ROBUST BAYES ESTIMATION[1]

by

James Harry Albert
Purdue University

Department of Statistics
Division of Mathematical Statistics
Mimeograph Series #79-9

June 1979

---

[1]Research in this paper was partially supported by the National Science Foundation under Grant #MCS 78-02300.

CHAPTER 1

INTRODUCTION

1. Notation

This section contains the notation that will be used and defines the relevant terms. Let $X = (X_1,\ldots,X_p)$ represent an observable vector valued random variable with values x in a sample space $\mathcal{X}$. Assume X has a probability distribution depending on a parameter θ (which may be vector valued) and there exists a density for X with respect to Lebesgue measure denoted by $f(x|\theta)$. Assume θ is unknown and let Θ represent the parameter space of all possible values of θ.

We are interested in estimating θ using an estimate d and the loss in estimating θ by d will be denoted $L(\theta,d)$. A (nonrandomized) estimator which is a function of the observation X will be denoted by $\delta(X)$, and the risk of δ for a particular value of θ is defined to be

$$R(\theta,\delta) = E_\theta^X[L(\theta,\delta(X))]$$

$$= \int_{\mathcal{X}} L(\theta,\delta(x))dF(x|\theta).$$

Superscripts in an expectation refer to the random variables over which the expectation is taken, while subscripts denote parameter values at which the expectation is taken. $F(x|\theta)$ is the cumulative distribution function of X.

If we have two estimators δ_1 and δ_2 that are both functions of X, then δ_1 is said to be better than δ_2 if $R(\theta,\delta_1) \leq R(\theta,\delta_2)$ for all $\theta \in \Theta$, with strict inequality for some value of θ. An estimator δ is called admissible if there does not exist an estimator better than δ; an estimator δ is called inadmissible if there does exist a better estimator.

A principle commonly used in choosing an estimator is the minimax principle. For an estimator δ, consider the quantity $\sup_{\theta\in\Theta} R(\theta,\delta)$. An estimator δ_1 is preferred to δ_2 by the minimax principle if $\sup_{\theta\in\Theta} R(\theta,\delta_1) < \sup_{\theta\in\Theta} R(\theta,\delta_2)$. An estimator which minimizes $\sup_{\theta\in\Theta} R(\theta,\delta)$ among all estimators with finite risk (that is, all rules δ such that $R(\theta,\delta) < \infty$ for all θ) is called a minimax rule.

We will be concerned with prior information about the parameter θ, that is, information available before X is observed. One convenient way of describing information about θ is by means of a probability distribution on Θ. We denote the corresponding prior density with respect to Lebesgue measure (assuming it exists) by $\pi(\theta)$. Given an estimator $\delta(X)$, the Bayes risk of δ is defined to be

$$r(\pi,\delta) = E^{\pi}[R(\theta,\delta)] = \int_{\Theta} R(\theta,\delta)dG(\theta),$$

where $G(\theta)$ is the cumulative distribution function of θ. If there exists an estimator δ^{π} which minimizes the Bayes risk among all estimators with finite risk, then δ^{π} is called the Bayes estimator or Bayes rule.

A Bayes estimator can typically be found from the posterior distribution of θ given x. Let $h(x,\theta)$ denote the joint density of

θ and X and let m(x) denote the marginal or unconditional distribution of X, namely

$$m(x) = \int_{\Theta} f(x|\theta)dG(\theta).$$

We will let $\pi(\theta|x)$ denote the conditional density of θ given x, defined (for $m(x) \neq 0$) by

$$\pi(\theta|x) = \frac{h(x,\theta)}{m(x)} .$$

2. Simultaneous estimation

2.1. History

First consider the situation where $X = (X_1,\ldots,X_p)$ has the p-variate normal distribution with mean vector $\theta = (\theta_1,\ldots,\theta_p)$ and identity convariance matrix (i.e. $X \sim N_p(\theta,I)$), and we wish to estimate θ using an estimator $\delta = (\delta_1,\ldots,\delta_p)$. The classical estimator is $\delta^0(X) = X$, which is the maximum likelihood estimate (MLE) and the minimum variance unbiased estimate (MVUE). For the loss function $L(\delta,\theta) = \sum_{i=1}^{p} (\delta_i-\theta_i)^2$, James and Stein (1961) showed that δ^0 is inadmissible when $p \geq 3$. They showed that the estimator

$$\hat{\delta}(X) = (I - \frac{p-2}{\sum_{i=1}^{p} X_i})X$$

has uniformly (for all θ) smaller risk than δ^0, with the largest improvement in risk occurring near the origin of the parameter space. Since the work of James and Stein, there have been many generalizations of this result. In this situation, δ^0 is minimax under quadratic loss, so finding better estimators than δ^0 is

equivalent to finding classes of minimax estimators. Large classes of minimax estimators have been found by Berger (1976), Bock (1975) and others. One of the more general results is that of Berger, et. al. (1976) where δ^0 is shown to be inadmissible when X is multivariate normal with an unknown covariance matrix.

Estimators improving upon the usual estimator (MVUE) have also been found for a variety of other distributions. Hudson (1978) considered estimation in the general continuous exponential family and found estimators that improved on the MVUE for squared error loss when three or more parameters were estimated. Berger (1978) developed a general technique for improving on standard estimators in the continuous exponential family for a variety of loss functions. In the simultaneous estimation of means from p independent Poisson distributions, Peng (1975), Zidek and Clevenson (1975) and Tsui (1978) all found estimators improving upon the MVUE for various loss functions and large enough p. Thus, for many underlying distributions of X and appropriate loss functions, the obvious estimator can be improved upon when several parameters are estimated.

Some general observations can be made about the above improved estimators. First, many of the estimators are like the James-Stein estimator $\hat{\delta}$, in that they shrink the MVUE towards a point, and the improvement in risk of the estimators is most dramatic at that point. Second, the amount of improvement over the MVUE becomes more substantial when the number of parameters estimated increases. Finally, one reason why many of these estimators improve on the MVUE

seems to be that the estimators are taking advantage of some similarities between the parameters estimated. This will be illustrated by empirical Bayes arguments in the next section.

2.2. Empirical Bayes approach

The general idea of the empirical Bayes approach (introduced by Robbins (1955)) is to use auxiliary data in constructing Bayes rules. There are two major applications of this approach. In the first, it is desired to estimate a single parameter, and data that has been observed in the past is used to help construct a prior distribution and obtain a Bayes rule. In the second application, inferences are made concerning p parameters simultaneously, and the current data is used in constructing the prior and the rule. This latter situation is referred to as the compound decision problem.

We first briefly review how empirical Bayes methods are implemented in the compound decision problem. Here $x_1,\ldots,x_n$ are observed, where $X_1,\ldots,X_n$ are independent with X_i having density $f(x_i|\theta_i)$, and it is of interest to make inferences about the group of parameters $\theta_1,\ldots,\theta_p$. One assumes that $\theta_1,\ldots,\theta_p$ come from a common unknown prior $\pi(\theta)$, and then uses the observations $x_1,\ldots,x_n$ to aid in the construction of the prior distribution. One easy way to perform this construction is to assume a particular functional form for $\pi(\theta)$, and then use the data in estimating any unknown parameters of the prior. Another method of using the data to obtain an estimator is to explicitly represent the Bayes rule in terms of the marginal distribution of X. Then $x_1,\ldots,x_n$ are used in estimating the

unknown marginal distribution and obtaining a rule (Robbins (1955)). We will only consider the first method.

As an example, the estimator proposed by James and Stein will be derived using the empirical Bayes approach (Efron and Morris (1973)). Assume $X_1,\ldots,X_p$ are independent and $X_i \sim N(\theta_i,1)$. The parameters $\theta_1,\ldots,\theta_p$ are assumed to come from a common $N(0,A)$ prior, where A is unknown. Under squared error loss, the Bayes rule for θ_i is

$$\delta_i^{\pi}(X) = (1 - \frac{1}{1+A})X_i.$$

The observations will be used to estimate A and obtain an empirical Bayes rule. First, note that marginally, $X_1,\ldots,X_p$ are independent, $X_i \sim N(0,1+A)$ and

$$\sum_{i=1}^{p} X_i^2/(1+A) \sim \chi^2(p).$$

Thus $\sum_{i=1}^{p} X_i^2/(p-2)$ can be used as an estimate of $1+A$, and after substitution, one obtains the James-Stein rule. It should be noted that the point 0, towards which the estimator shrinks X, is not special. If one assumes that the common prior for $\theta_1,\ldots,\theta_p$ is $N(\mu,A)$, and also that μ is known to the user, then the corresponding empirical Bayes rule is

$$\hat{\delta}(X) = \mu + (1 - \frac{p-2}{\sum_{i=1}^{p}(X_i-\mu)^2})(X-\mu).$$

It is of interest to study the behavior of an estimator when errors are made in the specification of the prior. Here, for example,

the sensitivity of $\hat{\delta}$ to the assumption that $N(\mu,A)$ is the common prior is of concern. If the true prior mean of θ_j, say, is much larger than μ, then X_j would likely be far from μ. In this case, the shrinking term $(p-2)/\sum_{i=1}^{p}(X_i-\mu)^2$ would be small and $\hat{\delta}(X) \tilde{=} X$. Thus $\hat{\delta}$ tends to ignore the prior information in the presence of an extreme observation. It will be shown that this behavior is an indication that $\hat{\delta}$ is robust with respect to uncertainty in the prior specification.

Many of the estimators that improve upon the usual estimator in simultaneous estimation may be derived by means of empirical Bayes arguments. These estimators perform much like the optimal Bayes rules, and they often are quite robust with respect to misspecification of prior information. The estimators that are discussed in this paper are closely related to empirical Bayes estimators.

2.3. Necessity of inputting prior information

We are interested in finding attractive alternatives to the usual estimator (the MVUE) in simultaneous estimation problems. The usual estimator is typically minimax, and so we cannot expect to find an estimator which substantially improves upon it (with respect to risk) over the entire parameter space. The James-Stein estimator shows the greatest improvement over the MVUE near the origin; likewise most alternative estimators show substantial improvement in only a particular region of the parameter space. Therefore if a user wants to find an estimator which is better than the usual one for his problem, he should specify a region in

which he would like the substantial improvement to occur. In other words, the input of prior information seems necessary in the development of good alternative estimators. If the user has virtually no prior information concerning the parameters to be estimated, then he may as well use the usual estimator, since any improved estimator will be unlikely to show much improvement at the true value of the parameter. In conclusion, to find an attractive alternative estimator to the MVUE, it seems necessary to be a Bayesian, and most of our work will be presented from a Bayesian point of view. This rationale for inputting prior information in improved estimators was discussed in detail by Berger (1977).

3. Robust Bayes estimation

3.1. Introduction

The subject of robustness of Bayes rules has received only sporadic attention in the literature. In James Berger's book Statistical Decision Theory, a section is devoted to this topic and many of the ideas that we will discuss can be found in that book.

In a Bayes decision problem, there are three main elements, the prior, $\pi(\theta)$, the sample density, $f(x|\theta)$, and the loss function, $L(\theta,\delta)$. The robustness of a Bayes estimator refers to the sensitivity of the estimator to the assumptions in the model about which there exists uncertainty. For example, if the form of the loss function for large errors is uncertain, then we would like the Bayes estimator to be insensitive to the selection of L. That is, if a new loss function is equivalent to the original one except for its specification for

large errors, it would not be desirable for the Bayes estimator to change significantly.

To completely discuss the robustness of a Bayes estimator, one should investigate the sensitivity of the estimator to the sample density, the loss, and the prior distribution. Sensitivity of any type of estimator (Bayesian or classical) to the density is an important topic and many authors, including Huber (1977) have developed robust classical estimators that are good for estimating parameters in distributions within a certain class. We will not discuss this type of robustness here, and refer the reader to Huber's book for a good discussion of the subject. Sensitivity of the estimator to the loss is important and any estimator should be evaluated with respect to different losses that seem appropriate for a particular estimation problem. The performance of our suggested rules with respect to different losses will be discussed, but a thorough investigation of this type of robustness will not be made. Our major concern is how sensitive the Bayes estimator is to uncertainty in the specification of the prior, and we discuss this topic in the next section.

3.2. Robustness with respect to the prior

We would like to investigate the sensitivity of the Bayes estimator to the prior distribution. Ideally, the prior distribution that is used by a statistician is accurately determined from past observations and subjective knowledge. Unfortunately, this is rarely the case, and the particular prior distribution used can never be more

than an approximation to the true prior for the problem. For example, the statistician may only be able to specify fractiles of the unknown prior distribution or a region which he thinks contains 90% of the distribution. In such cases he will be uncertain about the tail or extreme parts of the prior distribution. We are generally concerned about the sensitivity of the Bayes estimator to prior specifications that are uncertain, in this example, knowledge of the tail of the prior.

It will be shown in particular situations that Bayes estimators, especially those developed through conjugate priors, can be sensitive to uncertain parts of the prior specification. We want to develop Bayesian procedures which can incorporate prior knowledge, but are safe with respect to errors in the specification of prior knowledge. For example, if the tail of the prior distribution is uncertain, then the Bayes estimator should not perform much worse than the MVUE when errors are made in the specification of the tail of the prior. We next will discuss ways of measuring the robustness of estimators.

3.2.1. Posterior robustness

One method of analyzing the robustness of a particular Bayes estimator is to see how the estimator changes as we change the prior distribution. For example, it may be known that the prior distribution has particular fractiles but little else may be known about the distribution. Then one could consider the class of prior distributions with that set of fractiles, and see how the Bayes estimator changes within that class. If the Bayes estimator does not change significantly,

then the estimator is robust or insensitive to the prior information that is uncertain. This method of detecting robustness is probably the most natural from a Bayesian viewpoint, since one is investigating the posterior distribution and the Bayes decision directly.

Such an investigation of rules with regard to posterior robustness has been made by Edwards, Lindeman and Savage (1963). They consider the situation in which a random sample $X_1,\ldots,X_n$ has been taken from a distribution $f(x|\theta)$, and the likelihood, $\prod_{i=1}^{n} f(x_i|\theta)$, considered as a function of θ, is very concentrated or peaked about some value. A typical prior density will look flat in this region where the likelihood is most concentrated and, under suitable conditions, the authors show the posterior density may be approximated by the likelihood (suitably normalized to be a density for θ). In this situation, the Bayes rule will be a function primarily of the n observations and will essentially ignore the prior information. Thus for a wide range of priors, the Bayes decision will be the same, and this rule is very robust with respect to the prior distribution chosen. In other words, if the prior satisfies some mild conditions, the data will dominate the prior information when enough observations are taken. The authors refer to this situation as the principle of stable estimation and it displays one type of posterior robustness. Although this situation is of interest, we are primarily interested in posterior robustness when the data does not dominate the prior and the prior information is significant. Recall that we wish to use prior information in simultaneously estimating p parameters, and it is important for the Bayes estimator to use prior information if

significant improvement over the MVUE is desired.

3.2.2. Risk robustness

The second method of investigating robustness is to calculate the risk $R(\theta,\delta)$ of the Bayes estimator and observe its behavior over the parameter space. Since one is ultimately concerned with the Bayes estimator being a good alternative to the MVUE, the risk of the Bayes estimator is usually compared with the risk of the MVUE.

First, it is of interest to analyze the risk of the Bayes estimator when the prior information is correct. For example, if a single parameter is estimated and a prior mean and prior variance are inputted, then,if the prior information is correct, the parameter will lie within a few standard deviations of the prior mean. In this "prior region" of the parameter space the Bayes estimator is expected to have risk much smaller than the risk of the MVUE. This improvement in risk of the Bayes estimator over the MVUE would tell us that the Bayes estimator is making significant use of the given prior information. The user thus knows that if he specifies the prior mean and variance correctly, then,in repeated use of this estimator, his errors will be of a smaller magnitude than when the MVUE is used. Recall that our rationale for using prior information was to produce an estimator that improved upon the MVUE in a region of the parameter space. The inspection of the two risk functions in the prior region is the best way of checking this improvement.

Once a Bayes estimator has been evaluated in the prior region, one is interested in the behavior of its risk function outside of the

prior region. A user may be certain that 90% of the prior distribution occurs in a particular interval, but be unsure about the remaining 10%. If the prior has a fat tail, then values of the parameter that are far away from the prior region are possible. Also mistakes may be made in specifying fractiles of the prior, and the parameter may lie with high probability outside of the prior region. In either case one is interested in comparing the risk functions of the Bayes estimator and the MVUE outside of the region where the Bayes estimator shows improvement. If the Bayes estimator exhibits much higher risk than the MVUE in such a region, then the user must be concerned with the possibility of parameter values in that region and encountering errors of a large magnitude. If the risk of the Bayes estimator is approximately equal to the risk of the MVUE outside of the prior region, then in effect the estimator is ignoring the wrong prior information in this "extreme" region. The statistician using a robust Bayes estimator should feel safe in applying his prior knowledge, in that drastically wrong prior information will cause him to incur errors not much larger than the errors in using the MVUE. It will be shown that conjugate Bayes estimators often are very sensitive to misspecification of prior information, and this is indicated by a very large risk (compared to the risk of the MVUE) outside of the prior region. Finally from a classical point of view, comparing the risk functions of the Bayes estimator and MVUE over the entire parameter space allows one to evaluate how good the Bayes estimator is as an alternative to the MVUE.

3.2.3. Bayes risk robustness

The Bayes risk is minimized by the Bayes estimator under a particular model, and it is natural to consider Bayes risk as an appropriate measure of robustness. One knows that the Bayes risk of the Bayes estimator will be significantly smaller than the Bayes risk of the MVUE when the prior information has been specified correctly. But we are concerned with the sensitivity of the Bayes risk of the Bayes estimator to changes in the uncertain portion of the prior specification. For example, if uncertainty exists in the tail of the prior, then a robust Bayes estimator would have significantly smaller Bayes risk than the Bayes risk of the MVUE for priors which differ from the specified prior only in the tail. Most conjugate Bayes estimators are again not robust in this situation. Indeed it is possible for a conjugate Bayes estimator to have infinite Bayes risk under a true prior when the true prior differs from the specified prior only in the tail (see Berger (1979) - Chapter 4).

A convenient way of indicating "Bayes risk robustness" is to calculate the Bayes risk of the Bayes estimator under deviations from the specified prior information, and compare this with the Bayes risk of the MVUE. For "small" deviations from the prior model, we would hope that the Bayes risk of the Bayes estimator is still smaller than the Bayes risk of the MVUE. For large errors in prior specification, we hope that the Bayes risks of the two estimators would be about the same. The robust estimator we find may not be the "optimal" Bayes estimator under a particular prior, that is the

estimator may not minimize the Bayes risk among all rules with finite risk, but the estimator should have small Bayes risk under all priors which model the prior information well.

To perform a formal analysis of robustness with respect to Bayes risk, the Γ-minimax approach has frequently been used. In this approach, introduced by Robbins (1964), a class Γ of prior distributions is specified which contains all the possible priors for the given problem. For example, the first two moments of the prior distribution may be known with accuracy ϵ, but little else may be known concerning the prior. The set of possible priors is then

$$\Gamma = \{\pi: |\mu_i - \mu_i^0| < \epsilon,\ i = 1,2 \quad \text{where } \mu_i = E^{\pi}(\theta^i)\}.$$

The Γ-minimax approach evaluates an estimator by its worst Bayes risk in the class Γ, called its Γ-minimax risk. In a choice between two estimators, the estimator with the smaller Γ-minimax risk is preferred. The Γ-minimax risk is a reasonable measure of robustness when a class Γ is chosen which describes well the priors which model given prior information.

The major difficulty in using the Γ-minimax approach is that optimal rules are difficult to find for "good" classes of prior distributions. Usually prior information consists of probabilities attached to certain regions of the parameter space, and the class Γ should consist of priors which assign similar probabilities to the same regions. For example, a set of fractiles may be specified, and Γ could consist of all priors which have fractiles close to the given set.

These classes are very difficult to work with and there has been very little work in developing rules optimal with respect to such classes. There has been work deriving optimal rules when Γ consists of priors which have a set of moments close to a specified set (for one example, see Jackson, Donovan, Zimmer and Deely (1970)). The problem with using classes with such moment specifications is that prior moments are often not known to the user and the classes generally donot allow enough flexibility in certain parts of the prior information such as the tail of the prior. For a discussion of common types of prior information, the reader is referred to Chapter 3 of Berger (1979).

3.3. History

3.3.1. Normal estimation

Previous work in robust Bayes estimation has, for the most part, dealt with the multivariate normal mean. There has been a concern to develop a prior that realistically reflects the usual type of prior information. One prior that is often used is the conjugate normal prior, and Dawid (1973) and Anscombe (1963) argue that this is not a good prior for common types of prior information. In particular, this prior implies strong knowledge concerning the distance of the unknown parameter from its mean - it has a sharp tail or a tail which rapidly decreases as the absolute value of the difference between the parameter and its mean gets large. In this problem it is common to have significant information about the central portion of the prior but weak or vague information about the tail portion. The

normal prior can model the central portion well, but has tails much too sharp to reflect vague prior information in the tail.

The unsuitability of the normal prior can be seen by noting the behavior of the corresponding Bayes estimator when extreme data is observed. If one has vague prior knowledge in the tail, then extreme values of X are possible, and the Bayes estimator should be insensitive to this extreme data. But note that if $X \sim N(\theta,\sigma^2)$ and the prior for θ is $N(\mu,\tau^2)$, then the Bayes estimator under squared error loss is

$$\delta^B(X) = X - \frac{\sigma^2}{\sigma^2+\tau^2}(X-\mu).$$

Clearly when X is far from the prior mean μ, then δ^B is far from the MVUE X, which indicates that δ^B is not robust with respect to the prior tail.

Dissatisfaction with the behavior of the normal prior has led to discussion concerning the types of priors that should be used. Hill (1974) states that, in the situation where weak information exists about the prior tail, a prior which acts like a flat or uniform prior in the tails is appropriate. When extreme data is observed, the Bayes estimator will then be approximately the usual MVUE. For example, a t density has flatter tails than a normal density, so using a t prior would lead to an estimator which is more robust with respect to misspecification in the tail. Many authors have used t priors in the analysis of normal means, among them Anscombe (1963), Stein (1962), Hill (1969) and Dickey (1974). Rubin (1977) investigates the effect of choosing the wrong prior in

the normal problem, and finds that priors with "sufficiently flat" tails lead to Bayes procedures which are robust with respect to the prior tail. In the next section, we will discuss an explicit robust Bayes estimator of a multivariate normal mean which was developed using a flat tailed prior (Berger (1977a)).

3.3.2. A robust Bayes estimator of a multivariate normal mean

We summarize here the robust Bayesian analysis in Berger (1977a) concerning the estimation of a multivariate normal mean. The estimator developed in that paper is similar to the robust Bayesian estimators that will be analyzed in this paper. Let $X \sim N(\theta,\Sigma)$, where Σ is a known covariance matrix, and assume it is desired to estimate θ using a quadratic loss. Berger developed a generalized Bayes estimator for this situation which incorporates prior information in the form of a mean vector μ and a covariance matrix A. The prior that was used has extremely flat tails, so that the resulting estimator is very robust. The estimator which Berger found can be written as

$$\delta^*(X) = \mu+(I - \frac{r(||X-\mu||^2)}{||X-\mu||^2}\Sigma(\Sigma+A)^{-1})(X-\mu),$$

where $||X-\mu||^2 = (X-\mu)^t(\Sigma+A)^{-1}(X-\mu)$ and $r(\cdot) \leq (p-2)$ can be expressed in a closed form. When p, the dimension of the problem, is large and the prior information is correct, it can be shown that δ^* will perform like the Bayes estimator using the conjugate normal prior. But when the prior information has been misspecified, δ^* is much more robust than the conjugate Bayes estimator. One indication of this robustness

is that when one observation X_i deviates greatly from its corresponding mean μ_i, one can show that $r(||X-\mu||^2)/||X-\mu||^2$ becomes very small and $\delta^*(X) \cong X$. Thus bad or extreme observations tend to discredit the prior information. Berger showed δ^* to be very robust with respect to alternative priors when some parts of the prior information are incorrect. He also gave conditions for δ^* to have uniformly smaller risk than the risk of the MVUE δ^0, so that in many situations, especially those in which there exists some type of symmetry among the coordinates $\theta_1,\ldots,\theta_p$, δ^* is uniformly superior to δ^0 in terms of risk.

Using a normal approximation to the posterior distribution of θ, Berger derives a confidence ellipsoid for θ that is centered about δ^*. This ellipsoid is an improvement over the classical confidence ellipsoid both in terms of probability of coverage and size. Finally the above results are generalized to the case where $\Sigma = \sigma^2\Sigma_0$, Σ_0 a known matrix and σ^2 an unknown constant, and a generalized Bayes estimator similar to δ^* is found.

3.4. Summary

Generally, a robust Bayes estimator should be able to incorporate prior information and be safe when the prior information is wrong. This type of estimator is most useful when incomplete prior information is available. It should offer significant improvement over the MVUE when the prior information is correct, but lose very little compared to the MVUE when the prior information is misspecified. It can be thought of as a conservative application of the Bayesian method and it

may not be appropriate to use when very strong prior information does exist. Finally, a robust Bayes estimator should exhibit a risk $R(\theta,\delta)$ much smaller than that of the MVUE in the "prior region" of the parameter space and should not have a much larger risk elsewhere. Such an estimator will be an attractive alternative to the MVUE in simultaneous estimation.

