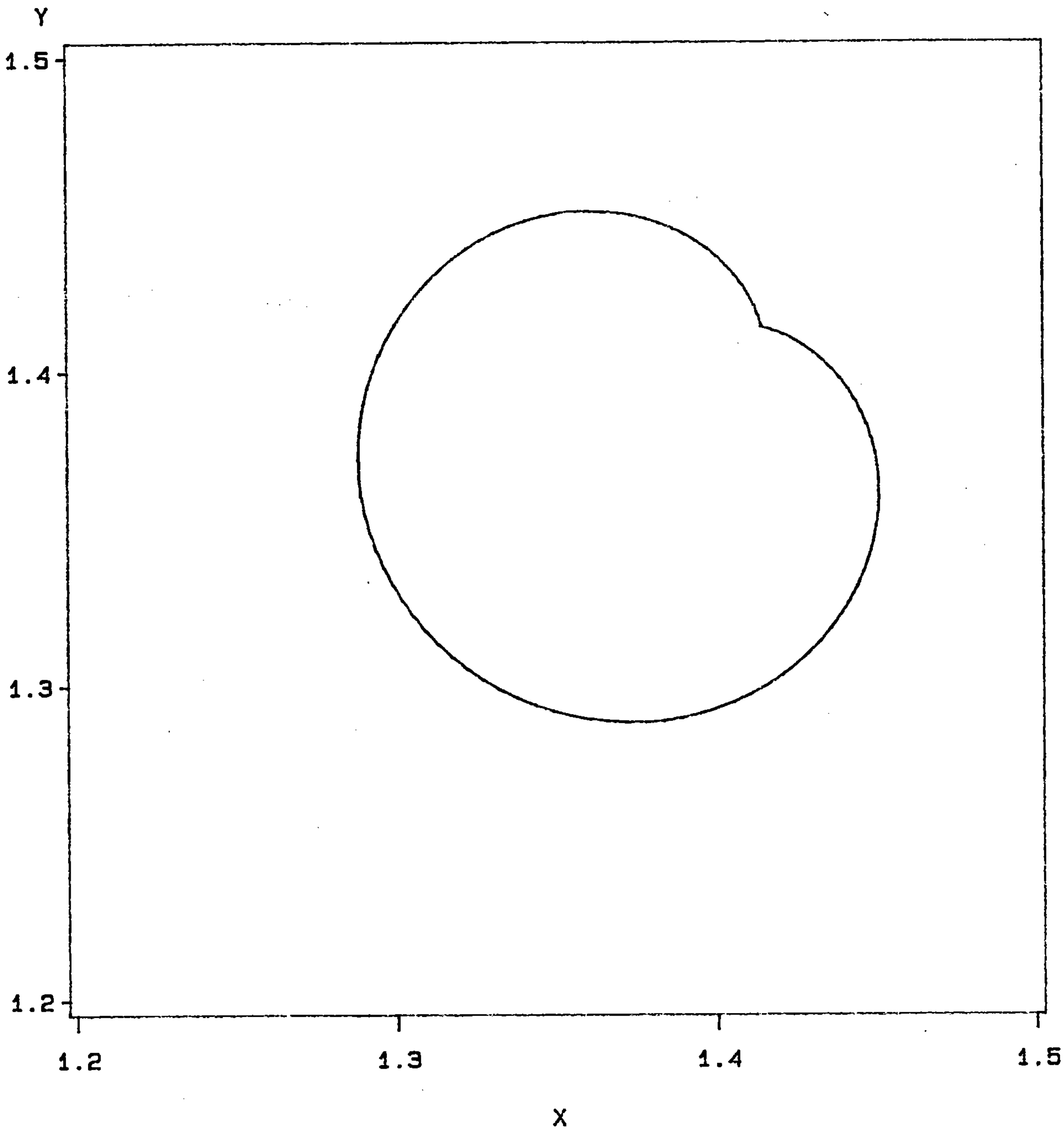


Figure 6

Set of posterior