4. Introduction to the work in this paper

In this paper two estimation problems are considered: the simultaneous estimation of means from independent Poisson distributions and the estimation of multinomial proportions. Robust Bayes estimators and associated confidence regions will be developed for both problems. In this section, the two problems are defined and the recommended estimation procedures are summarized.

In Chapter 2, simultaneous estimation of Poisson means is considered. Assume that $X_1,\ldots,X_p$ are independent and X_i is distributed Poisson with mean λ_i, $i = 1,\ldots,p$. It is desired to estimate $\lambda = (\lambda_1,\ldots,\lambda_p)$ using an estimator $\delta = (\delta_1,\ldots,\delta_p)$ and the loss $L_1(\delta,\lambda) = \sum_{i=1}^{p} (\delta_i-\lambda_i)^2$ will usually be considered. The usual estimator of λ is $\delta^0(X) = X$, which is the MLE and MVUE. The robust Bayes estimator that is developed incorporates a prior mean μ_i and a prior variance $\mu_i\beta_i$ for the component λ_i, $i = 1,\ldots,p$. This estimator is defined componentwise as

$$\delta_i^*(X)=\mu_i+\left(1 - \frac{1}{(\beta_i+1)} \min\left\{1,\frac{\sum_{j=1}^{p} X_j/(\beta_j+1)}{\sum_{j=1}^{p} X_j/(\beta_j+1)^2+ \sum_{j=1}^{p} ((X_j-\mu_j)/(\beta_j+1))^2}\right\}\right)(X_i-\mu_i),$$

$$i=1,\ldots,p.$$

It is shown in Chapter 2 that δ^* is an attractive alternative to δ^0 when prior information is available.

Chapter 3 considers confidence regions for the Poisson parameter λ. A classical confidence rectangle for λ is defined by

$$C^0(X) = \{\lambda: |X_i+z^2_{\alpha/2}/2-\lambda_i| \leq z_{\alpha/2}(X_i+z^2_{\alpha/2}/4)^{1/2},\ i = 1,\ldots,p\},$$

where z_α is the $(1-\alpha)100$th percentile of the standard normal distribution. The recommended robust Bayes confidence rectangle, based on δ^*, is defined by

$$C^*(X) = \{\lambda: |\delta^*_i(X)+(1-c^*_i(X))(z^2_{\alpha/2}-1)/3-\lambda_i| \leq z_{\alpha/2}(X_i+z^2_{\alpha/2}/4)^{1/2}(1-c^*_i(X))^{1/2},\ i = 1,\ldots,p\},$$

where

$$c^*_i(X) = \frac{1}{\beta_i+1}\min\{1,\ \frac{\sum_{j=1}^{p} X_j/(\beta_j+1)}{\sum_{j=1}^{p} X_j/(\beta_j+1)^2+\sum_{j=1}^{p}((X_j-\mu_j)/(\beta_j+1))^2}\}.$$

In Chapter 3, C^* is shown to be an attractive alternative to C^0 with respect to probability of coverage and size.

The multinomial estimation problem is considered in Chapter 4. Assume $X = (X_1,\ldots,X_p)$ has the multinomial distribution with parameters N and $\theta = (\theta_1,\ldots,\theta_p)$, that is, X has the density

$$f(x|\theta) = \binom{N}{x_1\ldots x_p}\prod_{i=1}^{p}\theta_i^{x_i},\quad x_i = 0,1,2,\ldots \text{ for all } i,\ \sum_{i=1}^{p} x_i = N,$$

where $0 \leq \theta_i \leq 1$ for all i, and $\sum_{i=1}^{p} \theta_i = 1$. Assume that N is known and it is desired to estimate θ using an estimator $\delta=(\delta_1,\ldots,\delta_p)$. The loss function that is considered is squared error, that is,

$$L_1(\theta,\delta) = N \sum_{i=1}^{p} (\delta_i-\theta_i)^2.$$

The classical estimator of θ is $\delta^0(X) = X/N$, the MLE and the MVUE. The two robust Bayes estimators that are developed incorporate a prior mean γ_i for the component θ_i, i = 1,...,p, and a parameter K, which reflects the accuracy of the prior means $\gamma_1,\ldots,\gamma_p$. The two estimators are defined componentwise as

$$\delta_i^*(X) = \gamma_i+(1-\min\{\frac{K}{N+K}, \frac{1-\sum_{j=1}^{p} \hat{\theta}_j^2}{1-\sum_{j=1}^{p} \hat{\theta}_j^2+N\sum_{j=1}^{p}(\hat{\theta}_j-\gamma_j)^2}\})(\hat{\theta}_i-\gamma_i),$$

and

$$\tilde{\delta}_i(X) = \gamma_i+(1-\min\{\frac{K}{N+K}, \frac{p-1}{N\sum_{j=1}^{p} \gamma_j^{-1}(\hat{\theta}_j-\gamma_j)^2}\})(\hat{\theta}_i-\gamma_i),$$

i = 1,...,p, where $\hat{\theta}_i = X_i/N$, i = 1,...,p. The estimators δ^* and $\tilde{\delta}$ are both shown in Chapter 4 to be attractive alternatives to δ^0.

Chapter 4 additionally considers confidence regions for the multinomial parameter θ. A classical confidence rectangle for θ is defined by

$$C^0(X) = \{\theta: |\hat{\theta}_j-\theta_j| \leq z_{\alpha/2}[\hat{\theta}_j(1-\hat{\theta}_j)/N]^{1/2}, \quad j \neq k,$$

$$|\hat{\theta}_k-\theta_k| \leq z_{\alpha/2} \sum_{j\neq k} [\hat{\theta}_j(1-\hat{\theta}_j)/N]^{1/2}, \text{ where } \hat{\theta}_k(1-\hat{\theta}_k)=\min_j\{\hat{\theta}_j(1-\hat{\theta}_j)\}\}.$$

Two recommended robust Bayes confidence rectangles, based respectively on the estimators δ^* and $\tilde{\delta}$, are defined by

$$C^*(X) = \{\theta: |\delta_j^*(X)-\theta_j| \le z_{\alpha/2}(1-c^*(X))^{1/2}[\hat{\theta}_j(1-\hat{\theta}_j)/N]^{1/2},\ j\neq k,$$

$$|\delta_k^*-\theta_k| \le z_{\alpha/2}(1-c^*(X))^{1/2}\sum_{j\neq k}[\hat{\theta}_j(1-\hat{\theta}_j)/N]^{1/2},\ \text{where}$$

$$\hat{\theta}_k(1-\hat{\theta}_k) = \min_j\{\hat{\theta}_j(1-\hat{\theta}_j)\}\},$$

and

$$\tilde{C}(X) = \{\theta: |\tilde{\delta}_j(X)-\theta_j| \le z_{\alpha/2}(1-\tilde{c}(X))^{1/2}[\hat{\theta}_j(1-\hat{\theta}_j/N]^{1/2},\ j\neq k,$$

$$|\tilde{\delta}_k-\theta_k| \le z_{\alpha/2}(1-\tilde{c}(X))^{1/2}\sum_{j\neq k}[\hat{\theta}_j(1-\hat{\theta}_j)/N]^{1/2},\ \text{where}$$

$$\hat{\theta}_k(1-\hat{\theta}_k) = \min_j\{\hat{\theta}_j(1-\hat{\theta}_j)\}\},$$

where

$$c_i^*(X) = \min\{\frac{K}{N+K},\ \frac{1-\sum_{j=1}^{p}\hat{\theta}_j^2}{1-\sum_{j=1}^{p}\hat{\theta}_j^2+N\sum_{j=1}^{p}(\hat{\theta}_j-\gamma_j)^2}\},$$

and

$$\tilde{c}_i(X) = \min\{\frac{K}{N+K},\ \frac{p-1}{N\sum_{j=1}^{p}\gamma_j^{-1}(\hat{\theta}_j-\gamma_j)^2}\}.$$

In Chapter 4, C^* and $\tilde{C}$ are shown to be attractive alternatives to C^0 when prior information is available.

CHAPTER 2

A ROBUST BAYES ESTIMATOR OF p POISSON MEANS

1. Introduction

1.1. History

As in the multivariate normal problem, the MVUE $\delta^0(X) = X$ has been found inadmissible in the simultaneous estimation of p Poisson means. Peng (1975) and Hudson (1978) considered this problem under the loss $L_1(\delta,\lambda) = \sum_{i=1}^{p} (\delta_i-\lambda_i)^2$ and each found estimators which possess uniformly smaller risk than δ^0 for $p \geq 3$. Peng's estimator, discussed in Section 3.2.2, is defined componentwise as

$$\delta_i^P(X) = X_i - \frac{(p-N_0-2)_+}{S} \sum_{j=1}^{X_i} j^{-1}, \quad i = 1,\ldots,p,$$

where

$$N_0 = \text{number of } \{X_j: X_j = 0\},$$

$$S = \sum_{i=1}^{p} \left(\sum_{j=1}^{X_i} j^{-1} \right)^2,$$

and

$$(a)_+ = \max\{0,a\}.$$

The estimators of Peng and Hudson shrink X towards the origin and therefore show most of their improvement over δ^0 near the origin. Tsui (1978) extended their results in

two ways. He found an estimator which uniformly improves upon δ^0 and shrinks X towards a positive integer K. This estimator is defined componentwise as

$$\delta_i^T(X) = X_i - \frac{(p - \sum_{n=0}^{K} N_n - 2)_+}{S^*} b_{X_i}, \quad i = 1,\ldots,p,$$

where N_n = number of $\{X_j: X_j = n\}$ and

$$b_j = 1 + \sum_{n=K+2}^{j} n^{-1} \quad \text{if } j \geq K+2$$

$$= 1 \quad \text{if } j = K+1$$

$$= 0 \quad \text{if } j = K$$

$$= -\mu' \quad \text{if } 0 \leq j < K \text{ and } K > 0,$$

$$\mu' > 0, \quad S^* = \sum_{i=1}^{p} b_{X_i}^2 .$$

The estimator δ^T will be discussed in Section 3.2.2. Tsui also found a similar estimator which shifts X towards a point determined by the data and uniformly improves upon δ^0 for $p \geq 4$.

Using the loss function $L_2(\delta,\lambda) = \sum_{i=1}^{p} \lambda_i^{-1}(\delta_i - \lambda_i)^2$, Clevenson and Zidek (1975) developed an estimator better than δ^0 for $p \geq 2$. Their estimator, also discussed in Section 3.2.2, is defined componentwise as

$$\delta_i^Z(X) = \left(1 - \frac{\gamma+p-1}{\sum_{j=1}^{p} X_j + \gamma + p - 1}\right) X_i, \quad i = 1,\ldots,p,$$

where $1 \leq \gamma \leq p-1$. Clevenson and Zidek showed δ^Z to be admissible and

generalized Bayes and gave it a Bayesian interpretation. Note that δ^Z shrinks X towards the origin and the authors recommend their estimator for use when the loss L_2 is appropriate and $\lambda_1,\ldots,\lambda_p$ are suspected small. Tsui and Press (1978a) generalized Clevenson and Zidek's result for the loss $\sum_{i=1}^{p} \lambda_i^{-k}(\delta_i-\lambda_i)^2$ and found a large class of estimators which have uniformly smaller risk than δ^0 for $p \geq 2$.

For the most part, the above work assumes that one observation has been taken from each Poisson population; Tsui and Press consider the case where n_i observations have been taken from the Poisson population with mean λ_i. It will be shown in Section 5.1 that this situation motivates the consideration of the weighted loss function $\sum_{i=1}^{p} c_i\lambda_i^{-k}(\delta_i-\lambda_i)^2$. They find better estimators than δ^0 under this loss.

We now summarize the Bayesian work that has been done in the Poisson estimation problem. The conjugate prior density for $\lambda_1,\ldots,\lambda_p$ is

$$g(\lambda_1,\ldots,\lambda_p) = \prod_{i=1}^{p} \frac{e^{-\lambda_i/\beta_i}\lambda_i^{\alpha_i-1}}{\beta_i^{\alpha_i}\Gamma(\alpha_i)}, \quad \lambda_1,\ldots,\lambda_p > 0, \quad \alpha_i,\beta_i > 0, \quad i = 1,\ldots,p.$$

That is, $\lambda_1,\ldots,\lambda_p$ are assumed independent with λ_i having a gamma distribution with parameters α_i and β_i. The Bayes estimator using this prior and under squared error loss will be discussed in the next section. When the parameters α_i and β_i are unknown and it is assumed that $\alpha_1 =\ldots= \alpha_p = \alpha$ and $\beta_1 =\ldots= \beta_p = \beta$, Leonard (1976) and Tsui and Press (1978b) find Bayes estimators when prior distributions

are placed on the parameters α and β. Leonard assumes $-\ell n\beta$ to be uniformly distributed over the real line and Tsui and Press additionally adopt various hypergeometric distributions for α. Bayesian estimates produced using these two stage priors are appropriate when $\lambda_1,\ldots,\lambda_p$ are close in size and can be thought to come from a common prior. In another paper, Leonard (1972) assumes that $\lambda_1,\ldots,\lambda_p$ are independent and identically distributed with $\ell n\ \lambda_i \sim N(\mu,\sigma^2)$, $i = 1,\ldots,p$. He then puts prior distributions on the parameters μ and σ^2 and proposes estimates based on the posterior distribution of $(\ell n\lambda_1,\ldots,\ell n\lambda_p)$.

When past data has been observed, a number of empirical Bayes methods have been proposed for estimating a Poisson mean. If m(x) denotes the marginal density of X with respect to the prior π, then one can represent the Bayes estimator of λ under squared error loss as

$$\frac{(x+1)m(x+1)}{m(x)}.$$

The past observations $X_1,\ldots,X_n$ are then used to estimate the marginal density m(x) and obtain a rule. This approach was introduced by Robbins (1955). Another empirical Bayes method uses the gamma (α,β) distribution for λ, as above, and in the expression of the Bayes estimator, estimates the unknown prior parameters from the past observations. One can write the Bayes estimator under squared error loss as

$$\delta^B(X) = \mu+(1 - \frac{1}{\beta+1})(X-\mu),$$

where $\mu = \alpha\beta$. One empirical Bayes estimator is found by substituting method of moments estimators for μ and β. These estimates are

$$\hat{\mu} = \sum_{i=1}^{n} X_i/n = \bar{X},$$

and

$$\hat{\beta} = \bar{X}^2/(s_1^2-\bar{X}), \quad \text{where } s_1^2 = \sum_{i=1}^{n} (X_i-\bar{X})^2/n.$$

The resulting rule is

$$\delta^{EB}(X) = \bar{X} + (1 - \frac{1}{\bar{X}^2/(s_1^2-\bar{X})+1})(X-\bar{X}).$$

(If $s_1^2 \leq \bar{X}$, then $\delta^{EB}(X) = \bar{X}$.) Other empirical Bayes methods approximate the prior π by the past observations and then obtain a Bayes rule. The reader is referred to Maritz (1969) for a summary of these techniques in Poisson estimation.

Hudson (1974) developed an empirical Bayes estimator for the simultaneous estimation of p means. For the loss L_1, he considered rules of the form

$$\delta_i(X) = \mu + (1-c(X))(X_i-\mu), \quad i = 1,\ldots,p,$$

where μ is the common prior mean of $\lambda_1,\ldots,\lambda_p$, and developed the estimator

$$\delta_i^H(X) = \mu + (1 - \frac{\sum_{i=1}^{p} X_i}{\sum_{i=1}^{p} (X_i-\mu)^2})(X_i-\mu), \quad i = 1,\ldots,p.$$

This rule is expected to do well when $\lambda_1,\ldots,\lambda_p$ are similar in size, and is similar to the rule that is developed in this paper.

1.2. Need for a robust Bayes estimator

As discussed in the introductory chapter, it is desired to find an estimator which allows the input of prior information but is robust or insensitive to parts of the prior knowledge that are uncertain. Before introducing a particular robust estimator, we now evaluate the performance of the conjugate Bayes estimator with regard to robustness criteria.

Let $\lambda_1,\ldots,\lambda_p$ have the conjugate prior density with parameters (α_i,β_i), $i = 1,\ldots,p$. Under the loss L_1, the Bayes estimator is componentwise

$$\delta_i^B(X) = \frac{(X_i+\alpha_i)\beta_i}{\beta_i+1}, \qquad i = 1,\ldots,p.$$

This estimator is easy to calculate and use and will substantially improve upon the classical estimator δ^0 when the prior density accurately reflects ones prior beliefs concerning $\lambda_1,\ldots,\lambda_p$.

As in the normal mean estimation problem, it is common here to have good knowledge about the central portion of the prior, but vague information about the tails of the prior. As mentioned in Chapter 1, a good prior should model our knowledge in the central region but have flat tails to be robust. The flat tails of the prior cause the corresponding Bayes estimator to ignore the prior information when outliers or extreme data are observed which make the prior implausible.

To see how extreme observations affect δ^B, note that δ_i^B may be written as

$$\delta_i^B(X) = X_i - \frac{X_i - \alpha_i\beta_i}{\beta_i+1},$$

where $\alpha_i\beta_i$ is the prior mean of λ_i. If X_i is an outlier, that is, if X_i is far from $\alpha_i\beta_i$ so that $|X_i-\alpha_i\beta_i|/(\beta_i+1)$ is large, then δ_i^B does not ignore the prior information. This lack of robustness is indicated by the fact that δ^B can perform substantially worse than δ^0 in terms of risk. The risk function of δ^B can be easily calculated to be

$$R(\delta^B,\lambda) = \sum_{i=1}^{p} \left(\frac{\beta_i}{\beta_i+1}\right)^2 \lambda_i + \sum_{i=1}^{p} \left(\frac{\lambda_i-\alpha_i\beta_i}{\beta_i+1}\right)^2.$$

Note that δ^0 has risk $R(\delta^0,\lambda) = \sum_{i=1}^{p} \lambda_i$. Observe that the dominant term in the risk of δ^B is $\sum_{i=1}^{p} ((\lambda_i-\alpha_i\beta_i)/(\beta_i+1))^2$, which increases quadratically as a function of λ_i, while $R(\delta^0,\lambda)$ increases linearly as a function of λ_i. Hence outside of a particular "prior region" of the parameter space about the prior mean, δ^B will show substantially worse risk than δ^0, and the decrement in risk becomes more severe as the distance from the prior mean increases.

Thus, using risk criteria, δ^B is not a robust Bayes estimator. The estimator δ^B will also perform poorly from a Bayes risk standpoint if the prior density is misspecified in that the "true" prior has tails much flatter than those of $g(\lambda)$. In this situation, the Bayes risk of δ^B will give much greater weight to the risk values corresponding to parameter values far from the prior mean. Thus it is possible for δ^B to have much larger Bayes risk than δ^0 when the true prior gives more weight to extreme values of λ.

In conclusion the conjugate Bayes estimator appears to be very sensitive to uncertainty in the tails of the prior distribution. The estimator that will be developed in the next section may be thought as an approximation to a Bayes estimator derived from a flat-tailed prior. It will be shown to be insensitive to values of the parameter far from the central region of the prior.

2. Development of the estimator

We would like our estimator to perform well in a particular prior region specified by the user. To this end, the Bayes estimator of λ is again considered under loss L_1, where $\lambda_1,\ldots,\lambda_p$ are assumed independent with λ_i having the gamma prior distribution. The Bayes estimator of λ_i may be written as

$$\delta_i^B(X) = \frac{(X_i+\alpha_i)\beta_i}{\beta_i+1}$$

$$= \alpha_i\beta_i + (1 - \frac{1}{\beta_i+1})(X_i-\alpha_i\beta_i).$$

Note that δ_i^B shrinks the observation X_i towards the prior mean $\alpha_i\beta_i$, and the amount of shrinkage is controlled by the inputted parameter β_i. An estimator is desired which shrinks towards the prior mean like δ^B, but restricts the amount of shrinkage when the observations appear to be inconsistent with the prior information. Also we would like to use all p observations $X_1,\ldots,X_p$ to estimate a particular component λ_i, since the improved estimators discussed in Section 1.1 use all of the observations. Therefore, consider estimators of the form

$$\hat{\delta}_i(X) = \mu_i + (1 - \frac{c(X)}{\beta_i+1})(X_i-\mu_i), \quad i = 1,\ldots,p,$$

where $\mu_i = \alpha_i\beta_i$ and $c(X)$ is a function of $X_1,\ldots,X_p$.

To find an appropriate $c(X)$, an argument similar to one in Hudson (1974) is used. If $c(X)$ is temporarily assumed to be a constant c, the risk of $\hat{\delta}$ under loss L_1 can be evaluated to be

$$R(\hat{\delta},\lambda) = \sum_{i=1}^{p} (1 - \frac{c}{\beta_i+1})^2\lambda_i + c^2 \sum_{i=1}^{p} (\frac{\lambda_i-\mu_i}{\beta_i+1})^2.$$

Minimizing the above expression with respect to c shows that the optimal c is

$$c' = \frac{\sum_{i=1}^{p} \lambda_i/(\beta_i+1)}{\sum_{i=1}^{p} \lambda_i/(\beta_i+1)^2 + \sum_{i=1}^{p} ((\lambda_i-\mu_i)/(\beta_i+1))^2}.$$

Although c' is a function of the unknown parameters $\lambda_1,\ldots,\lambda_p$, it can be estimated using the observations $X_1,\ldots,X_p$. In particular, if λ_i is estimated by its MLE X_i, we obtain the estimator

$$\delta_i'(X) = \mu_i + (1 - \frac{1}{\beta_i+1} \frac{\sum_{j=1}^{p} X_j/(\beta_j+1)}{\sum_{j=1}^{p} X_j/(\beta_j+1)^2 + \sum_{j=1}^{p} ((X_j-\mu_j)/(\beta_j+1))^2})(X_i-\mu_i),$$

$$i = 1,\ldots,p.$$

(It should be noted at this point that one can estimate c' by many different functions of $X_1,\ldots,X_p$ and produce different estimators. For example, Hudson estimates the numerator and denominator of c' separately using unbiased estimates. This particular estimate of c'

is chosen since it exhibits few singularities and leads to an estimator with many attractive properties.)

Using this method, Hudson derives an estimator similar to δ' with $\beta_1 = \ldots = \beta_p = 0$ and $\mu_1 = \ldots = \mu_p$. The estimator that is discussed here is an extension of Hudson's estimator in that it permits a different prior mean and variance input for each of the p Poisson parameters estimated.

Finally, since we would like our estimator to act like a Bayes estimator in a particular region of the parameter space, it is natural to restrict the shrinkage of X_i towards μ_i to the amount that δ^B shrinks X_i. δ' is then modified to

$$\delta_i^*(X) = \mu_i + \left(1 - \frac{1}{\beta_i+1} \min\left\{1, \frac{\sum_{j=1}^{p} X_j/(\beta_j+1)}{\sum_{j=1}^{p} X_j/(\beta_j+1)^2 + \sum_{j=1}^{p} ((X_j-\mu_j)/(\beta_j+1))^2}\right\}\right) \cdot (X_i-\mu_i), \quad i = 1,\ldots,p.$$

Although δ^* has not been developed explicitly as an empirical Bayes estimator, the above derivation is similar to an empirical Bayes derivation. In Efron and Morris (1974), some of the parameters in the prior density are unknown and the observations are used in estimating them. Here we want to keep known prior parameters in the estimator, and use the observations in two ways. First, when the prior parameters have been specified correctly, the observations are used to produce a shrinking constant which will ideally act like the shrinking constant of a natural Bayes estimator. Second,

the observations are used to make the estimator robust to errors in specifying the prior parameters.

The shrinking constant of δ_i^* is

$$1-c_i^*(X) = 1 - \frac{1}{\beta_i+1} \min\{1, \frac{\sum_{j=1}^{p} X_j/(\beta_j+1)}{\sum_{j=1}^{p} X_j/(\beta_j+1)^2 + \sum_{j=1}^{p} ((X_j-\mu_j)/(\beta_j+1))^2}\}.$$

To better understand how δ^* performs like the conjugate Bayes estimator, consider $1-c_i^*(X)$ when many Poisson means are estimated. Note that λ_i has prior mean μ_i and prior variance $\mu_i\beta_i$, and marginally X_i has mean μ_i and variance $\mu_i(\beta_i+1)$. Therefore, for large p, $1-c_i^*(X)$ may be approximated (using the law of large numbers) by

$$1 - \frac{1}{\beta_i+1} d,$$

where

$$d = \frac{\sum_{j=1}^{p} \mu_j/(\beta_j+1)}{\sum_{j=1}^{p} \mu_j/(\beta_j+1)^2 + \sum_{j=1}^{p} \mu_j/(\beta_j+1)}.$$

Note that $1/2 \leq d \leq 1$ and d will be close to one when moderate or large values of $\beta_1,\ldots,\beta_p$ are chosen. If d is approximately equal to one, the shrinking constant $1-c_i^*(X)$ is approximately equal to $1-(\beta_i+1)^{-1}$ and $\delta_i^* \tilde{=} \delta_i^B$. Therefore when the prior information is specified correctly, moderate values of $\beta_1,\ldots,\beta_p$ are used and p is large, δ^* should behave much like the conjugate Bayes estimator δ^B.

In the above paragraph, the behavior of the shrinking constant $1-c_i^*(X)$ is discussed when the prior information is correct. On the other hand, if the prior information has been misspecified, then at least one observation X_i will be far from its prior mean μ_i and $\sum_{j=1}^{p} ((X_j-\mu_j)/(\beta_j+1))^2$ will be large. In this case, for any i, $c_i^*(X)$ will be small and $\delta_i^*(X) \cong X_i$, so the estimator is in fact ignoring the wrong prior information.