means: priors in a density band, Example 12

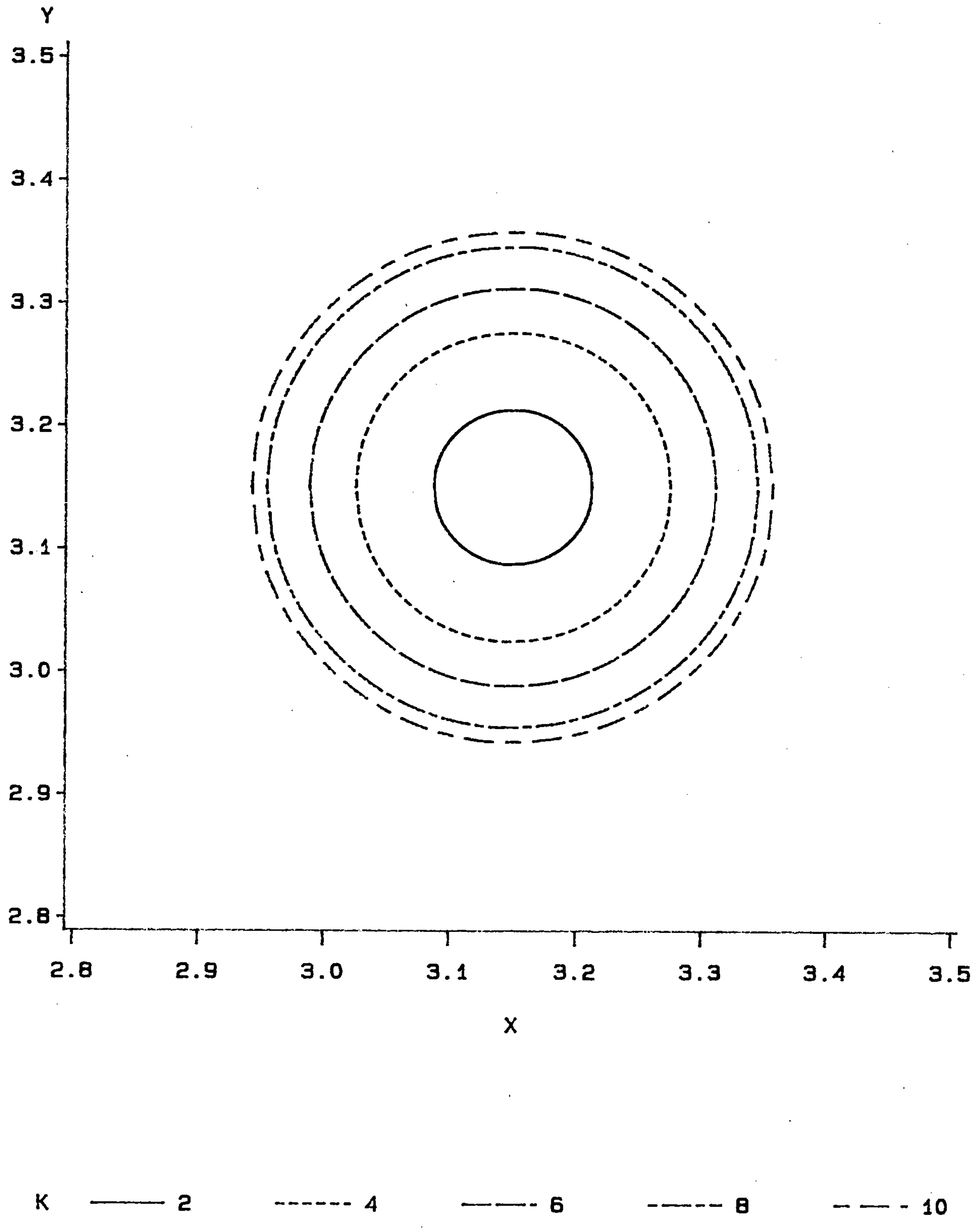
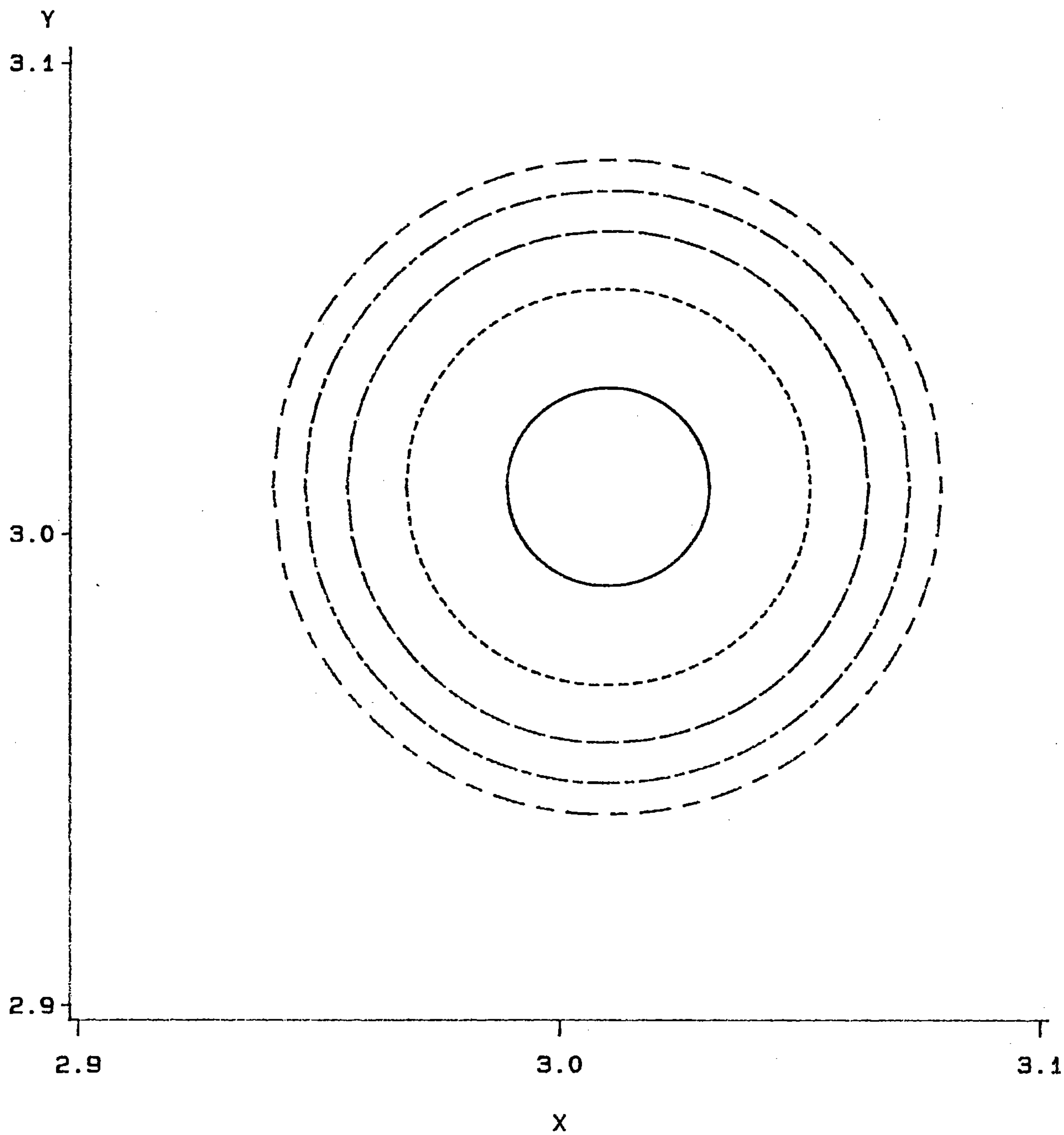