The above discussion indicates that δ^* is a candidate to be a robust Bayes estimator. If our prior information is correct, δ^* is an approximation to the Bayes estimator δ^B; otherwise if an observation casts doubt on the prior information, δ^* rejects the prior knowledge and approaches the MVUE. In the following section, the above statements will be made more precise and it will be argued that δ^* is an attractive alternative to δ^0.

3. Evaluation of the estimator

3.1. Methods of evaluation

The first section of our evaluation, Section 3.2, is concerned with the performance of δ^* when correct prior information is used. This section describes the different ways in which δ^* is an improved estimator over the usual one δ^0. In particular, we discuss (i) the performance in risk of δ^* in a prior region, (ii) the performance in Bayes risk of δ^*, and (iii) the risk performance of δ^* when the prior information has not been chosen symmetrically.

In Section 3.3, the robustness of δ^* to misspecified prior information is investigated. Generally it is shown that δ^* is a safe estimator with regard to uncertainty in the prior specification. We investigate (i) the performance in Bayes risk of δ^* when wrong prior information is used, (ii) the risk performance of δ^* for extreme values of $\lambda_1,\ldots,\lambda_p$, (iii) the risk performance of δ^* when very large prior means are used, and (iv) the risk performance of δ^* when $p = 1$.

In the first chapter, it is stated that one way to analyze the performance of an estimator with regard to the input of correct prior information is to look at the risk function in a prior region of the parameter space. In Section 3.2, the risk functions of δ^* and δ^0 will be compared for moderate values of the parameters $\lambda_1,\ldots,\lambda_p$, and since the risk of δ^* is not attainable in closed form, numerical studies will be used in our analysis. It is first shown how prior information is reflected in the location and the size of the improvement region, that is, the region where δ^* displays smaller risk than δ^0. It is shown that it is advantageous in particular situations to estimate a larger number of means simultaneously - the risk improvement of δ^* relative to δ^0 becomes more pronounced when more means are estimated.

Next the performance of δ^* will be compared with the performance of other estimators proposed to estimate p Poisson means simultaneously. Two of these estimators have the effect of shrinking the observation X towards zero, so as constructed they are not able to shrink the data toward nonzero means. We will compare δ^* with modified versions of

these estimators that can accept prior means. Next by means of numerical studies, Bayes risks of δ^* will be computed, and these will be compared with the optimal Bayes risks and Bayes risks of δ^0 when the prior information is correct.

Finally the performance of δ^* in the prior region is investigated when asymmetry exists in the prior information. Ideally δ^* should show substantial risk improvement upon δ^0 regardless of differences in prior means and variances selected. It is shown that particular prior components can dominate the others with respect to their influence on the risk of the estimator, but it will be argued that in many applications of this estimator, this problem does not occur.

In Section 3.2, δ^* is shown to use prior information like the Bayes estimator δ^B. In Section 3.3, δ^* is shown to be more robust than δ^B with respect to wrong prior information. First Bayes risks of δ^* and δ^B are found when the prior information is improperly specified, and δ^* is shown to be robust with respect to Bayes risk.

The remainder of our evaluation uses risk as a criterion of robustness. As mentioned in the first chapter, one way of evaluating the robustness of a particular estimator is to analyze its risk for values of the parameter away from the prior region. The first situation considered is that in which the prior means $\mu_1,\ldots,\mu_p$ are fixed and the parameters $\lambda_1,\ldots,\lambda_p$ get very large along a particular line. A theorem is stated which gives the asymptotic risk improvement of δ^* over δ^0. It has already been noted that in this situation, where $\lambda_1,\ldots,\lambda_p$ lie far from the prior region, the risk of δ^B is much

greater than the risk of δ^0. It is shown that, asymptotically, the risk of δ^* is no more than two units greater than the risk of δ^0, and along particular lines it can be smaller. Thus δ^* is more robust than δ^B in the situations where extreme values of the parameter (in the sense of being far away from the prior mean) are possible.

The second major result in this section concerns the selection of large prior means and their effect on the risk behavior of δ^*. It is known through the simulation work in Section 3.2 that δ^* performs well for moderate values of prior means. It is of interest to see how the risk function of δ^* compares with that of δ^0 when very large prior means are selected. We are primarily interested in how poorly δ^* can do relative to δ^0 outside of the prior region. First the proportional improvement in risk of δ^* over δ^0 is defined, and the situation is considered in which both $(\mu_1,\ldots,\mu_p)$ and $(\lambda_1,\ldots,\lambda_p)$ are getting large at the same rate. (If they increased at different rates, it would be the "large λ" case covered by the previously mentioned theorem.) It is shown that in this asymptotic situation, the proportional improvement of δ^* is equivalent to the risk improvement of a Stein-type estimator in the normal mean estimation problem.

The above mentioned result is used in two ways. First, the maximum proportional decrement in risk of δ^* compared to δ^0 is heuristically found in this asymptotic situation for all p (number of parameters estimated) and all selections of prior parameters. Second, the theorem is used in studying the behavior of δ^* in estimating one Poisson mean. To complete the discussion of the one-dimensional

estimator, the risk of δ^* for small values of λ is considered, and the maximum proportional decrement relative to δ^0 is found.

3.2. Incorporation of prior information

3.2.1. Introduction

Prior beliefs concerning the set $\{\lambda_1,\ldots,\lambda_p\}$ may be expressed through the parameters $\mu = (\mu_1,\ldots,\mu_p)$ and $\beta = (\beta_1,\ldots,\beta_p)$. In the Bayesian estimation model described earlier, μ_i and $\mu_i\beta_i$ are respectively the mean and variance of the prior distribution of λ_i. Thus larger values of β_i reflect a flatter and less informative prior distribution. The estimator δ^* shrinks the observation X towards μ, and the amount of shrinkage along the ith component is restricted to $|X_i-\mu_i|/(\beta_i+1)$, the amount that the Bayes estimator δ_i^B shrinks X_i.

3.2.2. Numerical studies - risk

In the figures that are to be presented, we show that δ^* has a significantly smaller risk than δ^0 in a region about μ and has a risk not much larger than δ^0 elsewhere. All of the risks of δ^* are found numerically using a computer, since the risk of δ^* generally is not expressible in a closed form. For the examples that are presented, at least 5000 occurrences of the random variable X are simulated, and the risks (or average losses) that are found have a standard error of approximately 5 per cent. On each graph, the risk of δ^0 (which is $\sum_{i=1}^{p} \lambda_i$ under loss L_1) is drawn to facilitate comparisons with δ^0.

We first look at the risk function of δ^* under squared error loss (loss L_1) in the simplest case, p = 1, when μ = 4 (Figure 1).

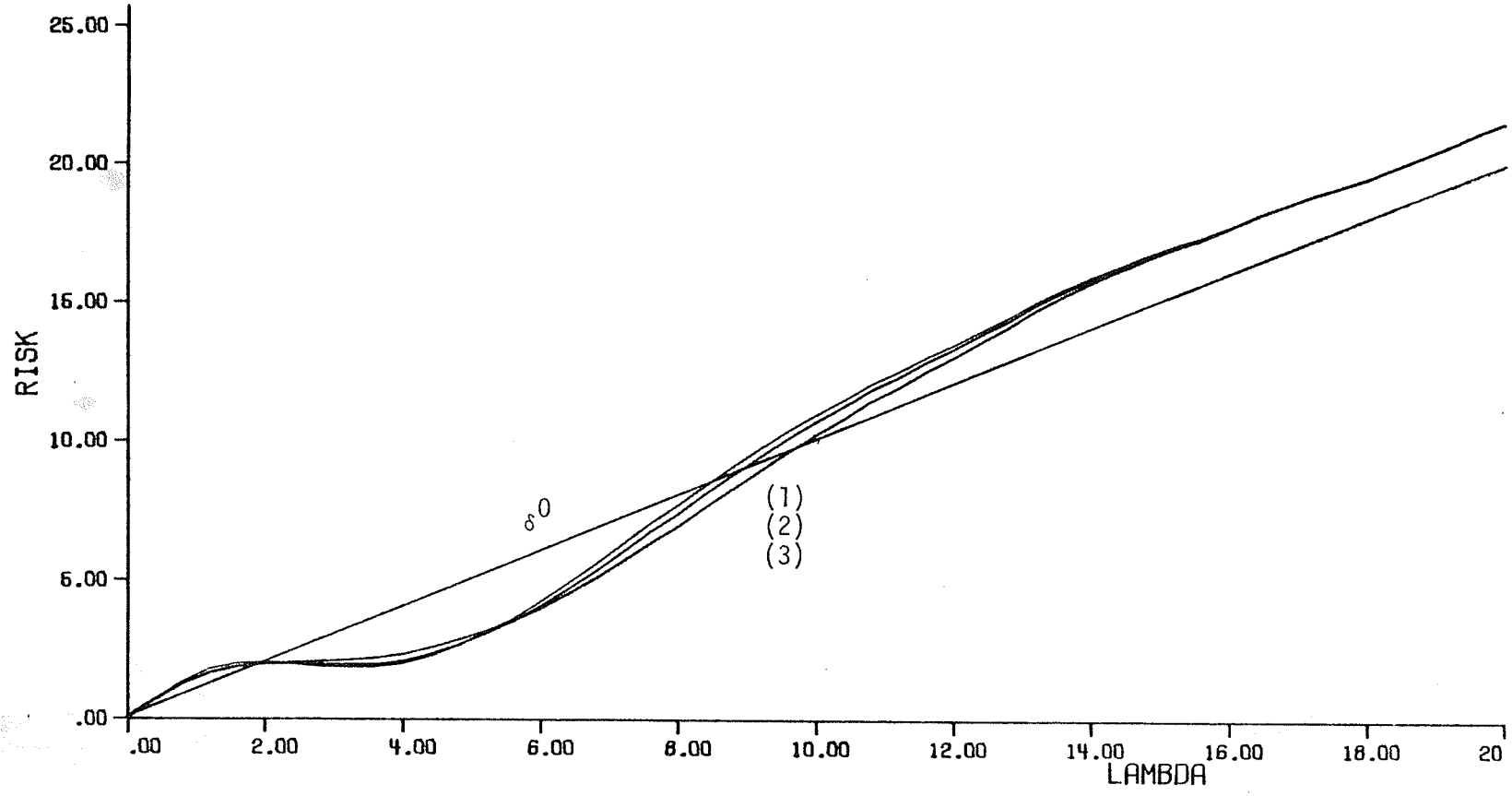


Figure 1

p=1. Risks of (1) δ^*, $(\mu,\beta)=(4,0)$, (2) δ^*, $(\mu,\beta)=(4,1)$, and (3) δ^*, $(\mu,\beta)=(4,2)$.

This graph shows the effect of increasing β from 0 to 2. For each value of β, the corresponding estimator achieves approximately a 50 per cent improvement in risk over δ^0 at $\mu = 4$ and at its worst, loses about one unit in risk ouside of the improvement region. The effect of increasing β is to widen the region of improvement while slightly lowering the amount of improvement close to $\mu = 4$. As will be shown in Section 3.3.2, asymptotically as λ approaches infinity, the one dimensional estimator loses one unit in risk compared to δ^0.

In dimensions greater than one, it is convenient to consider the proportional risk of δ^*, defined by

$$\frac{R(\delta^*,\lambda)}{R(\delta^0,\lambda)} = \frac{E[\sum_{i=1}^{p} (\delta_i^*(X)-\lambda_i)^2]}{\sum_{i=1}^{p} \lambda_i} .$$

Figure 2 presents a graph for p = 2 showing contours of constant values of proportional risk. Here the prior parameters $(\mu_1,\beta_1) = (\mu_2,\beta_2) = (4,0)$ are used. Keeping in mind that a proportional risk of less than one signifies improvement of δ^* over δ^0, one sees that the region of improvement is quite large. At this point, two comments should be made concerning this selection of prior information. First, by selecting $\beta_1 = \beta_2 = 0$, the MVUE X is shrunk as far as possible for our estimators towards the prior mean (4,4). Therefore the proportional risk value of .38 at the point (4,4) is lower than the minimum value realized when positive values of β_1 and β_2 are used. Second, from our experience it appears that the region of improvement is smallest when $\beta_1 = \beta_2 = 0$, so the improvement region should be

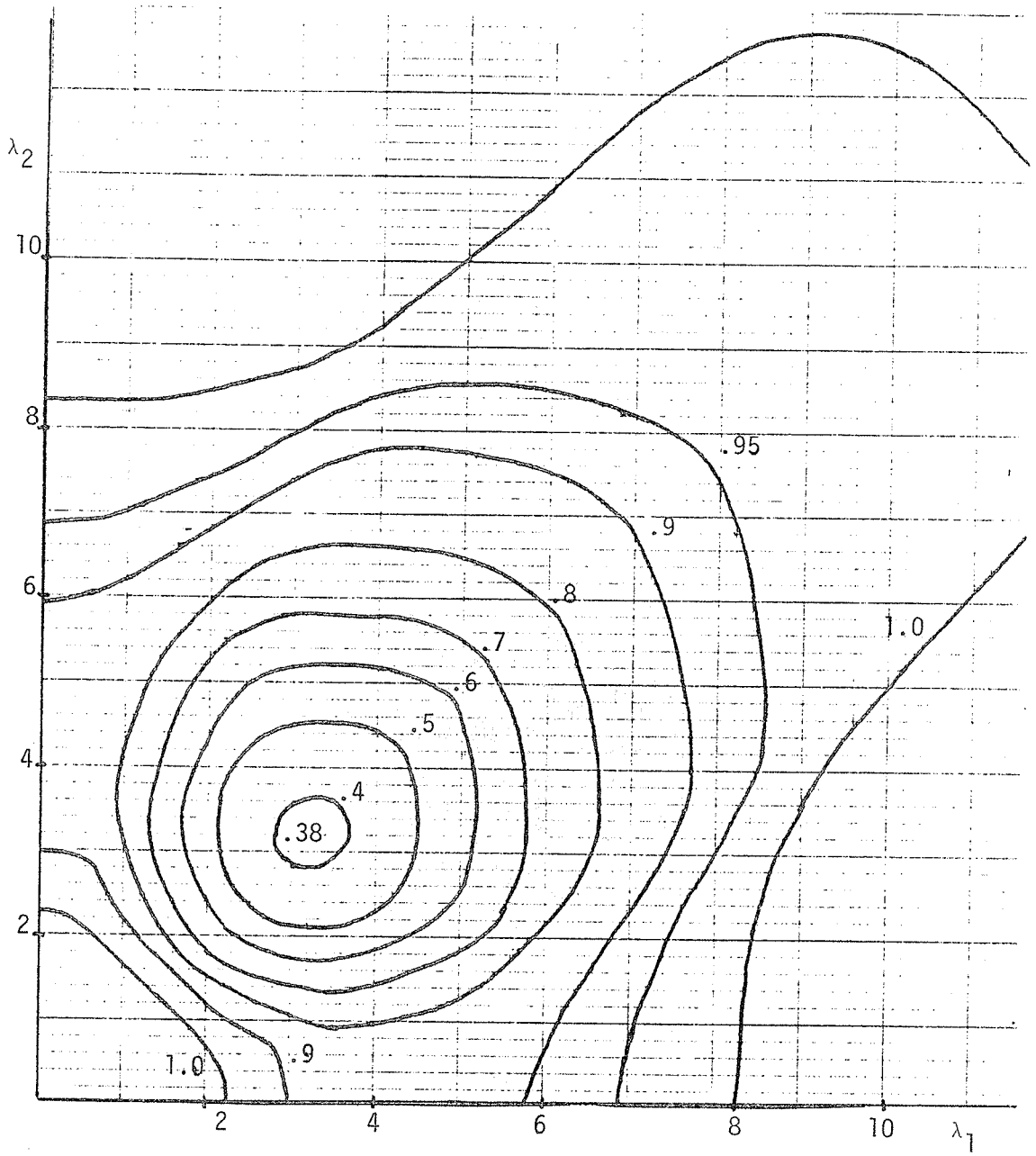


Figure 2

p=2. Contours of constant values of proportional risk of δ^*. Prior information: $(\mu_i,\beta_i)=(4,0)$, i=1,2.

satisfactory when positive β_1 and β_2 are used. Through our computer simulation and knowledge about the asymptotic behavior of δ^*, it seems that the proportional risk outside of the outer contour of 1 is bounded above by 1.1 and for small (λ_1,λ_2) beyond the inside contour of 1, the proportional risk appears to be bounded by

$$\lim_{\lambda_1,\lambda_2\to 0} \frac{R(\delta^*,\lambda)}{\sum_{i=1}^{2} \lambda_i} = 1.27.$$

In Chapter 1, it was noted that the amount of improvement of many alternative estimators over the MVUE becomes more substantial when more means are estimated simultaneously. Figures 3, 4 and 5 show that δ^* displays the same type of behavior. To understand Figure 3, consider the average risk, which is defined by

$$\frac{R(\delta^*,\lambda)}{p} = \frac{1}{p} E[\sum_{i=1}^{p} (\delta_i^*(X)-\lambda_i)^2].$$

When this average risk is calculated along the diagonal $\lambda_1 = \ldots = \lambda_p = \eta$, it is a good measure of the risk of δ^* in estimating a single coordinate of λ. Figure 3 shows the average risk (plotted against η) along this diagonal when p = 1,2 and 3 and the prior information is $(\mu_i,\beta_i) = (4,0)$ for each coordinate. The average risk of δ^0, $R(\delta^0,\lambda) = \eta$, is also plotted, so one can see the region of improvement and the amount of improvement of δ^*. One sees that the maximum decrement in average risk of δ^* compared to δ^0 for small η decreases as p increases, and the region of risk improvement increases for larger p. It should be

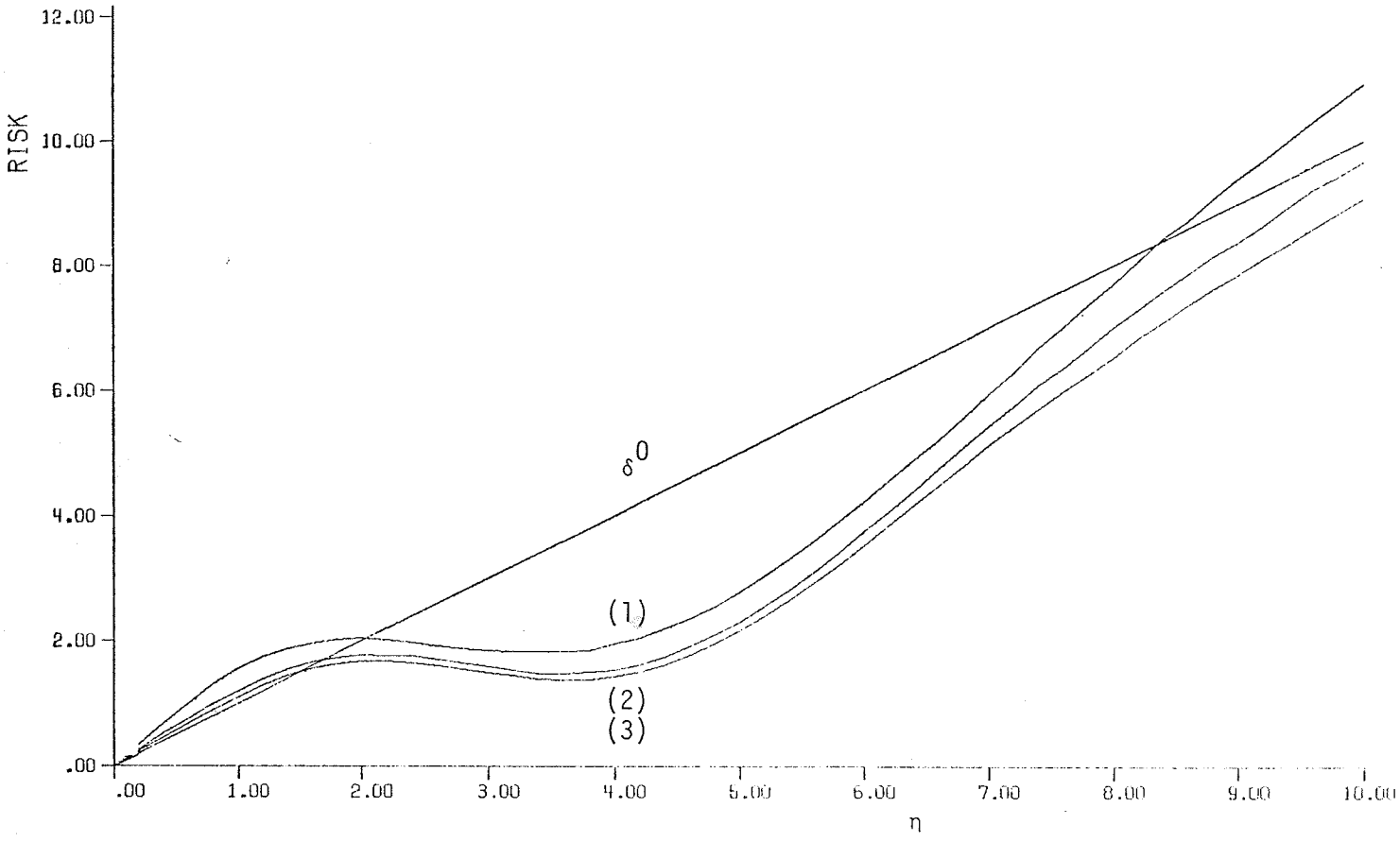


Figure 3

Average risks of (1) δ^*, p=1, (2) δ^*, p=2, (3) δ^*, p=3 along line $\lambda_1 = \ldots = \lambda_p = \eta$.

noted that the estimator δ^* is being viewed along the most favorable line, so the amount of improvement indicated here is not typical for all lines in the parameter space. Nevertheless Figure 3 illustrates the advantage of using all of the observations in estimating a single coordinate of λ when $\lambda_1,\ldots,\lambda_p$ are related.

Figures 4 and 5 show the risk function of δ^* in the cases $p = 3$ and $p = 6$ respectively, where $(\mu_i,\beta_i) = (4,4)$ for each coordinate. The risks in each case are plotted along three lines, L_1, where $\lambda_1 = \ldots = \lambda_p$, and L_2 and L_3, where there are moderate and severe differences between the λ_i's. It is expected that δ^* will perform well along line L_1 since the line goes through the shrinkage point $(4,\ldots,4)$, and also δ^* performs well asymptotically along this line (see Section 3.3.2). The risks are plotted as functions of $\sum_{i=1}^{p} \lambda_i$ so we can compare δ^* with δ^0. By looking at the two figures, we notice that in each case δ^* performs well compared to δ^0 along all three lines and offers some improvement along L_3 which is far removed from the prior mean. To see the effect of increasing p, first note that the prior variance of $\sum_{i=1}^{p} \lambda_i$ is $\sum_{i=1}^{p} \mu_i\beta_i$, and the corresponding prior standard deviation, $(\sum_{i=1}^{p} \mu_i\beta_i)^{1/2}$, can be used to compare graphs of different dimensions. In particular, the improvement region of δ^* can be viewed at fixed prior standard deviations (of $\sum_{i=1}^{p} \lambda_i$) away from the prior mean. In this example, it is not clear how to compare the two figures along lines L_2 and L_3, but a comparison can be made easily along line L_1. The prior standard deviation is $(\sum_{i=1}^{p} \mu_i\beta_i)^{1/2} = 4p^{1/2}$ in this case, and measured in these units, δ^* appears to have a

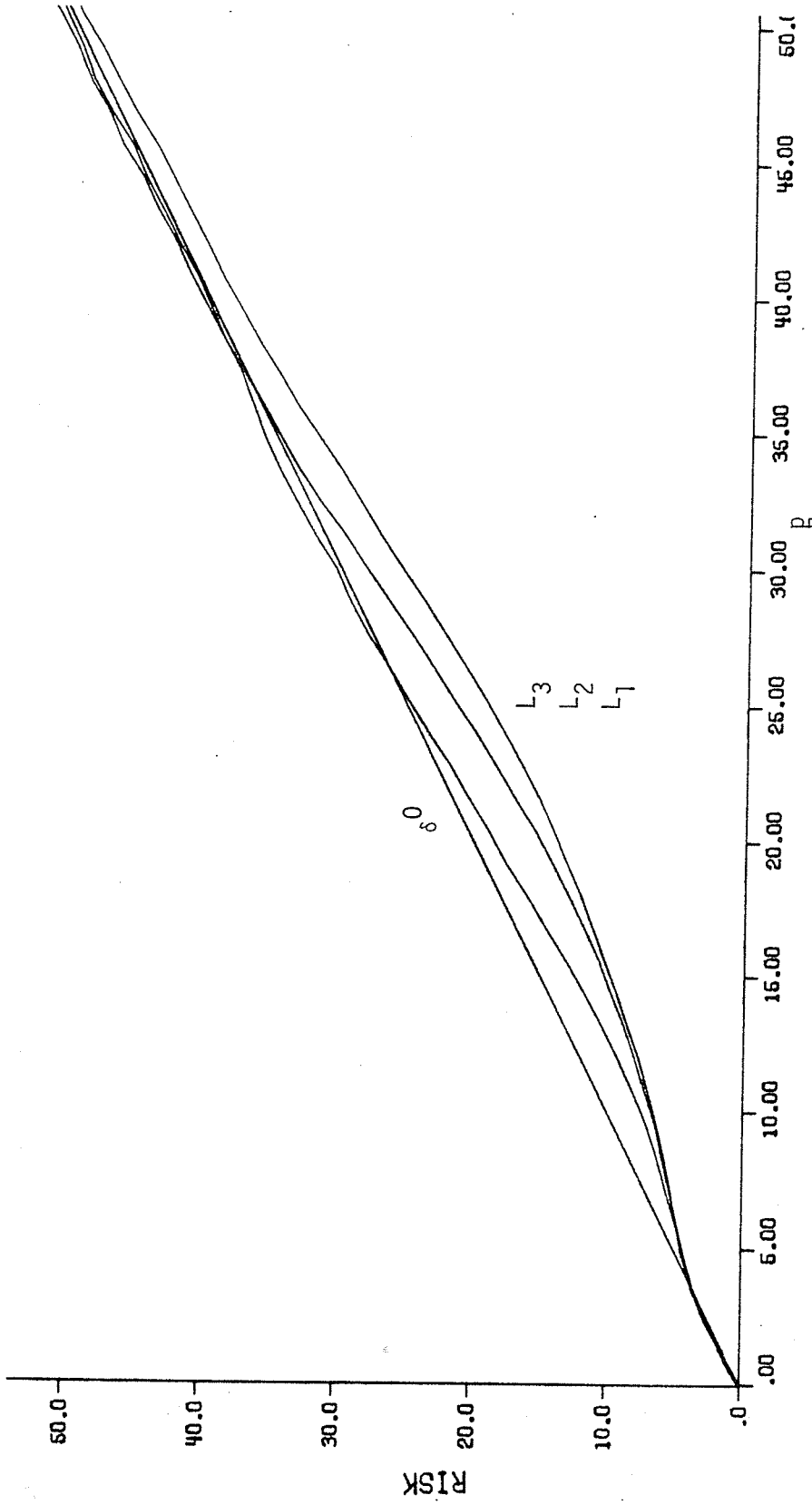


Figure 4

p=3. Risk of δ^* along lines (1) L_1: $\lambda_1=\lambda_2=\lambda_3$, (2) L_2: $\lambda_1=3\eta$, $\lambda_2=5\eta$, $\lambda_3=7\eta$, and (3) L_3: $\lambda_1=\eta$, $\lambda_2=10\eta$, $\lambda_3=19\eta$.