Figure 7

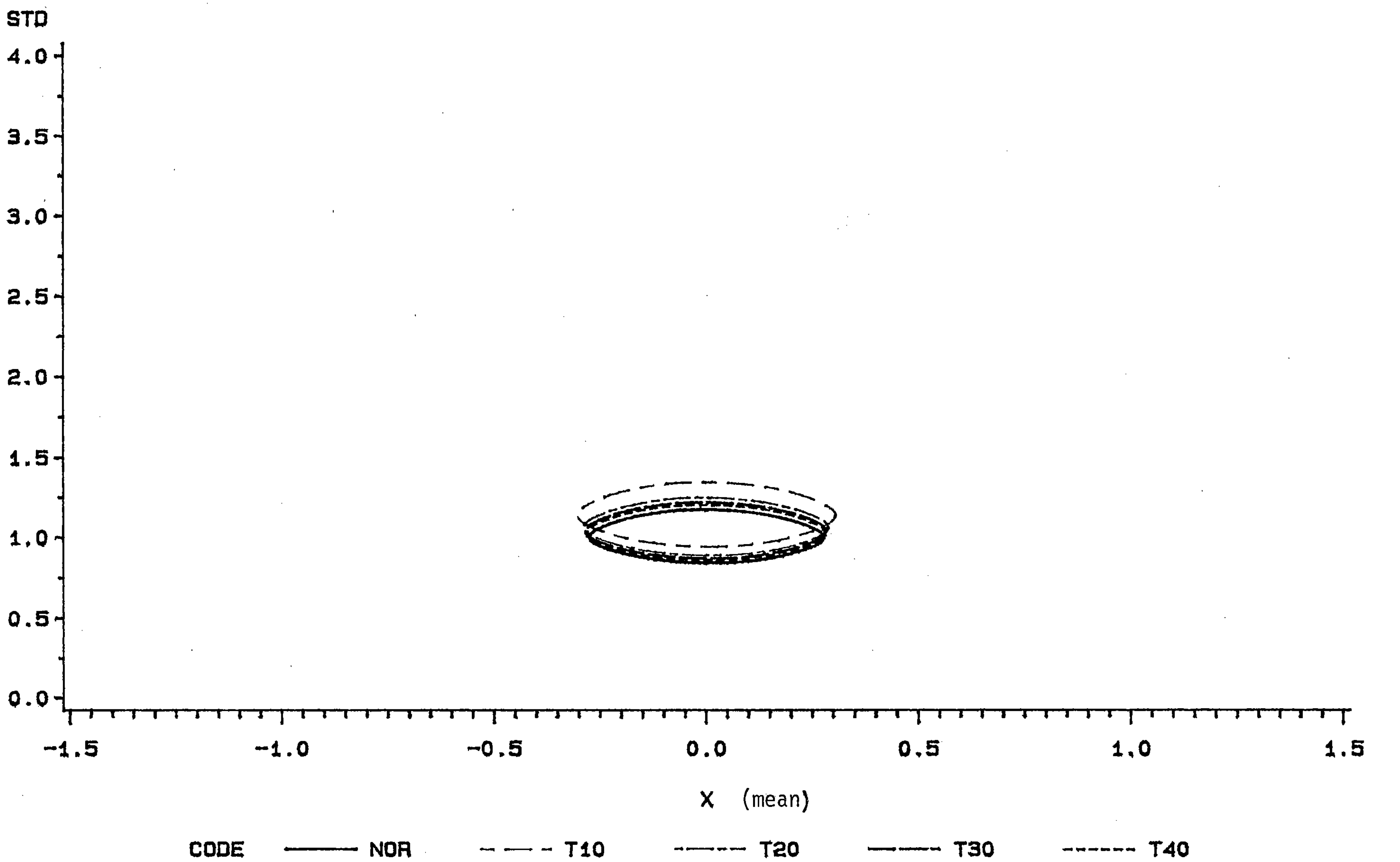
Set of posterior means: priors in a density band, Example 13