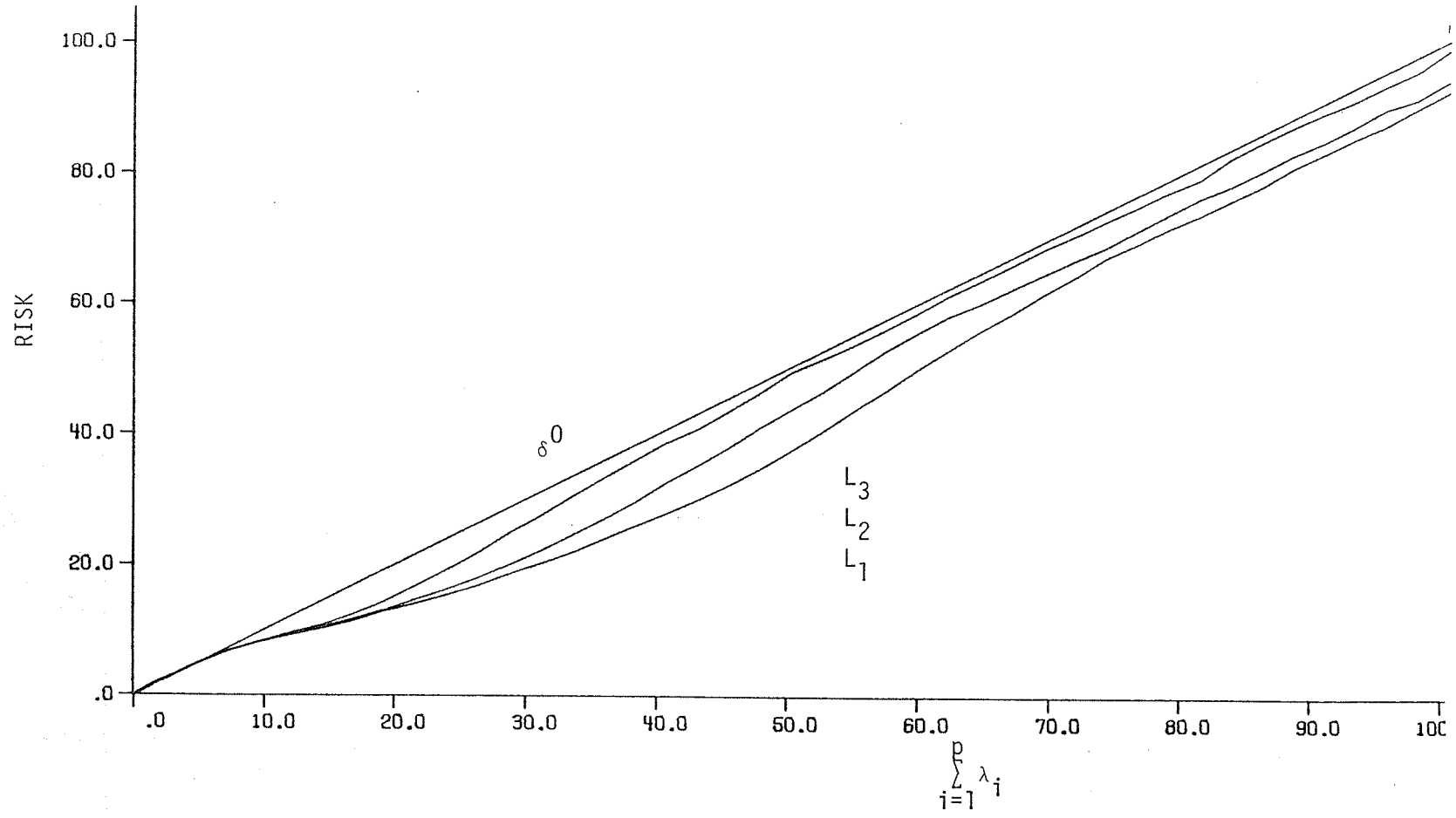


Figure 5

p=6. Risk of δ^* along lines (1) L_1: $\lambda_1=\ldots=\lambda_6$, (2) L_2: $\lambda_1=\eta$, $\lambda_2=2\eta$, $\lambda_3=4\eta$, $\lambda_4=5\eta$, $\lambda_5=5\eta$, $\lambda_6=7\eta$, and (3) L_3: $\lambda_1=\lambda_2=\lambda_3=\eta$, $\lambda_4=\lambda_5=\lambda_6=15\eta$.

larger region of improvement for p = 6 than for p = 3.

Table 1 shows values of the proportional risk of δ^0 ($R(\delta^*,\lambda)/R(\delta^0,\lambda)$) in the situation where $\mu_i = \beta_i = 4$ for all $i = 1,\ldots,p$ at the points $(2,\ldots,2)$, $(4,\ldots,4)$, etc. In this example, the prior standard deviation of $\sum_{i=1}^{p} \lambda_i$ is $(\sum_{i=1}^{p} \mu_i\beta_i)^{1/2} = 4p^{1/2}$ and the points represent fixed amounts of standard deviations from the prior mean. Although the proportional risk does not appear to decrease much at the shrinkage point $(4,\ldots,4)$ as p increases from 2 to 6, it does appear to decrease significantly at other points. This table further demonstrates how the region of improvement expands as p increases. It also shows the size of the improvement region measured in terms of prior standard deviations from μ.

Let us compare δ^* with three suggested estimators of p Poisson means, those proposed by Zidek and Clevenson (1975), Peng (1975) and Tsui (1978). These estimators have already been defined in Section 1.1. First δ^* will be compared with δ^P and δ^Z and the case p = 3 is considered. Since both Peng's and Zidek and Clevenson's estimators were designed to shrink X towards the orgin, we initially set $(\mu_i,\beta_i) = (0,0)$ for each component in δ^*.

Figure 6 compares the risk functions of δ^Z with $\gamma = 1$, δ^Z with $\gamma = p-1 = 2$, δ^P and δ^* along the diagonal line $\lambda_1 = \lambda_2 = \lambda_3$ for the loss L_1. Figure 7 compares the four estimators along the same line under the loss L_2. From the graphs, one sees that both versions of δ^Z do slightly better than δ^* close to the origin. But δ^* has smaller risk than δ^Z, $\gamma = 2$ outside of that region, and

Table 1

Values of proportional risk of δ^* for different values of p.

Prior information in δ^*: $\mu_i = \beta_i = 4$ for all i.

Point	Prior standard deviations away from prior mean	Values of $R(\delta^*,\lambda)/R(\delta^0,\lambda)$ p		
		2	3	6
(2)	.5	.824	.821	.762
(4)	0	.646	.646	.639
(6)	.5	.676	.671	.666
(8)	1	.776	.743	.724
(10)	1.5	.881	.849	.834
(12)	2.0	.965	.927	.892

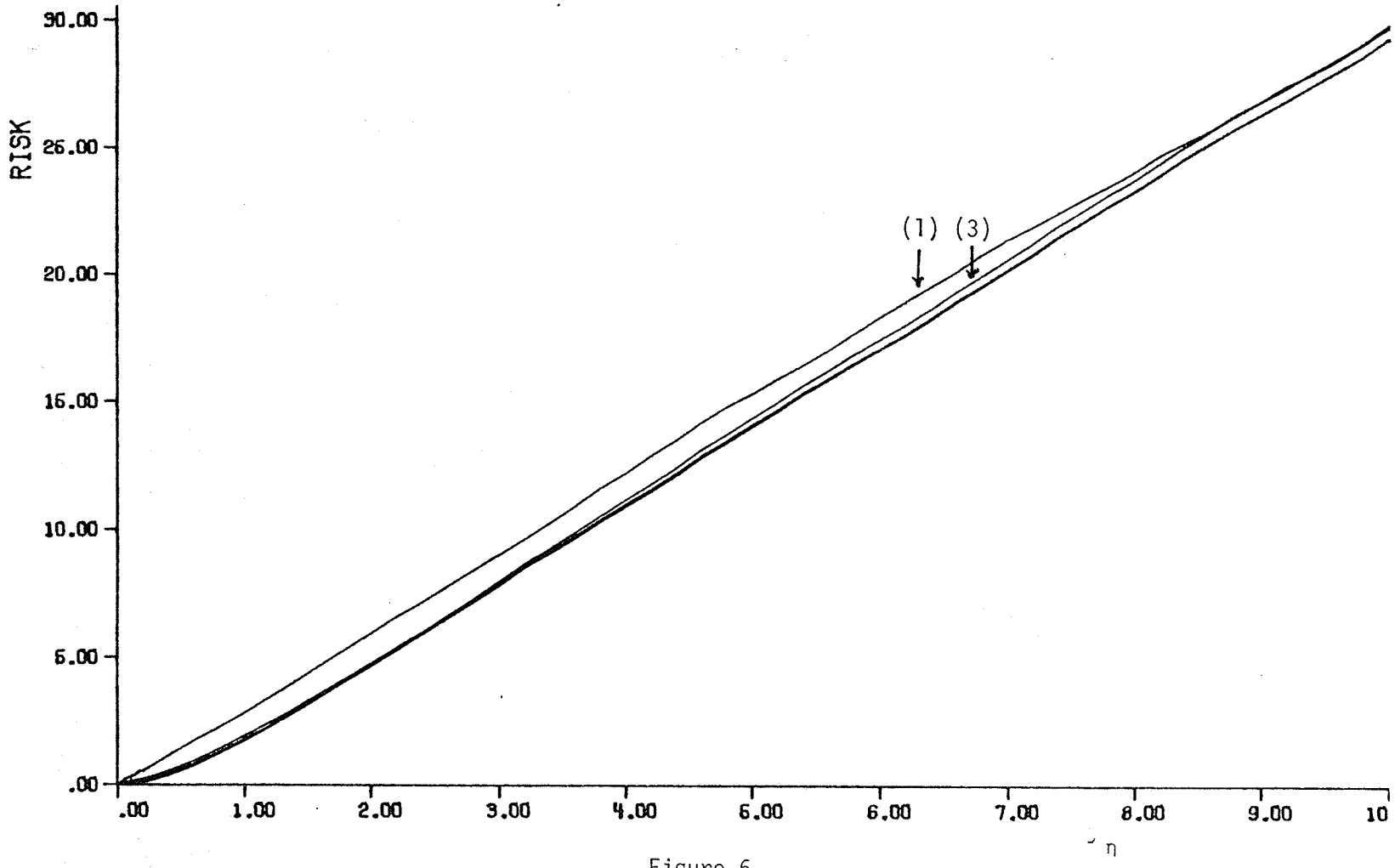


Figure 6

p=3. Prior information: $\mu_i=\beta_i=0$ for all i. Loss L_1. Risks of (1) δ^P, (2) δ^Z, $\gamma=1$, (3) δ^Z, $\gamma=2$, and (4) δ^* along line $\lambda_1=\lambda_2=\lambda_3=\eta$.

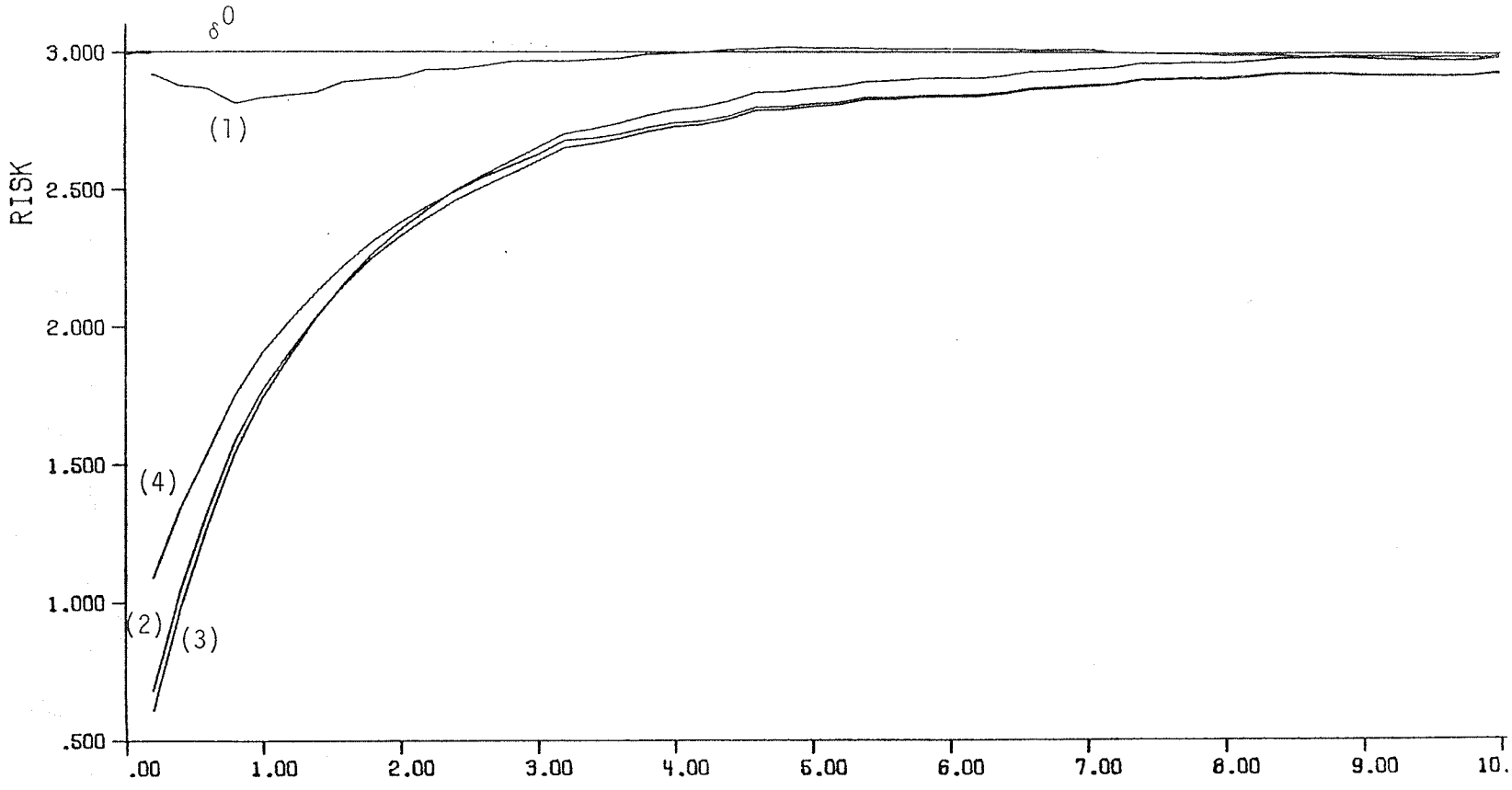


Figure 7

p=3. Prior information; $\mu_i=\beta_i=0$ for all i. Loss L_2. Risks of (1) δ^P, (2) δ^Z, $\gamma=1$, (3) δ^Z, $\gamma=2$, and (4) δ^* along line $\lambda_1=\lambda_2=\lambda_3=\eta$.

in general, Zidek's estimators and δ^* appear to be roughly equivalent in terms of risk. One also notes that δ^P only improves marginally on the MVUE - it does not shrink X significantly towards the origin.

Let us next adjust δ^Z and δ^P to shrink towards prior means in natural ways, and then compare these estimators with δ^*. Define componentwise the adjusted estimators

$$\delta_i^{Z'}(X) = \mu_i + \left(1 - \frac{\gamma+p-1}{\sum_{j=1}^{p} |X_j-\mu_j| + \gamma+p-1}\right)(X_i-\mu_i), \quad 1 \le \gamma \le p-1,$$

and

$$\delta_i^{P'}(X) = X_i - \frac{(p-N_0-2)}{S'}{}^{+} \left(\sum_{k=1}^{X_i} \frac{1}{k} - \sum_{k=1}^{[\mu_i]} \frac{1}{k}\right),$$

where $[\mu_i]$ = greatest integer $\le \mu_i$ and

$$S' = \sum_{i=1}^{p} \left(\sum_{k=1}^{X_i} \frac{1}{k} - \sum_{k=1}^{[\mu_i]} \frac{1}{k}\right)^2.$$

Consider the case p = 3 with prior means $\mu_1 = \mu_2 = \mu_3 = 4$. Figures 8 and 9 show the risk functions, under the two losses L_1 and L_2 respectively, of $\delta^{Z'}$ with $\gamma = 1$, $\delta^{Z'}$ with $\gamma = 2$, $\delta^{P'}$ and δ^*. We have set $\beta_1 = \beta_2 = \beta_3 = 0$ in the rule δ^*. Also note that, as before, the risks are plotted along the line $\lambda_1 = \lambda_2 = \lambda_3$. It is seen that all four estimators shrink towards the prior mean, and the estimators of Zidek and Clevenson and δ^* have similar risks in the region of greatest improvement. But the risks of $\delta^{Z'}$, $\gamma = 1$ and $\delta^{Z'}$, $\gamma = 2$ are much worse than those of $\delta^{P'}$ and δ^* when λ is close to the origin,

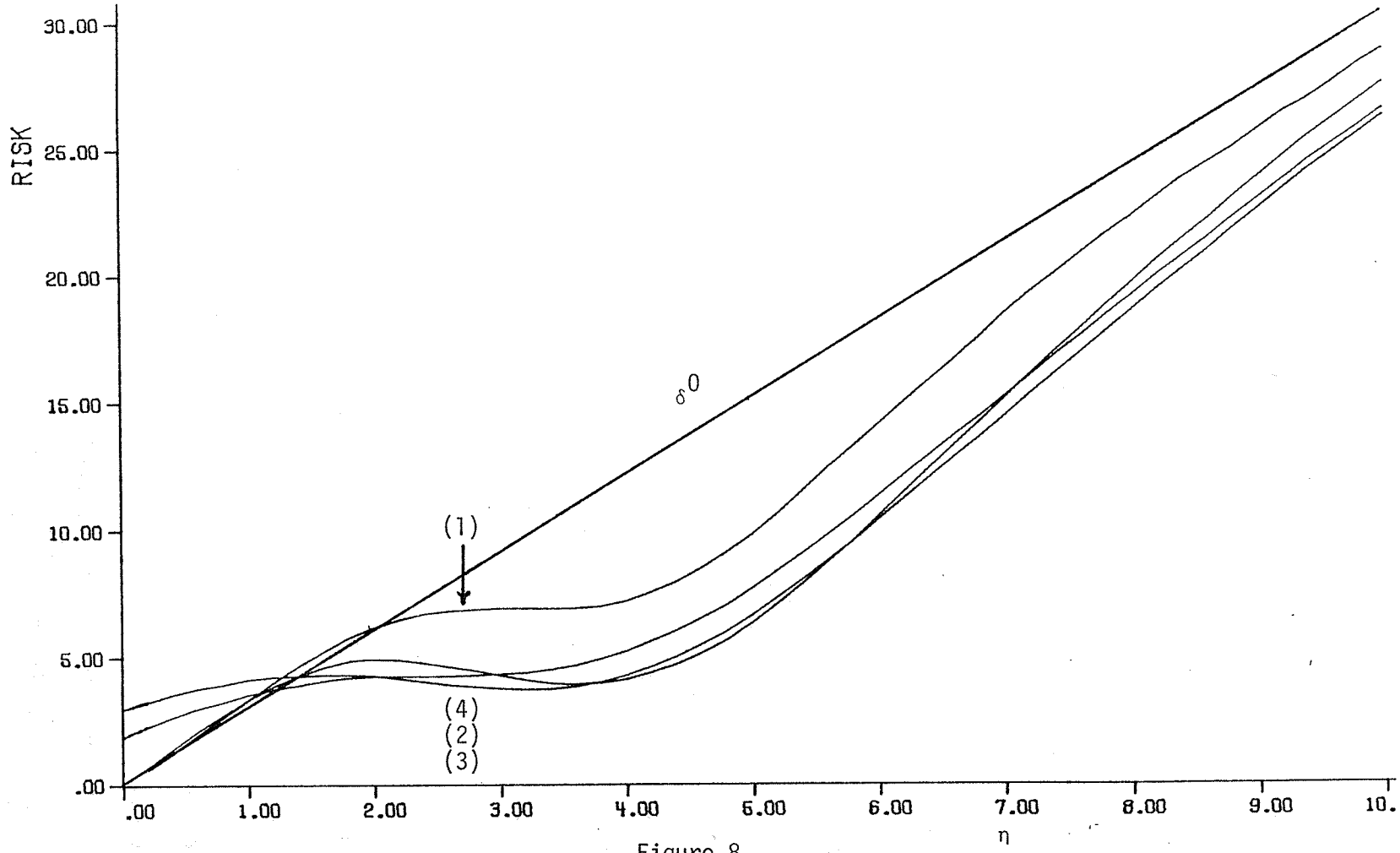


Figure 8

p=3. Prior information: $(\mu_i,\beta_i)=(4,0)$ for all i. Loss L_1. Risks of (1) $\delta^{P'}$, (2) $\delta^{Z'}$, $\gamma=1$, (3) $\delta^{Z'}$, $\gamma=2$, and (4) δ^* along line $\lambda_1=\lambda_2=\lambda_3=\eta$.

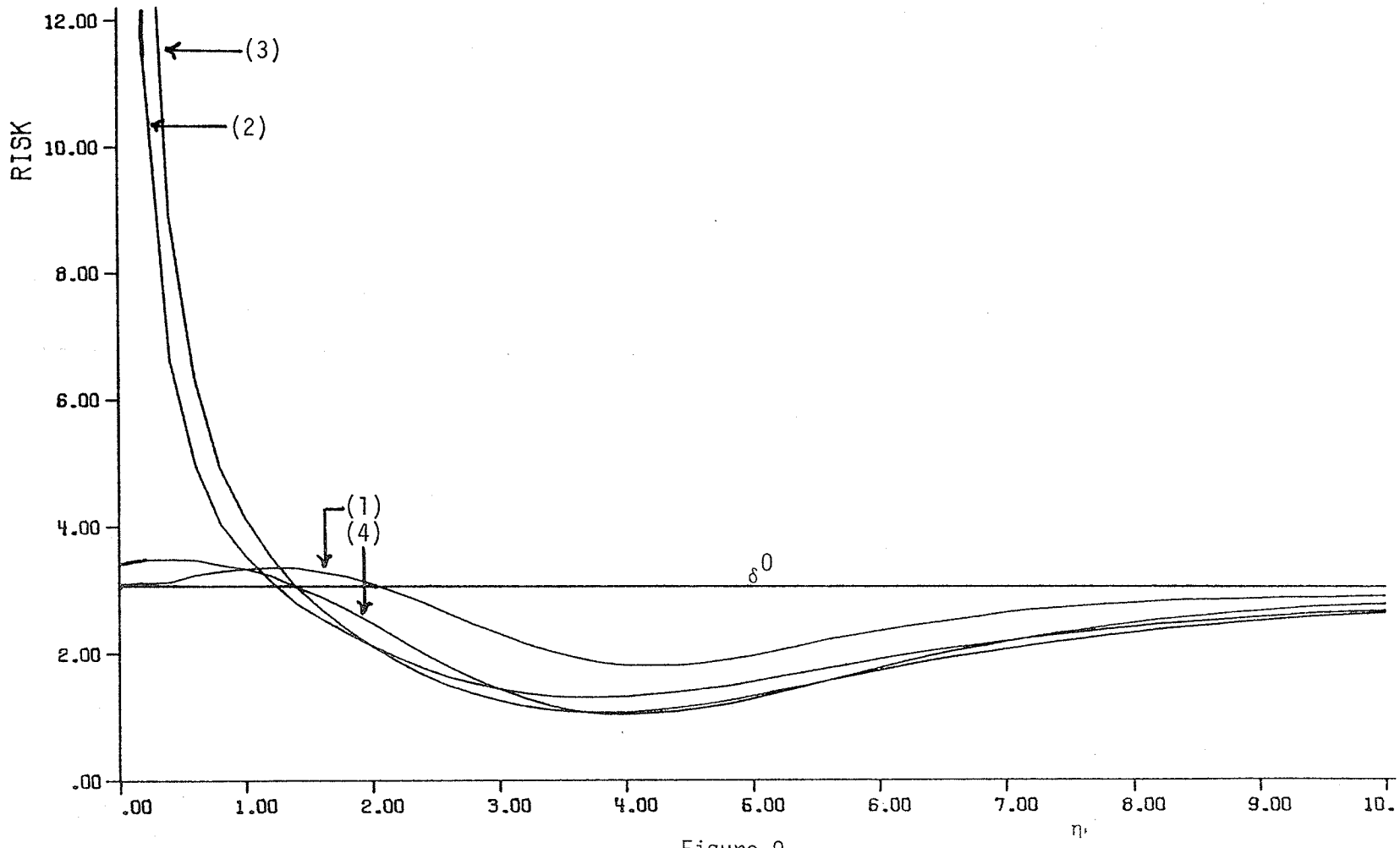


Figure 9

p=3. Prior information $(\mu_i,\beta_i)=(4,0)$ for all i. Loss L_2. Risks of (1) $\delta^{P'}$, (2) $\delta^{Z'}$, $\gamma=1$, (3) $\delta^{Z'}$, $\gamma=2$, and δ^* along line $\lambda_1=\lambda_2=\lambda_3=\eta$.

especially when viewed under loss L_2. This is due to the fact that the term in $\delta^{Z'}$,

$$\frac{\gamma+p-1}{\sum_{j=1}^{p}|X_j-\mu_j|+\gamma+p-1},$$

does not approach zero as $\lambda_1,\ldots,\lambda_p$ (and therefore $X_1,\ldots,X_p$) approach zero.

Next, δ^* is compared with δ^T, which does allow incorporation of a prior mean K. In this example, set K = 4 and $\mu' = 1$ in Tsui's estimator, and set $(\mu_i,\beta_i) = (4,0)$ for each component in δ^*. Figure 10 shows the risks of the two estimators for p = 6 along the line $\lambda_1 = \ldots = \lambda_p$. As in the case of δ^P, one notes that δ^T only offers marginal risk improvement over δ^0 near the prior mean and δ^* has a substantially smaller risk than δ^T near μ.

In this section it has been shown through numerical work how δ^* makes use of prior information as reflected in the region of risk improvement in the parameter space. It has been shown how this region is affected by the choice of the prior parameters μ and β and how the region can expand as one simultaneously estimates more means. Comparisons were made of the risk functions of δ^* and other proposed minimax estimators. The estimators of Peng and Zidek and Clevenson have the disadvantage of not being able to accept arbitrary prior means. Tsui's estimator does shrink toward an arbitrary prior mean, but it is not able to improve substantially upon the MVUE in the prior region. The robust Bayes estimator δ^*

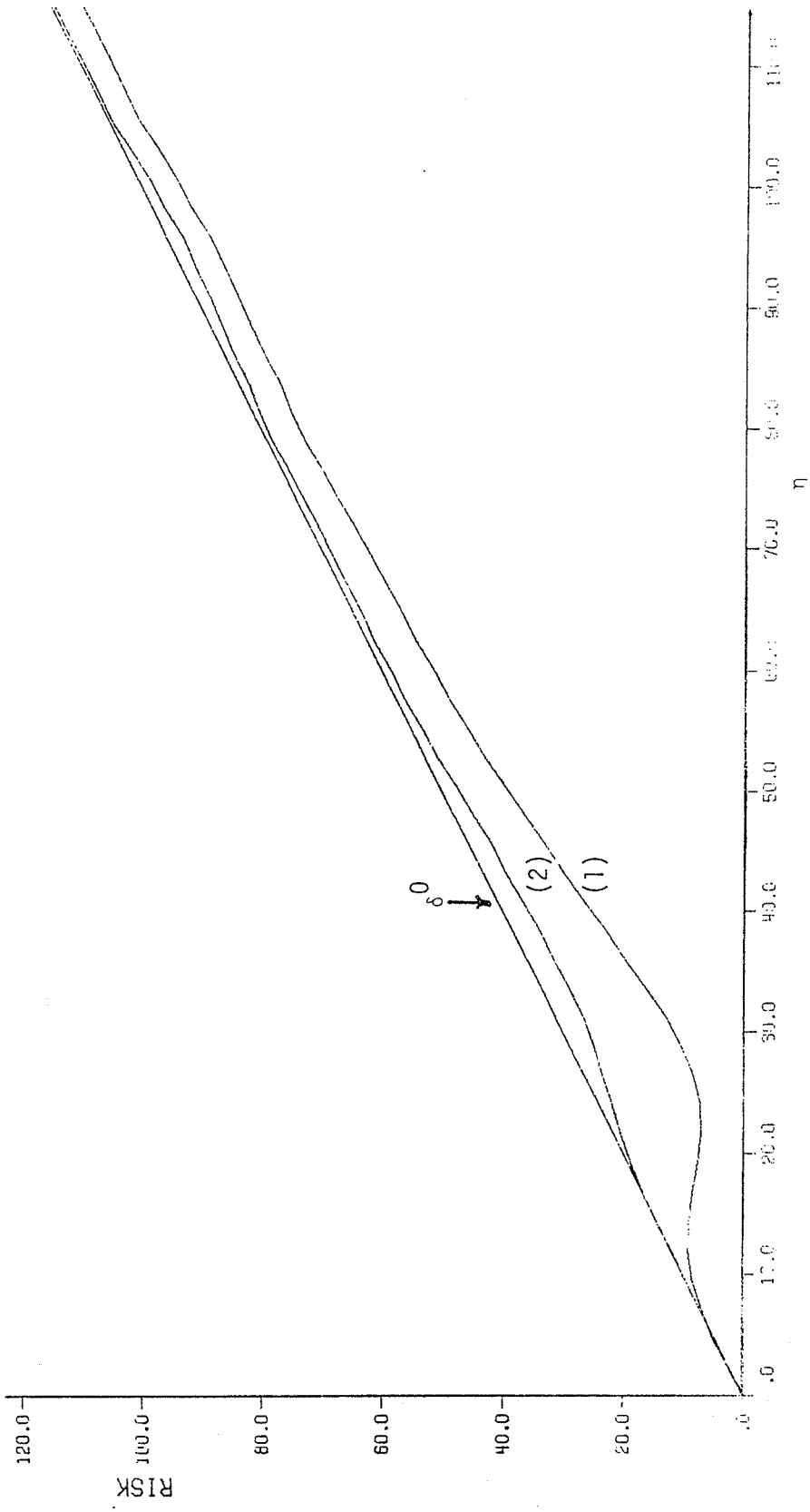


Figure 10

p=6. Risks of (1) δ^*, $(\mu_i,\beta_i) = (4,0)$ for all i, and (2) δ^T, K=4, $\mu'=1$ along line $\lambda_1=\ldots=\lambda_6=\eta$.

is able to accept different prior means, and although it is not minimax, it does improve upon the MVUE significantly in a prior region.

3.2.3. Numerical studies - Bayes risk

In this section δ^* is compared with δ^0 and the Bayes estimator δ^B with regards to Bayes risk. In Table 2, three cases are considered: (i) $p = 1$ with prior information $\mu = 5$ and $\beta = 2.5$, (ii) $p = 2$ with $(\mu_1,\beta_1) = (5, 2.5)$, $(\mu_2,\beta_2) = (8,1)$, and (iii) $p = 6$ with $(\mu_i,\beta_i) = (10,5)$, $i = 1,2,3$ and $(\mu_i,\beta_i) = (5,1)$, $i = 4,5,6$. These examples will indicate that supplying the correct prior information in the estimator δ^* leads to a significant reduction of the Bayes risk as compared to the Bayes risk of δ^0. The Bayes risks of δ^* are found through simulation and the values presented have a standard error of approximately 5 per cent.

The $p = 6$ case is typical of the behavior of δ^* in all three cases. In this case, the Bayes risk of δ^0 is $\sum_{i=1}^{6} \mu_i = 45$ and the Bayes risk of δ^* when the correct prior information has been used is 37.4, a substantial improvement. When the prior means $\mu_1,\ldots,\mu_6$ in δ^* are increased by factors of two and five, the Bayes risk increases to 45.5 and 50.5 respectively, showing that one is penalized significantly when $\mu_1,\ldots,\mu_p$ are chosen far from their true values. On the other hand, the selection of $\beta_1,\ldots,\beta_p$ does not appear to play a great role in the estimator δ^*. The smallest Bayes risk occurs when $\beta_1,\ldots,\beta_p$ are chosen correctly, but when they are

Table 2

Bayes risks of δ^B and δ^* with properly and improperly specified prior information.

p=1. Parameters in prior: $\mu=5$, $\beta=2.5$.

Parameters in estimator		Bayes risk	
μ	β	δ^B	δ^*
5	2.5	3.57	4.11
10	2.5	5.60	4.83
20	2.5	22.0	5.93
50	2.5	165	5.33
5	5	3.80	4.18
5	10	4.23	4.38
5	20	4.56	4.61
5	1.25	4.00	4.14
5	0	12.5	4.21

p=2. Parameters in prior: $(\mu_1,\beta_1)=(5,2.5)$, $(\mu_2,\beta_2)=(8,1)$.

Parameters in estimator				Bayes risk	
μ_1	β_1	μ_2	β_2	δ^B	δ^*
5	2.5	8	1	7.6	9.8
10	2.5	8	1	9.6	10.4
20	2.5	8	1	26	11.6
50	2.5	8	1	170	10.9
5	5	8	1	7.8	9.8
5	10	8	1	8.2	10.0
5	20	8	1	8.5	10.2
5	1.25	8	1	7.9	9.8
5	0	8	1	16.4	9.9

Table 2 (continued)

p=6. Parameters in prior: $(\mu_i,\beta_i)=(10,5)$, i=1,2,3,
$(\mu_i,\beta_i)=(5,1)$, i=4,5,6.

Parameters in estimator	Bayes risk	
	δ^B	δ^*
$(\mu_i,\beta_i)=(10,5)$, i=1,2,3 $(\mu_i,\beta_i)=(5,1)$, i=4,5,6	32.5	37.4
$(\mu_i,\beta_i)=(20,5)$, i=1,2,3 $(\mu_i,\beta_i)=(10,1)$, i=4,5,6	69.6	45.5
$(\mu_i,\beta_i)=(50,5)$, i=1,2,3 $(\mu_i,\beta_i)=(25,1)$, i=4,5,6	465.7	50.5
$(\mu_i,\beta_i)=(10,10)$, i=1,2,3 $(\mu_i,\beta_i)=(5,2)$, i=4,5,6	34.4	37.8
$(\mu_i,\beta_i)=(10,25)$, i=1,2,3 $(\mu_i,\beta_i)=(5,5)$, i=4,5,6	38.8	39.8
$(\mu_i,\beta_i)=(10,2.5)$, i=1,2,3 $(\mu_i,\beta_i)=(5,.5)$, i=4,5,6	35.9	37.6
$(\mu_i,\beta_i)=(10,0)$, i=1,2,3 $(\mu_i,\beta_i)=(5,0)$, i=4,5,6	165.0	38.7

chosen to be five times their true values, the Bayes risk increases to 39.8, a slight increase.

On the basis of simulations like those in Table 2, δ^* appears to be sensitive to the input of prior information, especially the prior means, and given correct selection of prior means, offers significant improvement in Bayes risk over the MVUE δ^0. Table 2 will also be used in Section 3.3.1, where the robustness of δ^* with respect to incorrect prior information will be discussed.

3.2.4. Further analysis in the case of asymmetric selection of prior information

In the preceding sections, δ^* has been shown to perform well when the prior means and variances have been chosen to be the same. That is, the cases where $\mu_1 = \ldots = \mu_p$ and $\beta_1 = \ldots = \beta_p$ have primarily been considered. In this section, the nonsymmetric situation is investigated, in which different prior means or different prior variances are chosen for the p components. It would be desirable for δ^* to exhibit a substantial risk improvement over δ^0 in the prior region for a wide range of selection of parameters $\mu_1,\ldots,\mu_p$ and $\beta_1,\ldots,\beta_p$.

In Figure 11, the risks of three versions of δ^* with different prior information are plotted in the case p = 2. The risks are shown along the line $\lambda_1 = \lambda_2$. Note that δ^{*1} with prior information $(\mu_1,\beta_1) = (\mu_2,\beta_2) = (4,4)$ performs best along this line. This is expected since the line $\lambda_1 = \lambda_2$ passes through the point (4,4), and it will be shown later that δ^{*1} performs well asymptotically along this line. In fact, in Section 3.3.2, it is shown that δ^{*1} has the

Figure 11

p=2. Risks of (1) δ^{*1}, $(\mu_i,\beta_i)=(4,4)$, i=1,2, (2) δ^{*2},$(\mu_1,\beta_1)=(0,4)$, $(\mu_2,\beta_2)=(8,4)$, and (3) δ^{*3}, $(\mu_1,\beta_1)=(4,0)$, $(\mu_2.\beta_2)=(4,8)$ along line $\lambda_1=\lambda_2=\eta$.

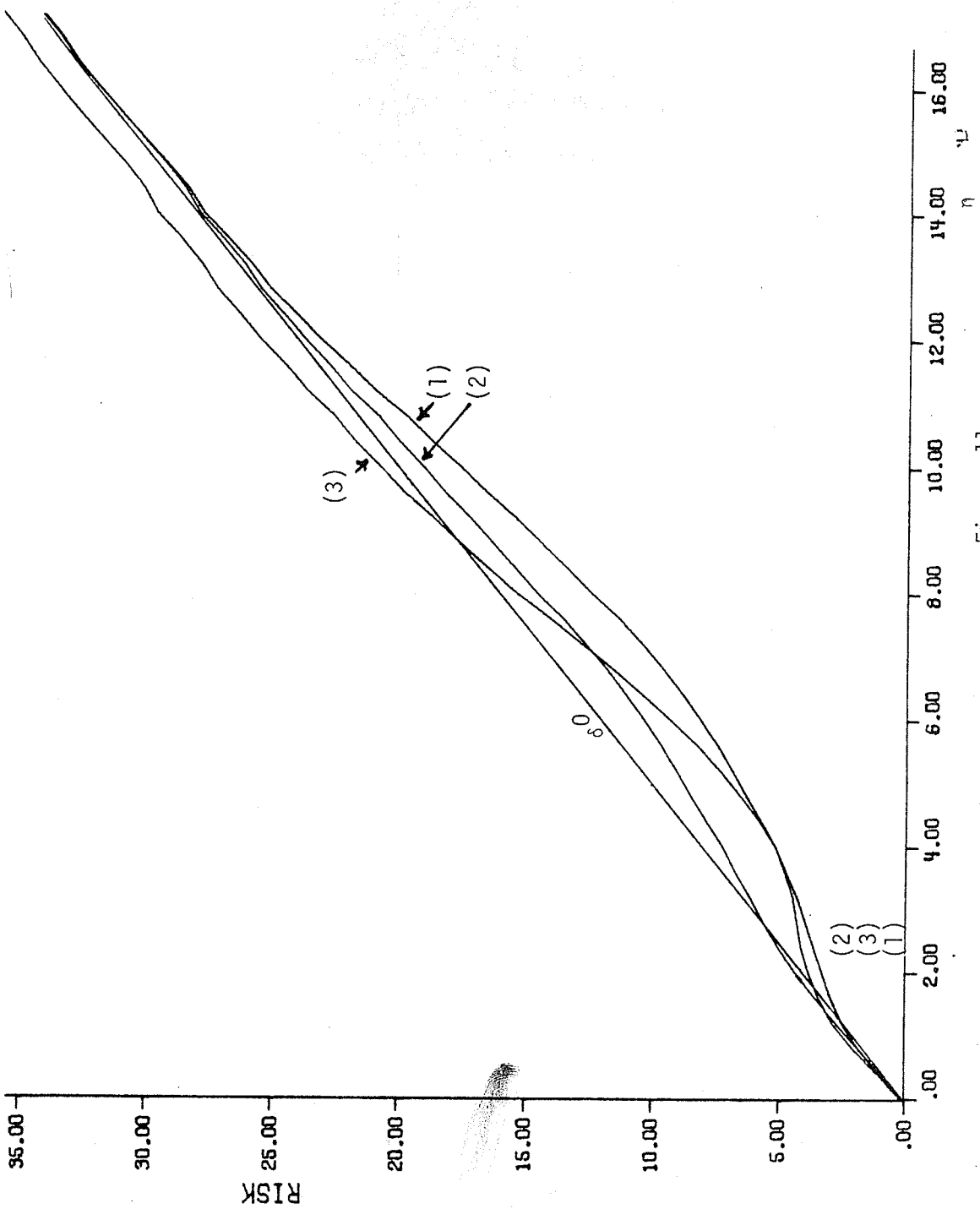

RISK
35.00
30.00
25.00
20.00
15.00
10.00
5.00
.00
.00
2.00
4.00
6.00
8.00
10.00
12.00
14.00
16.00
η
(1)
(2)
(3)
(2)
(3)
(1)
δ_0

Figure 11

same asymptotic (as $\lambda_1 = \lambda_2 \to \infty$) risk as δ^0 along the line $\lambda_1/(\beta_1+1) = \lambda_2/(\beta_2+1)$.

The second estimator δ^{*2} differs from δ^{*1} in that the former shrinks toward $\mu_1 = 0$ and $\mu_2 = 8$. The line $\lambda_1 = \lambda_2$ does not pass through the region where δ^{*2} shows the greatest improvement over δ^0, but δ^{*2} still shows some improvement in a large interval. One notes that δ^{*1} and δ^{*2} appear to be equivalent outside of the prior region. This is expected since β_1 and β_2 have not been changed and therefore, from the above remark, the asymptotic risk of δ^{*2} will be the same as that of δ^{*1}.

The estimator δ^{*3} differs from δ^{*1} in that the estimator uses $\beta_1 = 0$ and $\beta_2 = 8$. This causes δ^{*3} to obtain the same amount of improvement as δ^{*1} at the point (4,4), but the area of improvement is decreased significantly, and asymptotically outside of the prior region, δ^{*3} appears to have a risk one unit larger than the risk of δ^0. One explanation for this risk behavior is that the term

$$\frac{\sum_{i=1}^{2} X_i/(\beta_i+1)}{\sum_{i=1}^{2} X_i/(\beta_i+1)^2 + \sum_{i=1}^{2} ((X_i-\mu_i)/(\beta_i+1))^2} \cong \frac{X_1}{X_1+(X_1-4)^2},$$

due to the difference between β_1 and β_2, and thus δ^{*3} is approximately a one dimensional estimator. It will be shown that δ^* for $p = 1$ loses asymptotically one unit in risk compared to δ^0 outside of the prior region, and δ^{*3} displays a similar asymptotic behavior for p=2.

Let us investigate further the influence of prior information on the behavior of the estimator in the prior region. It was mentioned earlier that for large p and correct prior information, the shrinking constant of $\delta_i^*(X)$, $1-c_i^*(X)$, is approximately equal to

$$1 - \frac{1}{\beta_i+1} \frac{\sum_{j=1}^{p} \mu_j/(\beta_j+1)}{\sum_{j=1}^{p} \mu_j/(\beta_j+1)^2 + \sum_{j=1}^{p} \mu_j/(\beta_j+1)} .$$

It can happen that this shrinking constant is dominated by one set of prior components (μ_i,β_i) if the corresponding ratio, $\mu_i/(\beta_i+1)$, is very large. For example, if p = 3, $(\mu_1,\beta_1) = (10,1)$ and $(\mu_i,\beta_i) = (4,4)$ for i = 2,3, then $\mu_1/(\beta_1+1) = 5$, $\mu_i/(\beta_i+1) = .8$, i = 2,3, and the first set of prior components would have the greatest influence on the shrinking constant $1-c^*(X)$. When one set of prior components dominates the shrinking constant, δ^* appears to be performing like a one dimensional estimator in the prior region. In general, therefore, it appears that one should be concerned about nonsymmetric situations in which significant differences in the prior parameters exist. In Section 4, it will be shown that there do exist many nonsymmetric situations in which $\mu_i/(\beta_i+1)$ is approximately a constant for all i, and the above problem does not occur.

3.3. Robustness study

We now begin our study of δ^* with respect to robustness. First, in Section 3.3.1, the Bayes risk of δ^* is evaluated when the prior information has been misspecified. The remaining sections use risk as

a criterion. We evaluate the robustness of δ^* by comparing its risk to the risk of the MVUE for parameter values outside of the prior region.

3.3.1. Numerical studies - Bayes risk

With respect to Bayes risk, Table 2 shows the robustness of δ^* with regard to misspecification of the prior information. In the $p = 6$ example, when the prior means are chosen five times their true values, the Bayes risk of δ^* is 50.5, compared to the MVUE Bayes risk which is equal to 45. This is of sharp contrast to the risk of the Bayes estimator δ^B, which is 465.7 using the same prior information. It has already been remarked that δ^* is not very sensitive to the proper selection of $\beta_1,\ldots,\beta_p$ in the estimator; δ^* will offer substantial improvement over δ^0 even if $\beta_1,\ldots,\beta_p$ are chosen far from their true values. In contrast, the Bayes risk of δ^B increases rapidly if $\beta_1,\ldots,\beta_p$ are chosen much smaller than their true values.

3.3.2. Behavior of the estimator for large λ

In the numerical studies of Section 3.2, the MVUE δ^0 was compared with δ^* in and around the prior region for moderate values of the parameters $\lambda_1,\ldots,\lambda_p$. One situation that has not been considered is that in which $\lambda_1,\ldots,\lambda_p$ are large and far outside of the prior region. In Theorem 1, we consider the case in which the prior means $\mu_1,\ldots,\mu_p$ are fixed and $\lambda_1,\ldots,\lambda_p$ go to infinity along the line described by $\lambda_i = k_i\eta$, $i = 1,\ldots,p$, and the asymptotic risk improvement of δ^* over δ^0 is given.

Theorem 1

$$\text{Let } \delta_i^*(X) = \mu_i + \left(1 - \frac{1}{\beta_i+1} \min\left\{1, \frac{\sum_{j=1}^{p} X_j/(\beta_j+1)}{\sum_{j=1}^{p} X_j/(\beta_j+1)^2 + \sum_{j=1}^{p} ((X_j-\mu_j)/(\beta_j+1))^2}\right\}\right)$$

$$\cdot(X_i-\mu_i), \quad i = 1,\ldots,p.$$

Let $\lambda_i = k_i\eta$, $k_i > 0$, $i = 1,\ldots,p$. Then the asymptotic improvement as $\eta \to \infty$ of δ^* over δ^0 is

$$\lim_{\eta\to\infty} [R(\delta^0,\lambda)-R(\delta^*,\lambda)] =$$

$$\frac{[\sum_{j=1}^{p} k_j/(\beta_j+1)]^2}{\sum_{j=1}^{p} (k_j/(\beta_j+1))^2} - \frac{4\sum_{j=1}^{p} k_j/(\beta_j+1) \sum_{j=1}^{p} (k_j/(\beta_j+1))^3}{[\sum_{j=1}^{p} (k_j/(\beta_j+1))^2]^2} + 2.$$

Proof: See Appendix.

Consider the asymptotic improvement of δ^* over δ^0 given in Theorem 1. The risk improvement has been shown to be of the order of a constant for large $\lambda_1,\ldots,\lambda_p$, while the risk of δ^0 is $\sum_{j=1}^{p} \lambda_j$. Thus this risk improvement is insignificant compared to the risk of δ^0 for large $\lambda_1,\ldots,\lambda_p$. Note next that when $k_1/(\beta_1+1) = \ldots = k_p/(\beta_p+1)$, one can calculate $I = p-2$. Thus when many means are estimated simultaneously, δ^* can display smaller risk than δ^0 far outside of the prior region along particular lines.

In this asymptotic setting, it is of interest to investigate how poorly δ^* can perform relative to δ^0. Let $b_i = k_i/(\beta_i+1)$ for

$i = 1,\ldots,p$ and note without loss of generality that one can take $\sum_{j=1}^{p} b_j^2 = 1$. Then the asymptotic improvement in risk equals

$$I = -4 \sum_{j=1}^{p} b_j^3 \sum_{j=1}^{p} b_j + \left(\sum_{j=1}^{p} b_j \right)^2 + 2$$

$$\geq -4 \sum_{j=1}^{p} b_j + \left(\sum_{j=1}^{p} b_j \right)^2 + 2,$$

since $\sum_{j=1}^{p} b_j^2 = 1$ implies $\sum_{j=1}^{p} b_j^3 \leq 1$.