K —— 2 ------ 4 —— 6 ------ 8 — — 10

Figure 8

STANDARD DEVIATION VS. MEAN, K=2



FEB 1988

Figure 9

STANDARD DEVIATION VS. MEAN, K=4

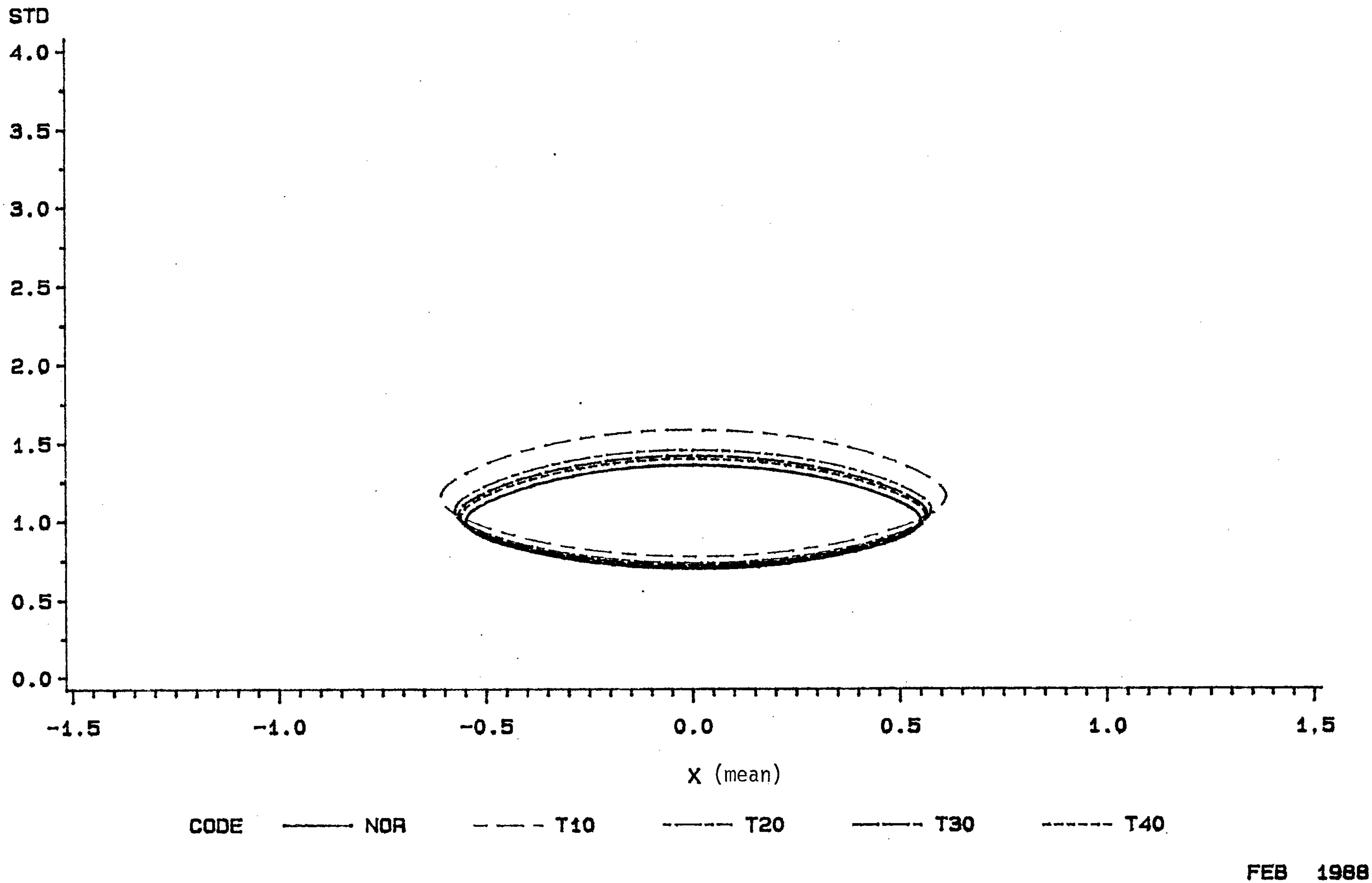


Figure 10

STANDARD DEVIATION VS. MEAN, K=6

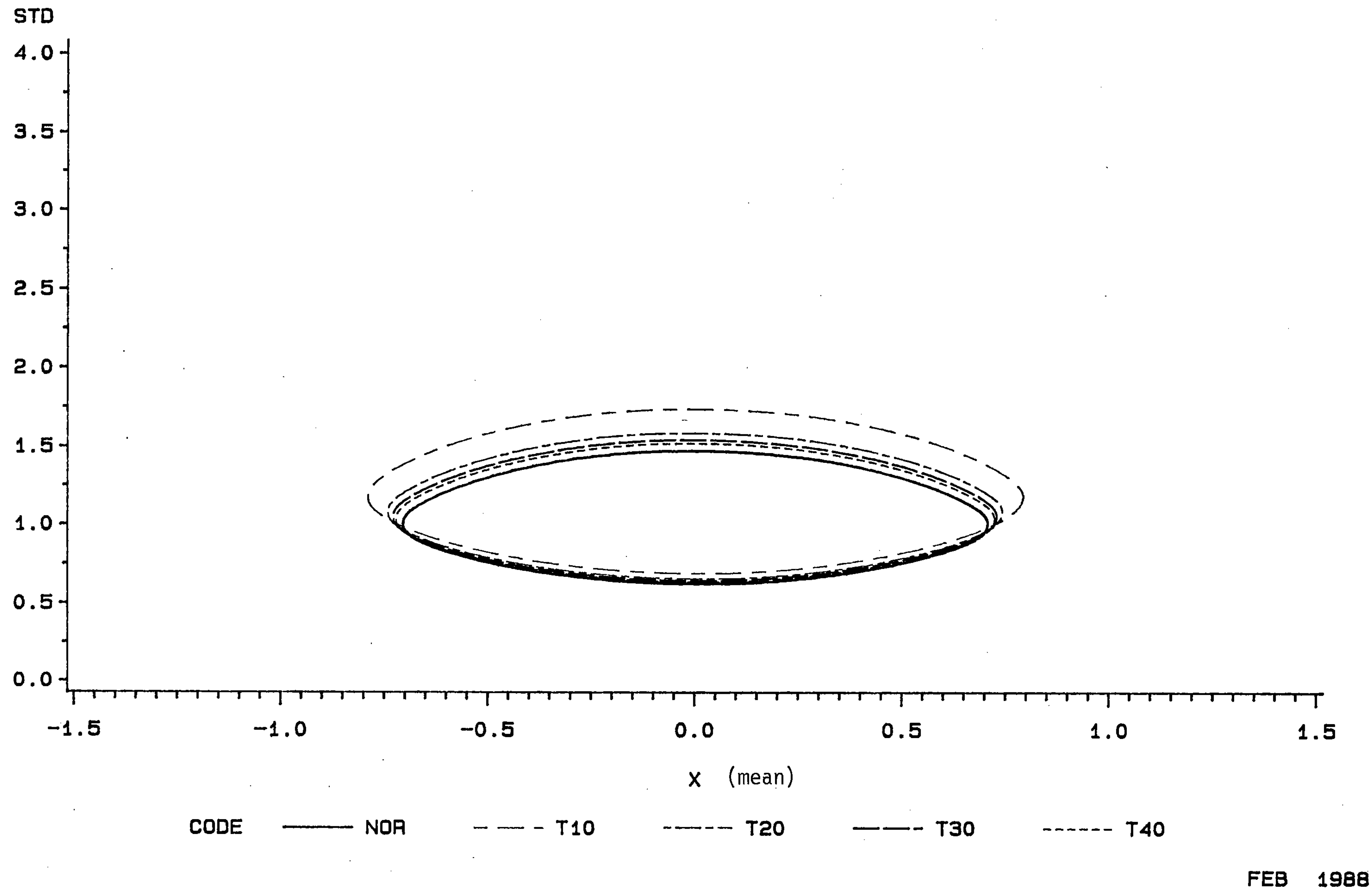