Now the last expression is a quadratic in $\sum_{j=1}^{p} b_j$ and achieves a minimum value of -2 at $\sum_{j=1}^{p} b_j = 2$. Thus

$$I \geq -2.$$

Therefore δ^* can not lose any more than two units of risk asymptotically compared to δ^0. Through computer simulation studies, it appears that δ^* will do worst asymptotically along lines very close to the edges of the parameter space. Along these lines, all but one of $\lambda_1,\ldots,\lambda_p$ are close to zero, and δ^* is performing much like a one dimensional estimator. Other ways will be described later in which the one dimensional case is a worst case for the estimator δ^*.

As an example to illustrate the robustness of δ^* compared to δ^B, consider the case $p = 1$ where the prior information is $\mu = 6$, $\beta = 2$. Figure 12 shows the risk functions of δ^* and δ^B in this situation. Note that the prior standard deviation is $(\mu\beta)^{1/2} = (12)^{1/2} \doteq 3.5$. One observes that the risk of δ^B is substantially

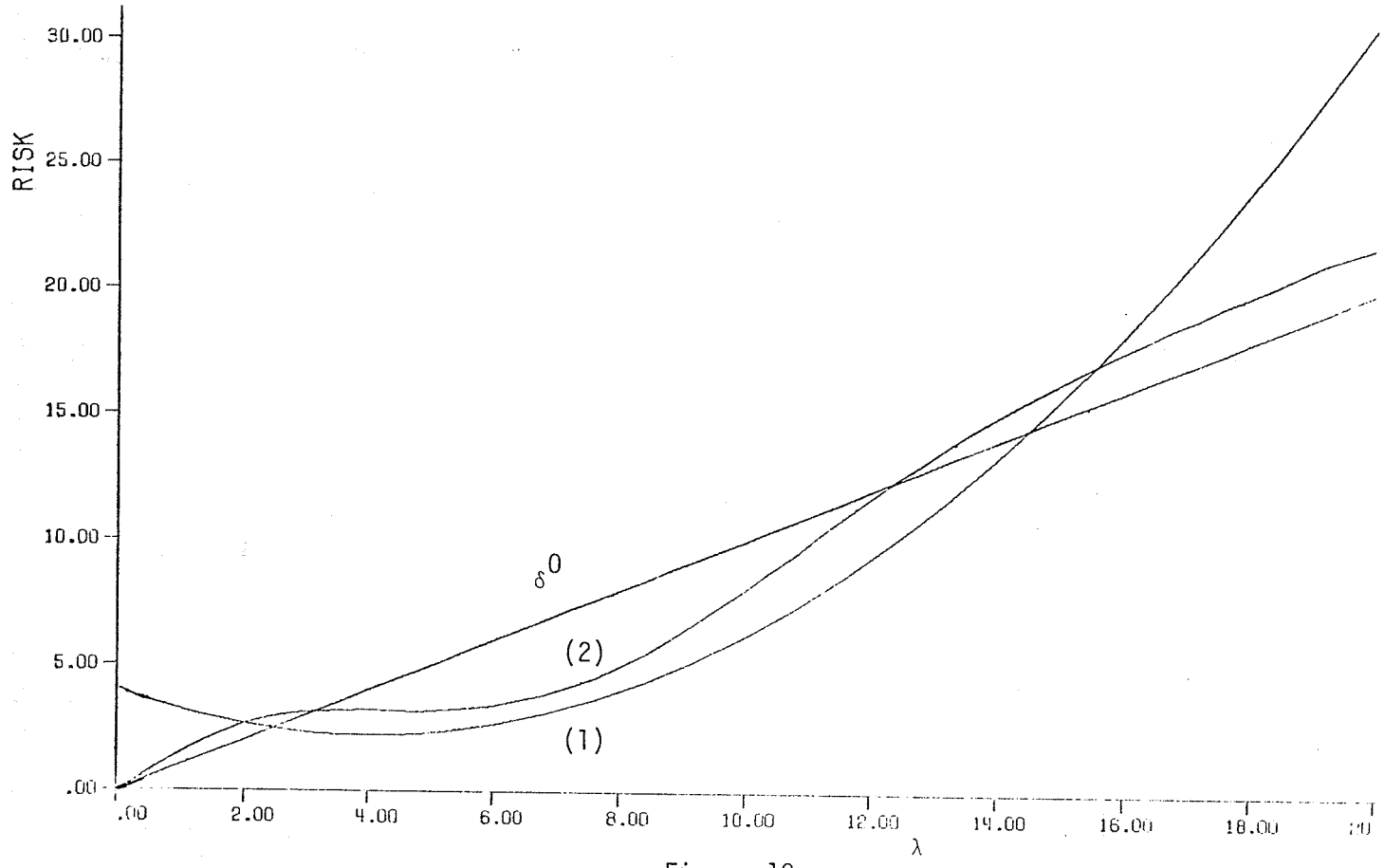


Figure 12

p=1. Risks of (1) δ^B, $(\mu,\beta)=(6,2)$, and (2) δ^*, $(\mu,\beta)=(6,2)$.

greater than the risk of δ^0 for $\lambda = 18$, only a few standard deviations away from the prior mean, and as λ increases the risk performance of δ^B becomes worse. The risk of δ^* can get larger than the risk of δ^0 outside of the prior region, but the amount of risk decrement appears to be bounded and approaches one as λ gets large. Clearly δ^* is a safer estimator to use than δ^B in the situation where parameter values of 18 or greater are likely.

3.3.3. Behavior of the estimator for large prior means

In the previous section, the situation was considered where the parameters $\lambda_1,\ldots,\lambda_p$ were far outside of the prior region, and the corresponding risk improvement of δ^* over δ^0 was found. Here the situation of using large prior means is considered. We are primarily interested in evaluating how poorly δ^* can perform (in terms of risk) relative to δ^0 outside of the improvement region.

From observing the risk curves of δ^* and δ^0 (for example, see Figure 1), one notes that compared to the risk of δ^0, the risk of δ^* appears to be worst just outside of the improvement region. As λ moves further away from the improvement region, the risk of δ^* appears to approach the risk of δ^0. In other words, the worst situation appears to occur a few prior standard deviations from the prior mean. Also note from Figure 1 that increasing the prior value of β seems to have the effect of flattening the risk function of δ^* towards the risk function of δ^0. Therefore for a given set of prior means $(\mu_1,\ldots,\mu_p)$, it appears likely that the maximum decrement in risk of δ^* relative to δ^0 is achieved when $\beta_1 = \ldots = \beta_p = 0$. In this case

$$\delta_i^*(X) = \mu_i + \left(1 - \frac{\sum_{j=1}^{p} X_j}{\sum_{j=1}^{p} X_j + \sum_{j=1}^{p} (X_j-\mu_j)^2}\right)(X_i-\mu_i), \quad i = 1,\ldots,p.$$

In this situation, when $\beta_1 = \ldots = \beta_p = 0$, it is no longer meaningful to talk about prior standard deviations, since the prior standard deviation of λ_i is $\mu_i\beta_i = 0$. But from our experience with risk curves of δ^* for $p = 1$, $\beta = 0$, and different values of μ, it appears that the worst situation for δ^* occurs at approximately a distance of $3\lambda^{1/2}$ from the prior mean. That is, if the proportional improvement in risk of δ^* over δ^0 is defined by

$$\rho^* = \frac{R(\delta^0,\lambda) - R(\delta^*,\lambda)}{R(\delta^0,\lambda)},$$

then in this situation ($p = 1$), ρ^* achieves approximately a minimum value at the points $\mu + 3\lambda^{1/2}$ and $\mu - 3\lambda^{1/2}$.

Let us consider this worst situation for δ^* when large prior means $\mu_1,\ldots,\mu_p$ are selected. This is one situation which has not been considered in the simulation work of Section 3.2. In the case where the prior means are going to infinity, let $\lambda_1,\ldots,\lambda_p$ also go to infinity, since we are interested in the performance of δ^* about the prior region. In particular since the risks of δ^0 and δ^* will be compared within standard deviations of the prior mean, let the μ_i's and λ_i's go to infinity such that $\mu_i - \lambda_i = O(\lambda_i^{1/2})$. This asymptotic setup will allow us to compare the risks of the two estimators in the region where δ^* is expected to do worst compared to δ^0. Theorem 2

gives the asymptotic value of ρ^* in this limiting situation, when $\lambda_1,\ldots,\lambda_p$ increase to infinity along a line from the origin.

Theorem 2

$$\text{Let } \delta_i^*(X) = X_i - \frac{(X_i-\mu_i)\sum_{j=1}^{p} X_j}{\sum_{j=1}^{p} X_j + \sum_{j=1}^{p}(X_j-\mu_j)^2}, \qquad i = 1,\ldots,p.$$

Let $\lambda_i = k_i\eta$, $i = 1,\ldots,p$ where $\sum_{i=1}^{p} k_i = 1$.

Let $\theta_i = \lambda_i - \mu_i$, and assume $\lim_{\eta\to\infty} \frac{\theta_i}{\eta^{1/2}} = \theta_i^*$, $i = 1,\ldots,p$. Then asymptotically, as $\eta \to \infty$,

$$(2.1)\quad \lim_{\eta\to\infty} \frac{R(\delta^0,\lambda)-R(\delta^*,\lambda)}{R(\delta^0,\lambda)} = E\Big[\sum_{i=1}^{p}\Big(2(W_i-\theta_i^*)\frac{W_i}{1+\sum_{j=1}^{p} W_j^2} - \Big(\frac{W_i}{1+\sum_{j=1}^{p} W_j^2}\Big)^2\Big)\Big],$$

where $W_i \sim N(\theta_i^*,k_i)$, $i = 1,\ldots,p$, and $W_1,\ldots,W_p$ are independent.

Proof: See Appendix.

In applying this theorem, it is useful to consider the normal estimation problem where $W_1,\ldots,W_p$ are independent with $W_i \sim N(\theta_i^*,k_i)$, $i = 1,\ldots,p$, and the vector $(\theta_1^*,\ldots,\theta_p^*)$ is to be estimated. Consider the Stein-type estimator

$$\hat{\delta}(W) = \Big(1 - \frac{1}{1+\sum_{j=1}^{p} W_j^2}\Big)W,$$

where $W = (W_1,\ldots,W_p)$, and assume $\hat{\delta}$ is to be compared with the MVUE $\delta^0(W) = W$ using the loss $\sum_{i=1}^{p}(\delta_i-\theta_i^*)^2$. The right hand side of (2.1),

$$E[\sum_{i=1}^{p}(2(W_i-\theta_i^*)\frac{W_i}{1+\sum_{j=1}^{p}W_j^2}-(\frac{W_i}{1+\sum_{j=1}^{p}W_j^2})^2)],$$

is the risk improvement of $\hat{\delta}$ upon δ^0. Thus in the asymptotic situation of Theorem 2, the proportional improvement of the Poisson estimator δ^* is equivalent to the improvement of the normal estimator $\hat{\delta}$. It is of interest to find the maximum proportional risk decrement of δ^*, or equivalently, the maximum risk decrement of $\hat{\delta}$.

To heuristically find the maximum risk decrement of $\hat{\delta}$, a lemma due to Stein will first be stated. This lemma allows us to obtain an unbiased estimator of the improvement in risk of an estimator of the form $W + f(W)$ over W in the normal situation.

Lemma: Let $W_1,\ldots,W_p$ be independent with $W_i \sim N(\theta_i,\sigma_i^2)$, $i = 1,\ldots,p$. Then for functions $f_1,\ldots,f_p$: $\mathbb{R}^p \to \mathbb{R}$ satisfying

$$E\{|\frac{\partial f_i}{\partial W_i}(W)|\} < \infty, \quad i = 1,\ldots,p,$$

it follows that

$$E\{(W_i-\theta_i)f_i(W)\} = E\{\sigma_i^2\frac{\partial f_i}{\partial W_i}\}, \quad i = 1,\ldots,p.$$

(The proof of this lemma in the case $\sigma_1^2 = \ldots = \sigma_p^2$ can be found in Hudson (1974).) Using this lemma, the improvement of the estimator $W+f(W)$ over W is

$$I = E[\sum_{i=1}^{p}(W_i-\theta_i)^2] - E[\sum_{i=1}^{p}(W_i+f_i(W)-\theta_i)^2]$$

$$= E[-2\sum_{i=1}^{p}(W_i-\theta_i)f_i(W) - \sum_{i=1}^{p}f_i(W)^2]$$

$$= E[-2\sum_{i=1}^{p}\sigma_i^2\frac{\partial f_i}{\partial W_i} - \sum_{i=1}^{p}f_i(W)^2],$$

assuming the conditions of the lemma are satisfied. For the estimator $\hat{\delta}$,

$$\hat{\delta}_i(W) = W_i - \frac{W_i}{1+\sum_{j=1}^{p}W_j^2}, \quad \sigma_i^2 = k_i, \quad i = 1,\ldots,p,$$

so that

$$\frac{\partial f_i(W)}{\partial W_i} = \frac{-(1+\sum_{j=1}^{p}W_j^2)+2W_i^2}{(1+\sum_{j=1}^{p}W_j^2)^2}.$$

Clearly the conditions of the lemma are satisfied and hence

$$I = E[2(\sum_{i=1}^{p}[\frac{\sigma_i^2}{1+\sum_{j=1}^{p}W_j^2} - \frac{2W_i^2\sigma_i^2}{(1+\sum_{j=1}^{p}W_j^2)^2}]) - \sum_{i=1}^{p}\frac{W_i^2}{(1+\sum_{j=1}^{p}W_j^2)^2}]$$

$$= E[\frac{1}{(1+\sum_{j=1}^{p}W_j^2)^2}(2-4\sum_{i=1}^{p}\sigma_i^2W_i^2 + \sum_{i=1}^{p}W_i^2)].$$

We would like to minimize I with respect to the parameters $\theta_1^*,\ldots,\theta_p^*$ and $\sigma_1^2,\ldots,\sigma_p^2$ (to find the "worst case"). This seems extremely difficult analytically, but we can argue heuristically as

follows. If $\sum_{j=1}^{p} W_j^2 = K$, then the expression inside the expectation is minimized when $W_m^2 = K$, where $\sigma_m^2 = \max_i \sigma_i^2$. Since we have the restriction, $\sum_{i=1}^{p} \sigma_i^2 = \sum_{i=1}^{p} k_i = 1$, it seems plausible that the expression is minimized when $\sigma_m^2 = 1$ and $\sigma_i^2 = 0$, $i \neq m$. In this case, I is the improvement of the one dimensional estimator

$$\tilde{\delta}(W) = (1 - \frac{1}{1+W^2})W$$

over the estimator $\delta^0(W) = W$ in the situation where $W \sim N(\theta^*,1)$.

Figure 13 shows the risk of $\tilde{\delta}(W)$ plotted as a function of θ^*; equivalently it shows the asymptotic proportional risk of δ^*, $R(\delta^*,\theta^*)/R(\delta^0,\theta^*)$, when $p = 1$ and $\beta = 0$. The constant line represents the risk of the estimator $\delta^0(W) = W$ in the normal problem. Note that

$$\max_{\theta^*} R(\tilde{\delta},\theta^*) = 1.27,$$

and

$$\min_{\theta^*} R(\tilde{\delta},\theta^*) = .48.$$

Relating this to the Poisson estimation problem, this indicates that asymptotically under the conditions of Theorem 2, the decrement in risk of δ^* can be no larger than 27% of the risk of the MVUE for all values of p.

Note also from Figure 13 that δ^* is much more robust than δ^B to misspecified prior information in the situation where large prior means are used. This figure shows the asymptotic proportional risk of δ^B plotted as a function of θ^* for $p = 1$. We have set $\beta = 1$ in

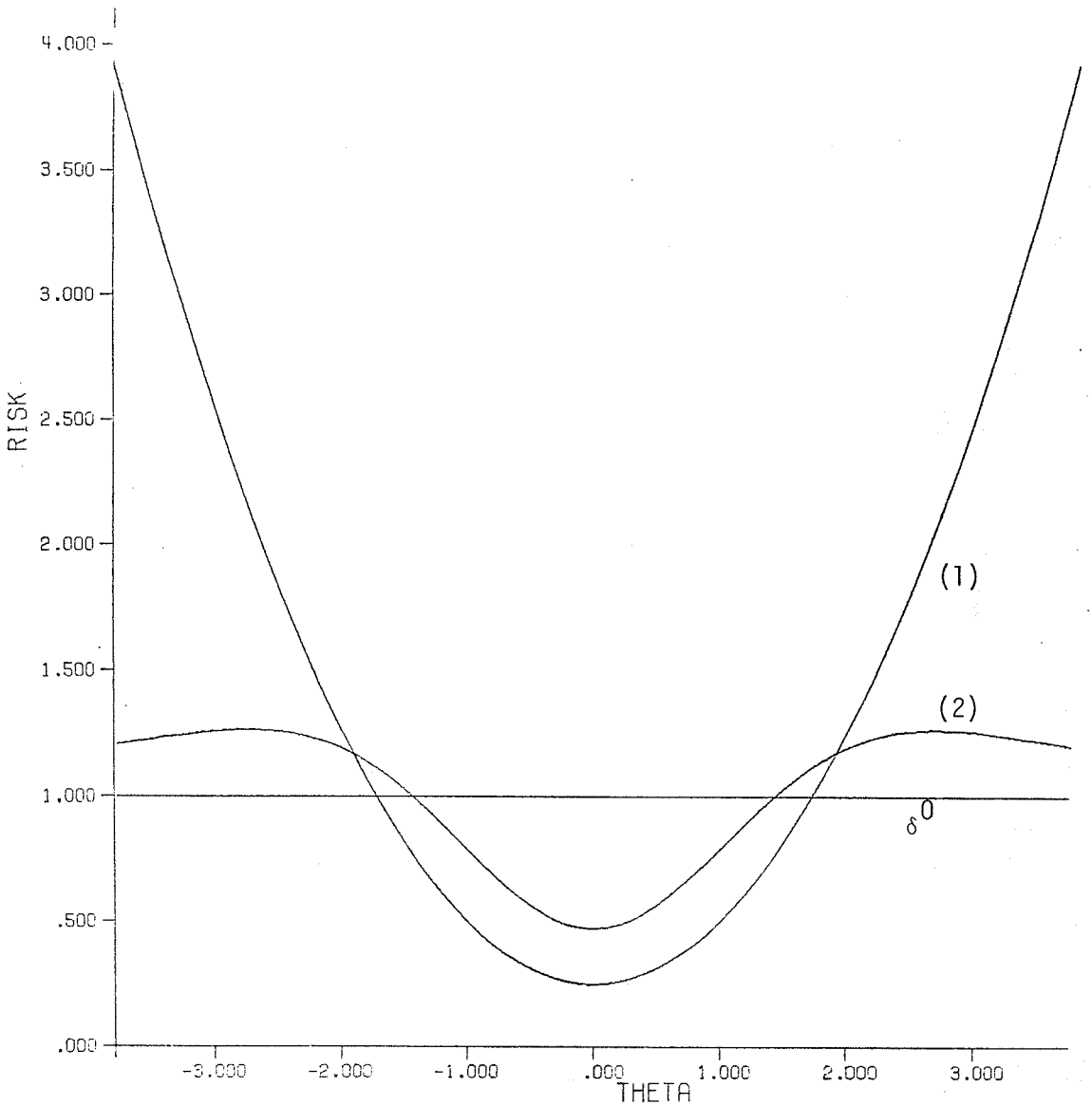


Figure 13

p=1. Proportional risks of (1) δ^B and (2) δ^* in limiting situation of Theorem 2.

the Bayes estimator δ^B so that δ^* and δ^B will perform similarly in the prior region. For large λ,

$$\theta^* \tilde{=} (\lambda-\mu)\lambda^{-1/2},$$

so θ^* represents the number of sample standard deviations that λ differs from the prior mean. Although both δ^* and δ^B display risk improvement over δ^0 near the prior mean ($\theta^* = 0$), outside of the improvement region the proportional risk of δ^B rapidly increases, while δ^* has a proportional risk bounded above by 1.27. This graph further demonstrates that δ^* is more robust than δ^B with respect to parameter values that occur outside of the prior region.

In summary, the situation of using large prior means has been considered. Our main concern was to find the situation involving large prior means in which δ^* performed the worst compared to δ^0. The proportional risk improvement of δ^* relative to δ^0 was defined, and this was heuristically shown to be minimized when only one parameter is estimated ($p = 1$), and the parameter β is set equal to zero. In this case, δ^* can possess a risk 27% larger than the risk of δ^0 outside of the prior region. It was shown that this "worst case" is much better than what can happen when δ^B, the conjugate Bayes estimator, is used with a very large prior mean. Finally the results in this section will be used to analyze thoroughly the risk of δ^* for $p = 1$.

3.3.4. Further analysis in the one dimensional case

In this section, the risk of δ^* is further discussed when $p = 1$, that is, when one Poisson mean is estimated. This case is of interest for two reasons. First we want to evaluate δ^* with respect to robustness to wrong prior information, and $p = 1$ appears to correspond to a worst case for the estimator. Second, it is reasonable to use δ^* in problems of estimating just one Poisson mean, and a complete investigation of this most common situation is desirable. Applications of Theorem 2 to the one dimensional estimator will be discussed, and the risk of δ^* for small values of λ will be evaluated.

Figure 13 shows the asymptotic proportional risk of δ^* when $\beta = 0$ and the prior mean μ and λ go to infinity at the rate described in Theorem 2. We are now interested in how large μ must be for the asymptotic risk behavior to be approximately valid. Figure 14 shows the proportional risk (calculated numerically), as a function of $(\lambda-\mu)\lambda^{-1/2}$ for prior mean values of 10, 30 and 50 and $\beta = 0$. One notes that when $\mu = 50$, the plot of the proportional risk is very similar to Figure 13. Thus for prior mean values of 50 or greater, the proportional improvement given in Theorem 2 seems to be a good approximation to the true proportional improvement.

What happens to the risk of δ^* when smaller means than 50 are chosen? Two comments can be made from observing Figure 14. First, the maximum proportional decrement in risk in the region where $(\lambda-\mu)\lambda^{-1/2} > 0$ increases as μ increases from 10 to 50. This suggests that the maximum proportional decrement in this region is bounded by the asymptotic amount, .27. In the region where $(\lambda-\mu)\lambda^{-1/2} < 0$, the

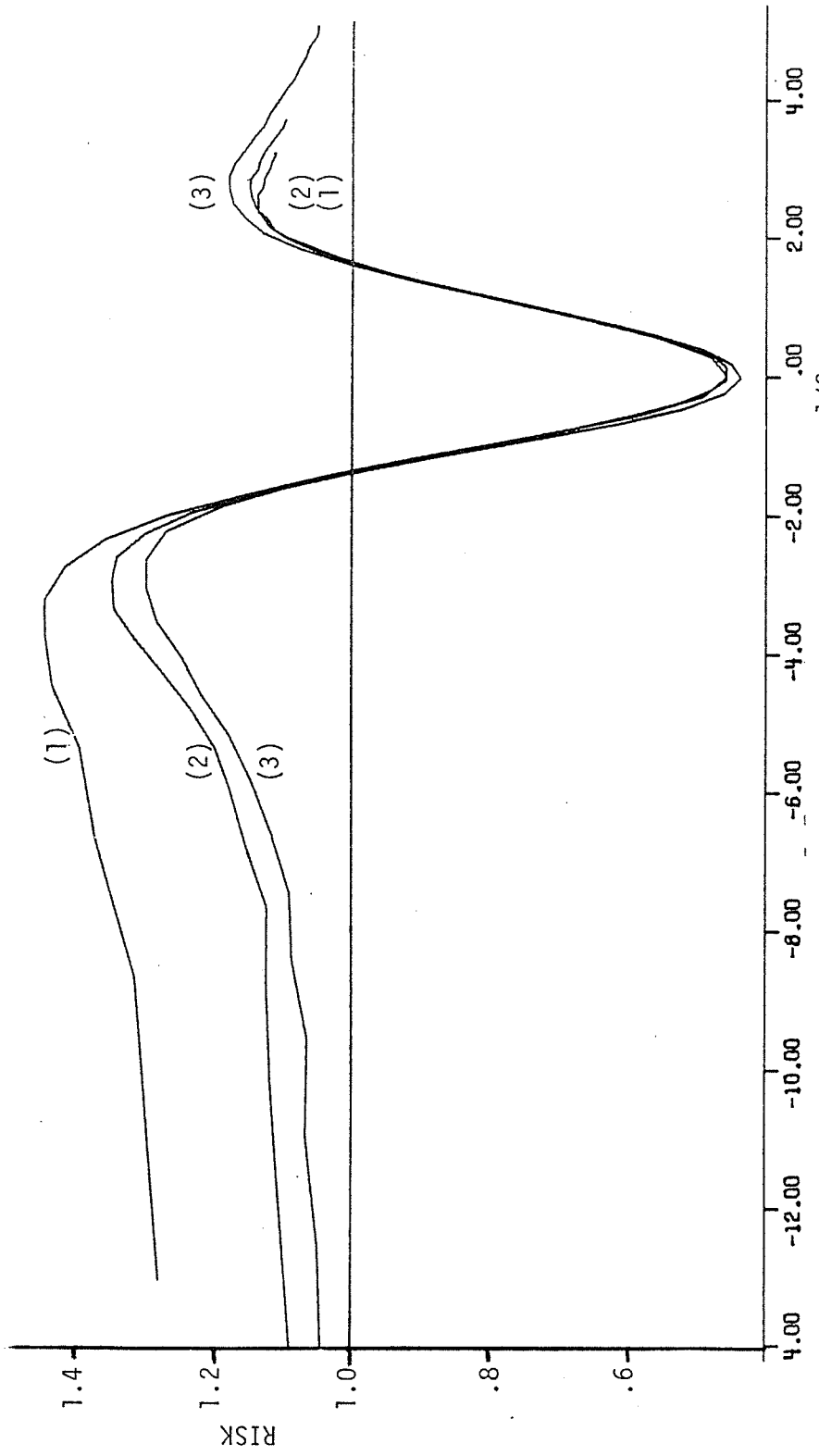


Figure 14

p=1. Proportional risks of (1) δ^*, $(\mu,\beta)=(10,0)$, (2) δ^*, $(\mu,\beta)=(30,0)$, and (3) δ^*, $(\mu,\beta)=(50,0)$.

proportional decrement appears to decrease as μ increases. In fact, the maximum proportional decrement when $\mu = 10$ is .4. This suggests that the worst case for the estimator δ^* in terms of maximum proportional decrement occurs as λ approaches zero.

For the one-dimensional estimator, we find the set of prior parameters (μ,β) which give the maximum proportional decrement in risk of δ^* as λ approaches zero. The proportional decrement in risk is

$$\frac{R(\delta^*,\lambda)-R(\delta^0,\lambda)}{R(\delta^0,\lambda)} = \frac{R(\delta^*,\lambda)}{\lambda} - 1$$

$$= \frac{E(\delta^*(X)-\lambda)^2}{\lambda} - 1.$$

Consider

$$\lim_{\lambda\to 0} \frac{E(\delta^*(X)-\lambda)^2}{\lambda},$$

which is also the limiting risk as λ approaches zero of δ^* under loss L_2. When $p = 1$,

$$\delta^*(X) = \mu+(1-\min\{\frac{1}{\beta+1}, \frac{X}{X+(X-\mu)^2}\})(X-\mu).$$

Now

$$E(\delta^*(X)-\lambda)^2 = (\delta^*(0)-\lambda)^2e^{-\lambda}+(\delta^*(1)-\lambda)^2\lambda e^{-\lambda}+o(\lambda)$$

$$= \lambda^2e^{-\lambda}+(\delta^*(1)-\lambda)^2\lambda e^{-\lambda}+o(\lambda).$$

Hence

$$\lim_{\lambda\to 0} \frac{E(\delta^*(X)-\lambda)^2}{\lambda} = \lim_{\lambda\to 0}[e^{-\lambda}(\lambda+\delta^*(1)^2-2\lambda\delta^*(1)+\lambda^2)]+o(1)$$

$$= \delta^*(1)^2.$$

Now

$$\delta^*(1)^2 = [1-\min\{\frac{1}{\beta+1}, \frac{1}{1+(1-\mu)^2}\}(1-\mu)]^2$$

$$= [1 - \frac{1-\mu}{\beta+1}]^2 \quad \text{if } (1-\mu)^2 \leq \beta$$

$$[1 - \frac{1-\mu}{1+(1-\mu)^2}]^2 \quad \text{if } (1-\mu)^2 > \beta.$$

To find the worst possible behavior of δ^* as λ approaches zero, we maximize $[\delta^*(1)]^2$ over all μ and β. Note:

(i) If $(1-\mu)^2 > \beta$, then through routine calculus,

$$\delta^*(1)^2 = [1 - \frac{1-\mu}{1+(1-\mu)^2}]^2$$

is maximized at $\mu = 2$, and

$$[1 - \frac{1-2}{1+(1-2)^2}]^2 = 2.25.$$

(ii) If $(1-\mu)^2 \leq \beta$ and $\mu \leq 1$, then

$$\delta^*(1)^2 = [1 - \frac{1-\mu}{\beta+1}]^2 \leq 1.$$

(iii) If $(1-\mu)^2 \leq \beta$ and $\mu > 1$, or equivalently $1 < \mu \leq 1+\beta^{1/2}$, then for any β,

$$\max_{\mu} \delta^*(1)^2 = \max_{\mu}[1 - \frac{1-\mu}{\beta+1}]^2$$

$$= \max_{\mu}[\frac{\beta+\mu}{\beta+1}]^2$$

$$= [\frac{\beta+1+\beta^{1/2}}{\beta+1}]^2,$$

and

$$\max_{\beta}\left[\frac{\beta+1+\beta^{1/2}}{\beta+1}\right]^2 = 2.25.$$

Therefore, $\max_{\mu,\beta}[\delta^*(1)]^2 = 2.25,$

and the maximum limiting proportional decrement in risk is 1.25. This means it is possible for δ^* to have risk that is 125% larger than the risk of δ^0 near $\lambda = 0$. Although the risk of δ^* can be much larger than the risk of δ^0 for small λ, the actual size of the risk decrement is small. In fact, it has already been shown numerically that the worst absolute risk decrement of δ^* is .27 λ. If the loss function $L_2(\delta,\lambda) = (\delta-\lambda)/\lambda$ is used instead of the squared error loss L_1 then one would be more concerned with errors made in estimating small λ. Under this loss, δ^* can perform poorly compared to δ^0 for given selection of μ and β.

It should be noted that for general p, one can find the limiting proportional risk decrement of δ^* as one approaches the origin on a particular line. As in the above work, it would be of interest to find the maximum proportional risk decrement over all selections of prior parameters $\mu_1,\ldots,\mu_p$ and $\beta_1,\ldots,\beta_p$. Unfortunately, it appears to be very cumbersome to perform the maximization for p larger than one.

4. Using the estimator

4.1. Choosing μ and β

It has been argued that δ^* is an attractive alternative estimator to δ^0 when certain vague prior information is available. In particular, for each Poisson mean estimated, one inputs two prior parameters μ_i and β_i. In this section, we discuss the type of prior information that is commonly known, and the ways of obtaining μ_i and β_i from this prior information.

The simplest way of obtaining μ and β is to guess at a mean and variance for each coordinate of λ. Since the prior mean and variance if λ_i are μ_i and $\mu_i\beta_i$ respectively, these guesses can be used to obtain μ_i and β_i. Unfortunately, although it may be easy to guess at a prior mean, a prior variance is harder to determine. Subtle characteristics of the prior distribution may greatly influence the variance, and as mentioned in Chapter 1, it is uncommon to have prior information concerning the tail.

As mentioned in Chapter 1, people can on the other hand specify fractiles of the prior distribution of λ_i or assign probabilities to particular areas of the parameter space. These assigned probabilities can lead to values of μ_i and β_i. For example, if the prior distribution of λ_i can be thought to be approximately normal in the central region, then μ_i and β_i can be calculated from the endpoints of an interval which is thought to contain a specific proportion of the prior distribution. If (a,b) is thought to be the interval which

contains the middle 50% of the distribution of λ_i, then by solving the equations

$$a = \mu_i - .675(\mu_i\beta_i)^{1/2}$$

$$b = \mu_i + .675(\mu_i\beta_i)^{1/2},$$

we can obtain the prior mean and standard deviation, μ_i and $(\mu_i\beta_i)^{1/2}$, and therefore β_i. In a similar fashion, values of μ_i and β_i may be obtained by fitting the central region of a gamma distribution to certain fractiles of the prior that are known. Although μ_i and $\mu_i\beta_i$ correspond to a prior mean and variance of λ_i, it is worth emphasizing that the prior information that is used is probabilities assigned to intervals of the parameter space. Prior knowledge of a specific form of the prior density or knowledge of the tails of the distribution are not necessary in order to use δ^*.

4.2. When to use δ^* in simultaneous estimation

It has been shown that δ^* possesses smaller risk than δ^0 in a prior region, and δ^* appears to perform best in this region when the quantities $\mu_1/(\beta_1+1),\ldots,\mu_p/(\beta_p+1)$ are similar in size. Also asymptotically outside of the prior region, δ^* has been shown to perform well (in terms of risk) compared to δ^0 along the line $\lambda_1/(\beta_1+1) = \ldots = \lambda_p/(\beta_p+1)$. When $\beta_1 = \ldots = \beta_p$, it appears that it is most appropriate to use δ^* in estimating means of a similar size. In this case, δ^* will be a superior estimator to the MVUE δ^0, which estimates the p means separately.

Let us consider a general setting of estimating p Poisson means where the use of the estimator δ^0 will be seen to be appropriate. Often, a Poisson mean is an approximation to the expected number of successes of a binomial random variable where N, the sample size, is large and θ, the probability of success is small. Thus p independent Poisson means frequently represent the expected number of successes for p independent binomial experiments. Let (N_i,θ_i) be the sample size and probability of success for the ith binomial experiment. The expected number of successes $(N_1\theta_1,\ldots,N_p\theta_p)$ will be estimated based on the observed number of successes in the corresponding samples $(X_1,\ldots,X_p)$.

An example of this situation was described by Zidek and Clevenson (1975). They discuss the problem of estimating mean numbers of oil well discoveries for different months simultaneously. Here the search for oil at a particular location can be considered a Bernoulli trial and searches at different locations may be assumed independent. Thus the mean numbers of discoveries are really the expected numbers of successes for different binomial experiments.

It has been stated that δ^* will perform best (in terms of risk) in the prior region when the prior parameters are such that $\mu_1/(\beta_1+1),\ldots,\mu_p/(\beta_p+1)$ are approximately equal. It will be shown that this is indeed frequently the case in the above situation, so that δ^* is frequently appropriate.

Often, the parameters $\theta_1,\ldots,\theta_p$ can be thought to come from a common prior distribution $\pi(\theta)$. Let ξ and σ^2 denote the prior mean and variance respectively of this prior. Note that

$$\frac{\mu_i}{\beta_i+1} = \frac{\mu_i^2}{\mu_i\beta_i+\mu_i} = \frac{[\text{prior variance}]^2}{\text{prior variance + prior mean}},$$

and in this binomial estimation problem, the prior mean and variance of $N_i\theta_i$ are $N_i\xi$ and $N_i^2\sigma^2$ respectively. Thus

$$\frac{\mu_i}{\beta_i+1} = \frac{N_i^2\xi^2}{N_i^2\sigma^2+N_i\xi}.$$

Now typically β_i is chosen larger than one for all i, and in this situation,

$$\frac{\xi^2}{2\sigma^2} < \frac{\mu_i}{\beta_i+1} < \frac{\xi^2}{\sigma^2}, \quad i = 1,\ldots,p.$$

Therefore, among the quantities $\mu_1/(\beta_1+1),\ldots,\mu_p/(\beta_p+1)$, it would be unlikely that one component would be much larger than the others. If indeed the β_i's are chosen large, then the above quantities would be approximately equal. In other words, when each prior variance $\mu_i\beta_i$ significantly exceeds the corresponding prior mean μ_i, then $\mu_i/(\beta_i+1)$ is approximately a constant for the above situation.

In the oil well example described above, it would be reasonable to have a common prior for the probabilities of finding wells during different months. From the above discussion, δ^* should thus perform well for many selections of the common prior.

5. Other topics

5.1. More than one observation taken from each population

In the above discussion, only one observation X_i was assumed taken from each Poisson distribution. Let us generalize to the case where

more than one observation is taken from each population. Suppose $(X_{i1},\ldots,X_{in_i})$ are observed from the ith Poisson distribution, $i = 1,\ldots,p$, and consider the sufficient statistics

$$Y_1 = \sum_{j=1}^{n_1} X_{1j},\ldots, Y_p = \sum_{j=1}^{n_p} X_{pj},$$

which are independently Poisson with means $n_1\lambda_1,\ldots,n_p\lambda_p$ respectively.

To estimate the vector $\lambda^* = (n_1\lambda_1,\ldots,n_p\lambda_p)$, first note that the prior mean and prior variance of $n_k\lambda_k$ are

$$E[n_k\lambda_k] = n_k\mu_k,$$

$$Var[n_k\lambda_k] = n_k^2\mu_k\beta_k = (n_k\mu_k)(n_k\beta_k),$$

and δ^* in this situation is defined componentwise by

$$\delta_i^*(Y) = n_i\mu_i + (1 - \frac{1}{n_i\beta_i+1}\, d(Y))(Y_i-n_i\mu_i),$$

where $d(Y) = \min\{1, \dfrac{\sum_{j=1}^{p} Y_j/(n_j\beta_j+1)}{\sum_{j=1}^{p} Y_j/(n_j\beta_j+1)^2+ \sum_{j=1}^{p} ((Y_j-n_j\mu_j)/(n_j\beta_j+1))^2}\}$.

The performance of δ^* has been discussed under the loss L_1, which here is $\sum_{i=1}^{p} (\delta_i-n_i\lambda_i)^2$. Usually however one is interested in estimating not $n_i\lambda_i$, but the λ_i themselves under the loss $\sum_{i=1}^{p} (\delta_i-\lambda_i)^2$. A possible estimator of $(\lambda_1,\ldots,\lambda_p)$ is

$$\delta^{**} = (\frac{\delta_1^*}{n_1},\ldots, \frac{\delta_p^*}{n_p}).$$

Note that

$$R(\delta^{**},\lambda) = E[\sum_{i=1}^{p} (n_i^{-1}\delta_i^*(Y)-\lambda_i)^2] = E[\sum_{i=1}^{p} n_i^{-2}(\delta_i^*(Y)-n_i\lambda_i)^2].$$

Hence if $n_1 = \ldots = n_p = n$, then

$$R(\delta^{**},\lambda) = n^{-2}R(\delta^{*},n\lambda),$$

and the performance of δ^{**} (relative to the MVUE) under the loss $\sum_{i=1}^{p}(\delta_i-\lambda_i)^2$ will be equivalent to the performance of δ^{*} under the loss $\sum_{i=1}^{p}(\delta_i-n\lambda_i)^2$.

If the n_i's are not all equal, then $R(\delta^{**},\lambda)$ becomes a weighted sum of the losses $(\delta_i^{*}(Y)-n_i\lambda_i)^2$, $i = 1,\ldots,p$. This motivates the consideration of the general loss function

$$\sum_{i=1}^{p} q_i(\delta_i-\lambda_i)^2,$$

when $\lambda = (\lambda_1,\ldots,\lambda_p)$ is to be estimated. We, therefore, briefly discuss the performance of δ^{*} with respect to a weighted loss.

First consider the extreme case where $q_1 = 1$ and $q_i = 0$ for $i \neq 1$. Then

$$\sum_{i=1}^{p} q_i(\delta_i-\lambda_i)^2 = (\delta_1-\lambda_1)^2.$$

Recall that the shrinking constant of δ_1^{*} is

$$1-c_1^{*}(X) = 1 - \frac{1}{\beta_1+1}\min\{1, \frac{\sum_{j=1}^{p} X_j/(\beta_j+1)}{\sum_{j=1}^{p} X_j/(\beta_j+1)^2+\sum_{j=1}^{p}((X_j-\mu_j)/(\beta_j+1))^2}\}.$$

If the parameters $\lambda_1,\ldots,\lambda_p$ are not close in size, it is possible for $1-c_1^{*}(X)$ to be dominated by the observations $X_2,\ldots,X_p$. That is, the coordinates of λ that are not important in the loss may exert an undersirable influence on the risk of δ^{*}.

Generally, δ^* does not appear to be appropriate in the case of weighted loss. Recall that in the derivation of δ^*, a symmetric loss gave rise to the shrinking constant $1-c^*(X)$, and a weighted loss would imply a shrinking constant of a different form. That is, if the derivation of Section 2.2 is applied to the loss $\sum_{i=1}^{p} q_i(\delta_i-\lambda_i)^2$, then the resulting estimator is defined componentwise by

$$\hat{\delta}_i(X)=\mu_i+\left(1 - \frac{1}{\beta_i+1} \min\left\{1, \frac{\sum_{j=1}^{p} q_j X_j/(\beta_j+1)}{\sum_{j=1}^{p} q_j X_j/(\beta_j+1)^2+\sum_{j=1}^{p} q_j((X_j-\mu_j)/(\beta_j+1))^2}\right\}\right)$$

$$\cdot(X_i-\mu_i), \quad i = 1,\ldots,p.$$

Note that in the extreme case where $q_k = 1$ and $q_i = 0$ for $i \neq k$, $\hat{\delta}_k$ is equivalent to the one dimensional form of δ^*. Thus it appears that $\hat{\delta}$ would not exhibit the same problem as δ^* in this situation, and this example suggests that $\hat{\delta}$ may be a good alternative to δ^0 under weighted loss. However, a true Bayes estimator (from this loss) would not depend on the weights $q_1,\ldots,q_p$. Therefore, from a Bayesian standpoint, $\hat{\delta}$ is not desirable in that it does depend on the particular weights used.

Berger (1977b) suggested another method of dealing with nonsymmetric loss functions. The general idea is to divide the original problem into a number of subproblems. Let

$$L(\delta,\lambda) = \sum_{i=1}^{p} q_i L_i(\delta_i,\lambda_i)$$

be a convex loss for the problem, where all $q_i \geq 0$. Also, if z is a vector, let

$$z^j = (z_1,\ldots,z_j).$$

The subproblems that are considered are those in estimating the λ^j. Define the loss function for the jth subproblem of estimating λ^j by

$$L^{(j)}(\delta,\lambda^j) = \sum_{i=1}^{j} \alpha_i^j q_i L_i(\delta_i,\lambda_i),$$

where $0 \leq \alpha_i^j \leq 1$, $\alpha_i^j = 0$ for $j < i$, and $\sum_{j=1}^{p} \alpha_i^j = 1$ for all i. For each subproblem, find an estimator $\delta^{(j)}$ which has uniformly smaller risk than the usual estimator $\bar{\delta}^j$. Then the estimator defined componentwise as

$$\delta_i'(X) = \sum_{j=1}^{p} \alpha_i^j \delta_i^{(j)}(X), \quad i = 1,\ldots,p,$$

will have uniformly smaller risk than $\bar{\delta}^p$ for the loss L in the original problem (for a proof of this statement, see Berger).

As one application of this method, Berger assumes without loss of generality that $q_1 \geq q_2 \geq \ldots \geq q_p$, and he defines

$$\alpha_i^j = \begin{cases} 0, & j < i \\ (q_j - q_{j+1})/q_i, & j \geq i, \end{cases}$$

where q_{p+1} is defined to be zero. Then

$$L^{(j)}(\delta,\lambda^j) = (q_j - q_{j+1}) \sum_{i=1}^{j} L_i(\delta_i,\lambda_i).$$

In our case, $L_i(\delta_i,\lambda_i) = (\delta_i-\lambda_i)^2$, and although $\delta^{*(j)}$ does not have uniformly smaller risk than $\delta^{0(j)}$ under loss $L^{(j)}$, it has been shown that $\delta^{*(j)}$ is an attractive alternative to $\delta^{0(j)}$. So the suggested estimator of λ under loss L is

$$\delta_i'(X) = \sum_{j=1}^{p} \alpha_i^j \delta_i^{*(j)}(X), \quad i = 1,\ldots,p,$$

where α_i^j is defined above.

As an example, take p = 3. The suggested estimator of λ under loss $L(\delta,\lambda) = \sum_{i=1}^{3} q_i(\delta_i-\lambda_i)^2$ is componentwise

$$\delta_1'(X) = \frac{q_1-q_2}{q_1}\,\delta_1^{*(1)}(X) + \frac{q_2-q_3}{q_1}\,\delta_1^{*(2)}(X) + \frac{q_3}{q_1}\,\delta_1^{*(3)}(X)$$

$$\delta_2'(X) = \frac{q_2-q_3}{q_2}\,\delta_2^{*(2)}(X) + \frac{q_3}{q_2}\,\delta_2^{*(3)}(X)$$

$$\delta_3'(X) = \delta_3^{*(3)}(X).$$

Consider again the extreme case when $q_1 = 1$ and $q_i = 0$ for $i > 1$. Here $\delta_1'(X) = \delta_1^{*(1)}(X)$, the one dimensional form of δ^*, and δ' behaves like the rule $\hat{\delta}$ in this situation. However, unlike a true Bayes estimator, δ' will depend on the weights of the loss function.

In summary, in this section we have considered the problem of taking more than one observation from each population. If the numbers of observations from each population are equal, then an estimator was proposed which will improve upon the MVUE just as δ^* improves upon the MVUE in the one observation case. However if unequal numbers of observations are taken, δ^* will not always

perform well under loss L_1. A general asymmetric loss was considered, and two estimators were proposed which should display better risk behavior than δ^*.

5.2. Unknown prior means

In our discussion, it was assumed that prior parameters $\{(\mu_i,\beta_i),\ i = 1,\ldots,p\}$ are known to the user. Let us now consider the situation where $\lambda_1,\ldots,\lambda_p$ are known to come from a common prior, but no information exists about that prior. First, if $\lambda_1,\ldots,\lambda_p$ come from a common prior, then $\mu_1 =\ldots= \mu_p = \mu^*$, and $\beta_1 =\ldots= \beta_p = \beta^*$, and δ^* becomes

$$\delta_i^*(X) = \mu^*+(1-\min\{\frac{1}{\beta^*+1}, \frac{\sum_{j=1}^{p} X_j}{\sum_{j=1}^{p} X_j+\sum_{j=1}^{p}(X_j-\mu^*)^2}\})(X_i-\mu^*),\ i = 1,\ldots,p.$$

Now μ^* and β^* are unknown, but they can be estimated from the observations. First, marginally $E(X_i) = \mu^*$ for all i, so a natural estimator of μ^* is

$$\hat{\mu} = \sum_{j=1}^{p} X_j/p.$$

Second, in Section 2.3.1, it has been stated that $\sum_{j=1}^{p} X_j/(\sum_{j=1}^{p} X_j + \sum_{j=1}^{p}(X_j-\mu^*)^2)$ is approximately equal to $1/(\beta^*+1)$ when μ^* is known and p is large. Similarly it can be shown that for large p

$$\frac{\sum_{j=1}^{p} X_j}{\sum_{j=1}^{p} X_j + \sum_{j=1}^{p} (X_j-\hat{\mu})^2} \cong \frac{1}{\beta^*+2}.$$

Hence $\sum_{j=1}^{p} X_j/(\sum_{j=1}^{p} X_j + \sum_{j=1}^{p} (X_j-\hat{\mu})^2)$ is a reasonable estimator of

$$\min\{\frac{1}{\beta^*+1}, \frac{\sum_{j=1}^{p} X_j}{\sum_{j=1}^{p} X_j + \sum_{j=1}^{p} (X_j-\mu^*)^2}\}.$$

Thus when μ^* and β^* are unknown, the following estimator is suggested:

$$\tilde{\delta}_i(X) = \hat{\mu}+(1 - \frac{\sum_{j=1}^{p} X_j}{\sum_{j=1}^{p} X_j + \sum_{j=1}^{p} (X_j-\hat{\mu})^2})(X_i-\hat{\mu}),\ i = 1,\ldots,p.$$

Note that $\tilde{\delta}$ corresponds to the usual empirical Bayes estimator as discussed in Chapter 1.

Let us briefly evaluate $\tilde{\delta}$ in the case p = 2. Figure 15 shows contours of constant values of proportional risk $(R(\tilde{\delta},\lambda)/R(\delta^0,\lambda))$ over the plane of λ_1 and λ_2 values. This graph can be compared to Figure 2, where the data is shrunk towards a predetermined point rather than a point that is determined from the data. In Figure 15, the proportional risk has approximately a value of .65 on the $\lambda_1=\lambda_2$ line and appears to have a maximum value of 1.15 near the boundaries of the parameter space. In this example, $\tilde{\delta}$ appears to be an attractive alternative to δ^0 when λ_1 and λ_2 are thought to come from a common unknown prior.

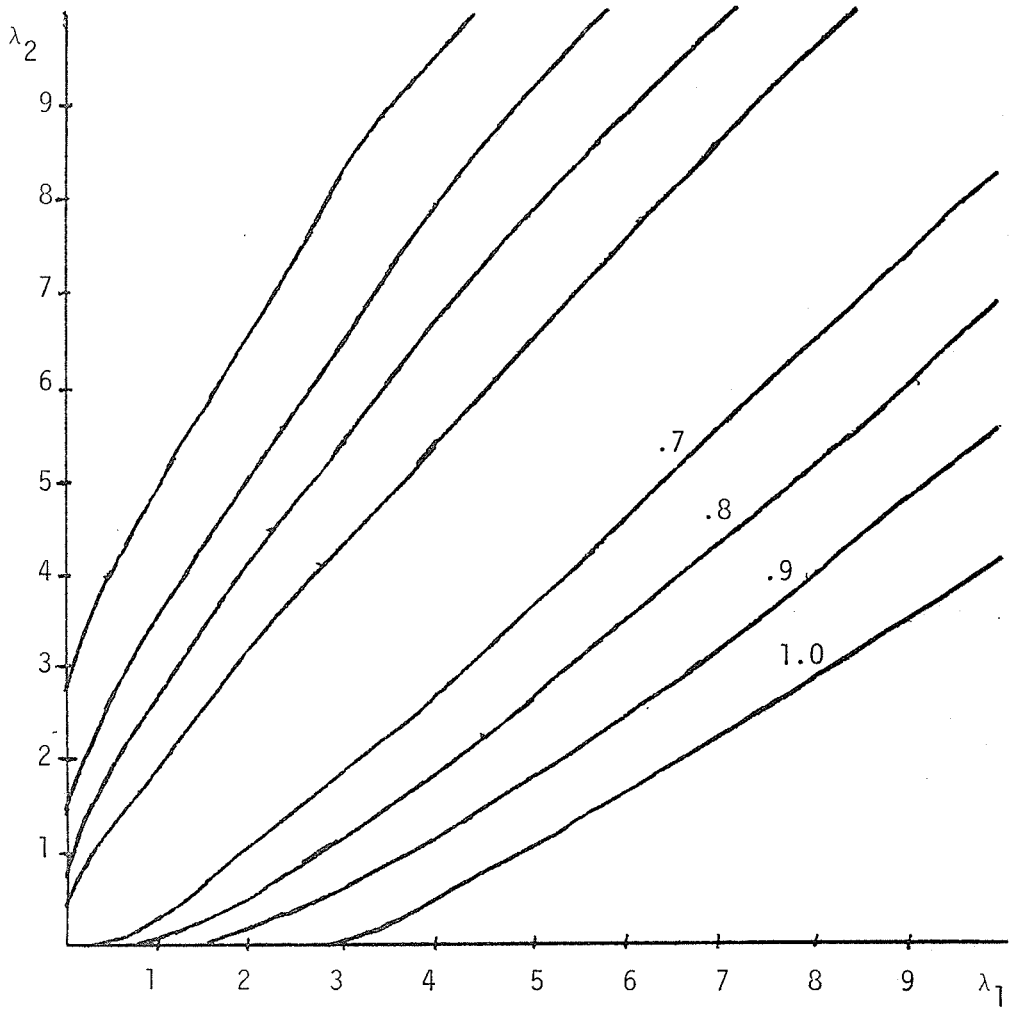


Figure 15

p=2. Contours of constant values of proportional risk of $\tilde{\delta}$.