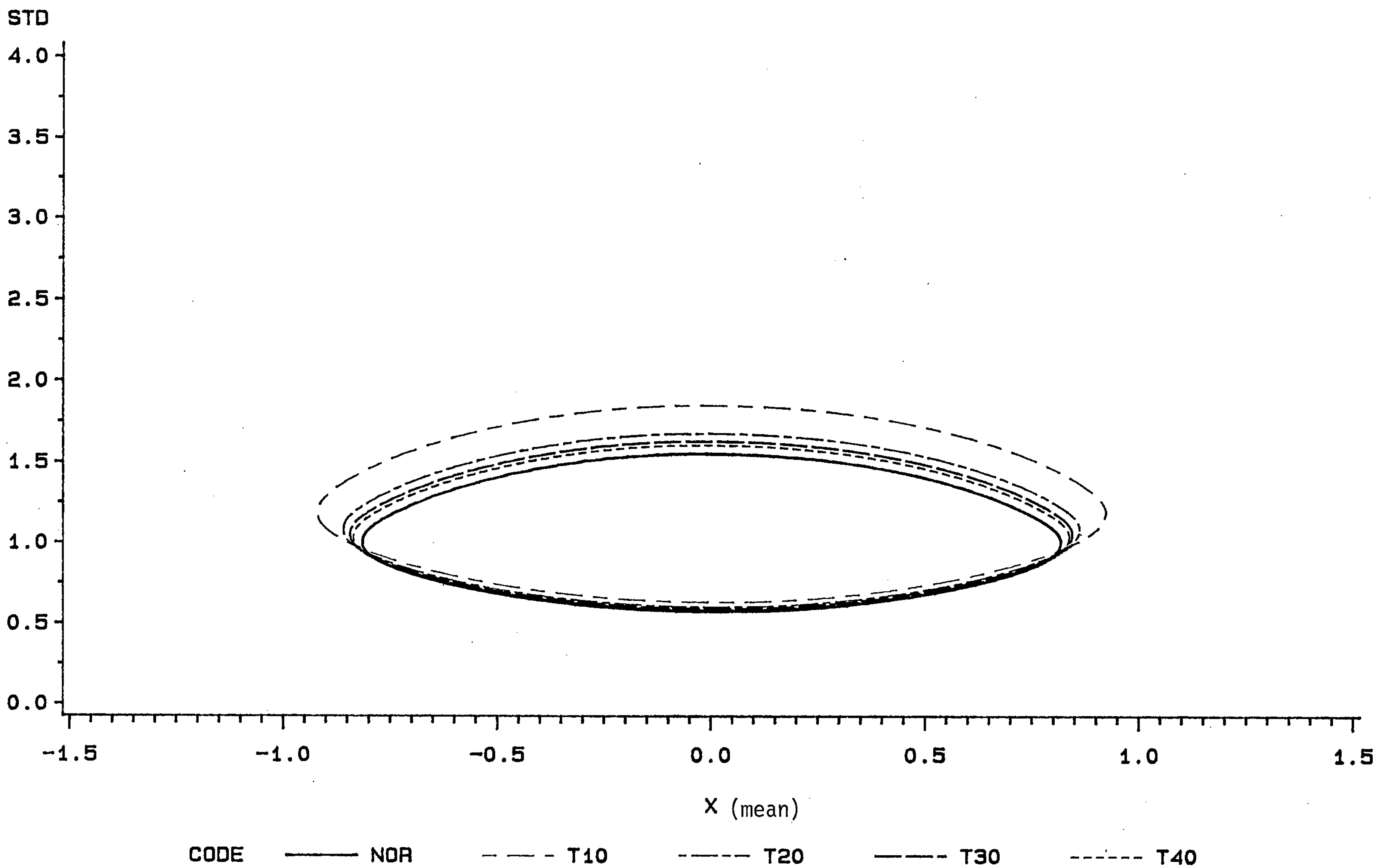


Figure 11

STANDARD DEVIATION VS. MEAN, K=8



FEB 1988

Figure 12

STANDARD DEVIATION VS. MEAN, K=10