CHAPTER 3

A ROBUST BAYES CONFIDENCE REGION FOR p POISSON MEANS

1. Introduction

1.1. Improved confidence regions

To motivate the consideration of improved simultaneous confidence regions for p Poisson means, the work that has been done in the normal mean estimation problem will first be discussed. Again assume $X \sim N_p(\theta,\Sigma)$, where Σ is known. Assume it is desired to find a $(1-\gamma)100\%$ confidence region for θ. The usual confidence ellipsoid is

$$C^0(X) = \{\theta: (X-\theta)^t\Sigma^{-1}(X-\theta) \leq \chi^2(\gamma)\},$$

where $\chi^2(\gamma)$ is the $100(1-\gamma)$ percentile of the chi squared distribution with p degrees of freedom.

One reasonable method of evaluating the goodness of a particular confidence region is to formally define a loss in using the confidence region. Let $L(C(X),\theta)$ denote the loss in using $C(X)$ to estimate θ. Also define the corresponding risk

$$R(C,\theta) = E_\theta L(C(X),\theta).$$

One risk that has been considered is of the form

$$R(C,\theta) = c_1 E_\theta h(C(X)) + c_2(1-P_\theta(\theta \in C(X))),$$

where $h(C(X))$ is some measure of the size of the region $C(X)$, $P_\theta(\theta \in C(X))$ is the probability of coverage of C, and c_1, c_2 are appropriate constants. With regard to this particular risk, Brown (1966) and Joshi (1967) both showed the region C^0 is inadmissible for $p \geq 3$. The regions that were shown to have uniformly smaller risk than C^0 differ from C^0 only in the centering term. In fact, one of the better regions may be expressed as

$$C^*(X) = C^0(\tilde{\delta}(X)),$$

where $\tilde{\delta}(X)$ has a similar form to an estimator that has uniformly smaller risk than $\delta^0(X) = X$ (under quadratic loss) in point estimation of a normal mean.

Other authors have suggested different alternative confidence regions to C^0 in this normal problem. When $\Sigma = I$, Stein (1962) suggested the confidence ellipsoid

$$C^S(X) = \{\theta: |\theta-(1 - \frac{p}{\sum_{j=1}^{p} X_j^2})X|^2 \leq (1 - \frac{p}{\sum_{j=1}^{p} X_j^2})\chi^2(\gamma)\}.$$

This region was developed assuming p is large and differs from C^0 in both the centering and width terms. The centering term $(1-p/\sum_{j=1}^{p} X_j^2)X$ is almost identical to the James-Stein estimator discussed in Chapter 1, and the factor $(1-p/\sum_{j=1}^{p} X_j^2)$ in the width term is just the shrinking constant of the centering term. This extra factor in the width term is appealing from a Bayesian standpoint,

since if one believes that θ is in a particular region, the Bayesian interval using this prior information is shorter than the corresponding classical interval. A more sophisticated form of C^S was later developed by Stein (1974), and differs from C^S in the expression of the width term. Berger (1977a) developed a confidence ellipsoid centered at his robust Bayes estimator (see Chapter 1), and argued that it is an improved region (over C^0) in terms of both size and probability of coverage. Thus analogous to point estimation, the classical normal mean confidence region C^0 has been improved upon when $p \geq 3$. The improved regions are based on estimators that are related to the improved estimators in point estimation of a normal mean.

1.2. Confidence regions for p Poisson means

Methods of constructing confidence rectangles for the Poisson parameter $\lambda = (\lambda_1,\ldots,\lambda_p)$ will be discussed in this section. Rectangles are considered instead of ellipsoids, because there exist classical techniques for constructing a confidence interval for a component of λ, and simultaneous confidence intervals for $\lambda_1,\ldots,\lambda_p$ are equivalent to a confidence rectangle for λ.

The usual classical procedure for constructing a confidence interval for a component of λ, say λ_i, is based on the fact that as λ_i approaches infinity, $\lambda_i^{-1/2}(X_i-\lambda_i)$ is asymptotically normally distributed with mean zero and variance one. By solving the equations

$$\lambda_i^{-1/2}(X_i-\lambda_i) = \pm z_{\gamma/2},$$

where $z_{\gamma/2}$ is the $100(1-\gamma/2)$ percentile of the standard normal distribution, one obtains the confidence interval

$$X_i + z^2_{\gamma/2}/2 \pm z_{\gamma/2}(X_i+z^2_{\gamma/2}/4)^{1/2}.$$

This interval clearly has asymptotic probability of coverage of $1-\gamma$ as λ_i approaches infinity. The confidence rectangle formed by p of the above intervals is defined by

$$C^0(X) = \{\lambda: |X_i+z^2_{\gamma/2}/2-\lambda_i| \leq z_{\gamma/2}(X_i+z^2_{\gamma/2}/4)^{1/2}, \ i = 1,\ldots,p\}.$$

Asymptotically, C^0 has probability $(1-\gamma)^p$ of covering λ. C^0 is the usual confidence region based on the MVUE δ^0, and our suggested robust Bayes region will be compared with C^0.

Since prior information will be used in our confidence region for λ, a "standard" Bayes confidence region, or credible region, will now be considered. This will be the (approximate) credible rectangle for λ based on the conjugate prior. This credible region will be used as the starting point for the development of our robust Bayes region.

Let us first derive an approximate credible interval for the component λ_i. As before, X_i given λ_i is distributed Poisson (λ_i), λ_i is distributed gamma (α_i,β_i), and the posterior distribution of λ_i given X_i is gamma $(X_i+\alpha_i,\beta_i/(\beta_i+1))$. A $(1-\gamma)100\%$ credible interval will contain $(1-\gamma)100\%$ of the posterior distribution of λ_i. Usually it is desirable to find the smallest interval containing $(1-\gamma)100\%$ of the posterior probability. This interval is called the

$(1-\gamma)100\%$ HPD (highest posterior density) credible interval.

Since the posterior distribution is gamma in this case, we first find an approximation for an arbitrary percentile of the gamma distribution. Using the Wilson-Hilferty approximation, the 100ϵ percentile of the chi squared distribution with v degrees of freedom can be approximated by

$$v[z_{1-\epsilon}(2/(9v))^{1/2} + 1-2/(9v)]^3,$$

where $z_{1-\epsilon}$ is the 100ϵ percentile of the standard normal distribution (Johnson and Kotz (1970)). If X is chi squared with v degrees of freedom, it is well known that $Y = \beta X/2$ has a gamma distribution with parameters $v/2$ and β. Hence the 100ϵ percentile of the gamma distribution with parameters α and β is approximately

$$\alpha\beta(z_{1-\epsilon}/(3\alpha^{1/2}) + 1-1/(9\alpha))^3.$$

It follows that an approximation to the 100ϵ percentile of the posterior distribution of λ_i is

$$\frac{(X_i+\alpha_i)\beta_i}{\beta_i+1}\left(z_{1-\epsilon}\frac{1}{3(X_i+\alpha_i)^{1/2}} + 1 - \frac{1}{9(X_i+\alpha_i)}\right)^3.$$

It would be desirable to use the above approximate 100ϵ percentile in finding the HPD credible interval. Unfortunately, it is difficult in general to specify the HPD credible interval in terms of the percentiles of the distribution. We will instead use these approximate percentiles to find the interval which approximately contains the middle $(1-\gamma)100\%$ of the posterior distribution, and then

argue that the length of this interval is close to the length of the HPD credible interval. In the above percentile formula, set $\epsilon_1 = \gamma/2$ and $\epsilon_2 = 1-\gamma/2$, and expand the resulting expressions, using $z_{1-\gamma/2} = -z_{\gamma/2}$. After some algebra, one finds that the interval which approximately contains the middle $(1-\gamma)100\%$ posterior area of λ_i is

$$\frac{(X_i+\alpha_i)\beta_i}{\beta_i+1} + \frac{\beta_i}{\beta_i+1}\left[z^2_{\gamma/2}\left(\frac{1}{3} - \frac{1}{27(X_i+\alpha_i)}\right) - \frac{1}{3} + \frac{1}{27(X_i+\alpha_i)} - \frac{1}{729}\frac{1}{(X_i+\alpha_i)^2}\right]$$

$$\pm z_{\gamma/2}\frac{(X_i+\alpha_i)^{1/2}\beta_i}{\beta_i+1}\left[1 + \frac{1}{27(X_i+\alpha_i)}z^2_{\gamma/2} + \frac{1}{81(X_i+\alpha_i)^2} - \frac{2}{9}\frac{1}{(X_i+\alpha_i)}\right].$$

Denote by C^B the confidence rectangle for λ formed by p of the above intervals.

The above interval will generally be longer than the optimal HPD credible interval for λ_i. The HPD credible interval will be significantly shorter when the gamma posterior distribution is skewed towards the left, that is, when the quantity $X_i+\alpha_i$ is small. When $(X_i+\alpha_i) \geq 2$, and the posterior distribution is more symmetric, the two intervals are similar in length. From numerical studies, it appears that the above interval is less than 12% longer than the HPD credible interval when $X_i+\alpha_i = 2$, and as $X_i+\alpha_i$ increases, the length of the above interval approaches the length of the HPD credible interval. Since the prior parameter α_i is frequently chosen to be greater than zero, $X_i+\alpha_i$ is often larger than two. Therefore, in many applications, the

credible interval derived above will have length approximately equal to the length of the HPD credible interval. Modifications to the intervals in C^B will now be made to derive a robust Bayes confidence region.

2. Development of a robust Bayes confidence region

As in the point estimation problem, one would like a confidence region to perform like a natural Bayesian confidence region in an area of the parameter space. Let us consider the approximate credible interval for λ_i derived in the last section. Note that when $X_i+\alpha_i \geq 1$, certain terms in this interval may be ignored, since they contribute little to the location and width of the interval. If these terms are removed, then the approximate credible region is

$$\frac{(X_i+\alpha_i)\beta_i}{\beta_i+1} + \frac{\beta_i}{\beta_i+1}\frac{z^2_{\gamma/2}-1}{3} \pm z_{\gamma/2}\left[\frac{(X_i+\alpha_i)\beta_i}{\beta_i+1}\right]^{1/2}\left(\frac{\beta_i}{\beta_i+1}\right)^{1/2}.$$

This interval will be approximately equivalent to the original interval in most applications if $X_i+\alpha_i$ is not small. Note that this interval contains two important expressions, $(X_i+\alpha_i)\beta_i/(\beta_i+1)$, the Bayes estimator of λ_i under squared error loss and $\beta_i/(\beta_i+1)$.

To construct a robust Bayes interval, two substitutions are made. We substitute our recommended estimator δ_i^* for the Bayes estimator $(X_i+\alpha_i)\beta_i/(\beta_i+1)$, and substitute

$$1-c_i^*(X) = 1 - \frac{1}{\beta_i+1} \min\{1, \frac{\sum_{j=1}^{p} X_j/(\beta_j+1)}{\sum_{j=1}^{p} X_j/(\beta_j+1)^2 + \sum_{j=1}^{p} ((X_j-\mu_j)/(\beta_j+1))^2}\}$$

for $1-1/(\beta_i+1) = \beta_i/(\beta_i+1)$, as was done in the derivation of δ^*. With these two substitutions, the interval becomes

$$\delta_i^*(X) + (1-c_i^*(X))(z_{\gamma/2}^2-1)/3 \pm z_{\gamma/2}(\delta_i^*(X))^{1/2}(1-c_i^*(X))^{1/2}.$$

The confidence rectangle based on these p intervals will be denoted by C^R. As will be shown later, it is necessary to make one further adjustment to the intervals in C^R so that the resulting confidence region has a probability of coverage close to the nominal level. That adjustment is to replace $\delta_i^*(X)$ in the width term of the ith interval by $X_i+z_{\gamma/2}^2/4$. The width term of the resulting interval will then only differ from the width term of the classical interval by a factor of $(1-c_i^*(X))^{1/2}$. This change has the effect of generally increasing the length of the ith interval and therefore the volume of the confidence rectangle. With this final adjustment, the recommended robust Bayes rectangle, denoted C*, is formed from the intervals

$$\delta_i^*(X) + (1-c_i^*(X))(z_{\gamma/2}^2-1)/3 \pm z_{\gamma/2}(X_i+z_{\gamma/2}^2/4)^{1/2}(1-c_i^*(X))^{1/2}.$$

Before the confidence region C* is formally evaluated, one can see from its form that it seems to behave like a robust Bayes region. In Chapter 2, it was argued that when the prior information is correctly specified, p is large, and $\beta_1,\ldots,\beta_p$ are chosen to be

of moderate size, then $c_i^*(X) \tilde{=} (\beta_i+1)^{-1}$ and $\delta_i^*(X) \tilde{=} \delta_i^B(X)$. In this situation, C* will not be an approximation to C^B, because of the above adjustment. But the centering terms of the intervals in C* will be approximately equal to the centering terms of the intervals in C^B. Also in this case, the widths of the intervals in C* will be significantly smaller than the corresponding interval widths in C^0. If, on the other hand, the prior information has been misspecified, then at least one observation is far from its corresponding prior mean, $c_i^*(X) \tilde{=} 0$, and $\delta_i^*(X) \tilde{=} X_i$. In this case, C* will resemble the classical region C^0 based on the MVUE.

From inspection, we expect C* to improve upon C^0 in a prior region, but not perform much worse than C^0 outside of the prior region. In the next section, the criteria for evaluating a confidence region will be defined, and comparisons will be made between C* and C^0 with respect to this criteria.

3. Evaluation

3.1. Methods of evaluation

The two main criteria that will be used to evaluate the goodness of a particular confidence interval are probability of coverage and size. In Chapter 2, δ^* was shown to be an attractive alternative to δ^0 when vague prior information is available. Here it is shown that C*, the confidence region based on δ^*, is an attractive alternative to C^0. Also C* is shown to be more robust to uncertainty in the prior specification than the approximate conjugate Bayes confidence region C^B.

In Section 3.2, probability of coverage is considered. To evaluate a particular confidence region, usually probabilities of coverage of a region are compared against some nominal level. To obtain this nominal level, a confidence of $(1-\gamma)100\%$ will be assigned to the ith interval covering λ_i, $i = 1,\ldots,p$. Using independence, this will imply a confidence of $(1-\gamma)^p 100\%$ assigned to the rectangle covering λ. From a classical viewpoint, it is desirable for a confidence rectangle to possess a probability of coverage of at least the nominal level $((1-\gamma)^p)$ over the entire parameter space.

In two examples, probabilities of coverage are found through simulation for the regions discussed above. It is seen that C* has higher probabilities of coverage than C^0 in a region about the prior mean. Outside of the prior region, C* has probabilities of coverage close to the nominal level, and asymptotically, as λ approaches infinity, C* is shown to have exactly the nominal level probability of coverage. Comparisons are also made between the regions C* and C^R with respect to probabilities of coverage, and it is seen why adjustments were made to C^R in arriving at C*. Finally, C^B is shown to be more sensitive than C* to misspecification of the prior information.

In Section 3.3, it is shown that C* is generally an improvement upon C^0 with respect to size. The measure of size that is used is the volume of the rectangle. The volume of C* is significantly smaller than the volume of C^0 in a region about the prior mean and can not be larger elsewhere. In one of the examples above, the expected size of C* is compared with the expected size of C^0 in a large region about the prior mean.

3.2. Probability of coverage

In this section, comparisons are made between the classical region C^0, the Bayes region C^B, and the robust Bayes regions C^R and C* with respect to probabilities of coverage. First consider the case p = 2. By means of computer simulation, probabilities of coverage are found for the four confidence regions over the plane of λ_1 and λ_2 values (Figures 16, 17 and 18). In these simulations, at least 10,000 random variables X were generated, and the probabilities of coverage found have a standard error of approximately .003. In this example, the prior information $(\mu_i,\beta_i) = (4,2)$, i = 1,2 is used for the Bayes procedures. Thus the prior standard deviation of $\sum_{i=1}^{2} \lambda_i$ is $(\mu_1\beta_1+\mu_2\beta_2)^{1/2} = 4$, and this value can be used in specifying distances from the prior mean. In this example, $\gamma = .05$, so $z_{\gamma/2} = 1.96$ is used in the intervals, and the nominal level probability of coverage is $(.95)^2 = .9025$.

From looking at Figure 16, one notes that the probabilities of coverage of the classical region C^0 are a little small for small values of λ, but otherwise are uniformly close to .9. This behavior is expected, since the intervals in C^0 were derived assuming that $\lambda_1,\ldots,\lambda_p$ are large (see Section 1.2). Figure 18 shows the probabilities of coverage of the robust Bayes region C^R and the recommended region C*. Both regions obtain high probabilities of coverage near the prior mean (4,4) and they appear to approach the nominal level for values far from the prior mean. The problem in achieving a uniform probability of coverage of .9025 occurs in

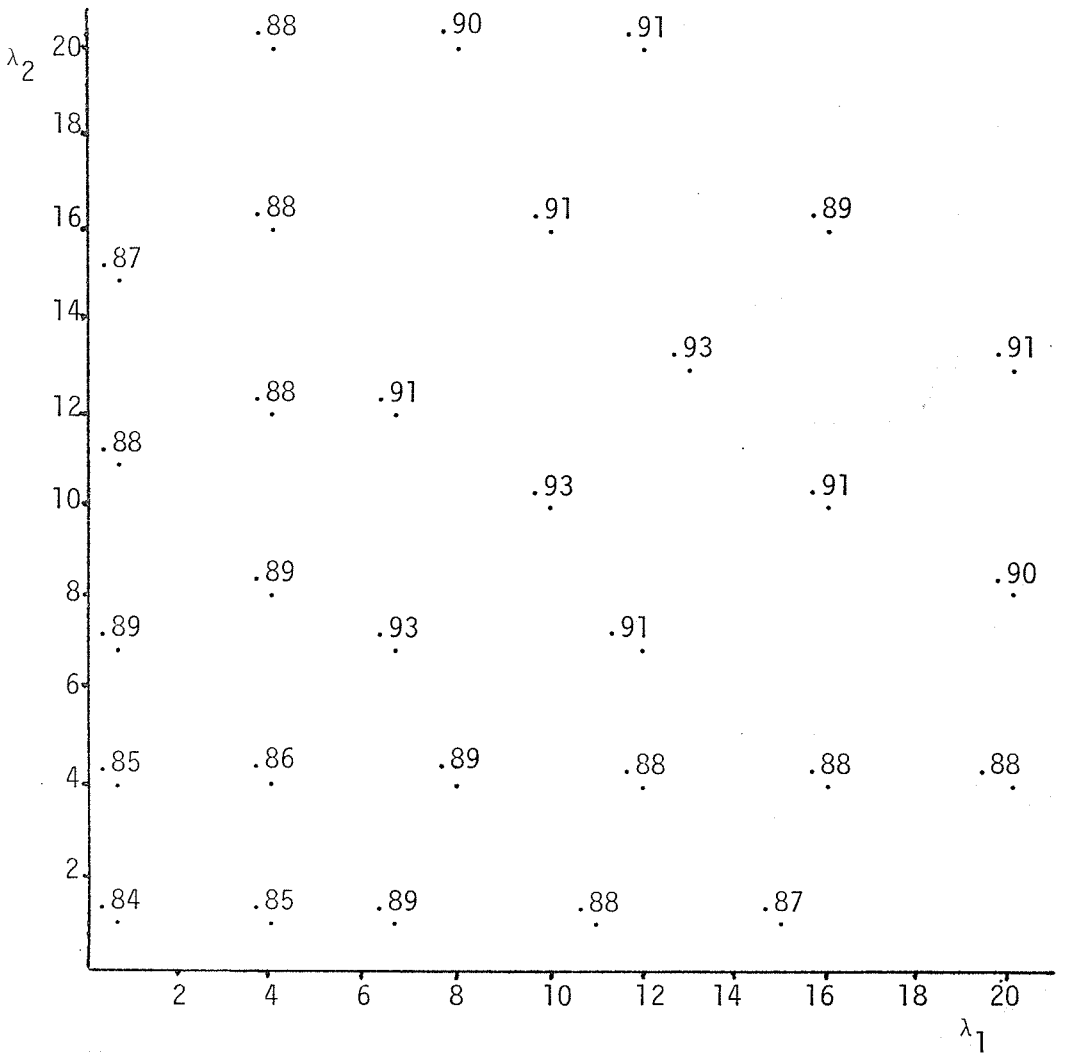


Figure 16

p=2. Probabilities of coverage of C^0.

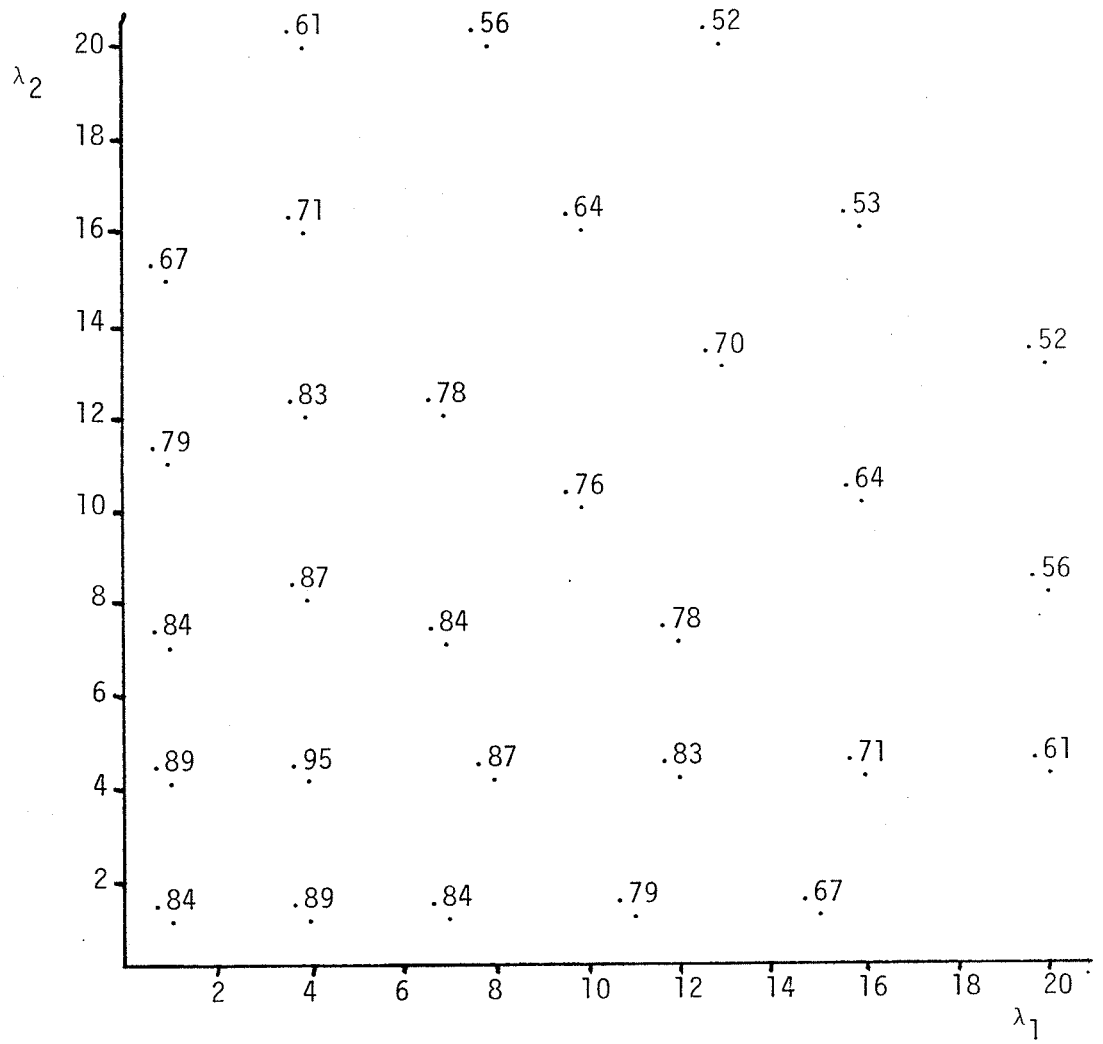


Figure 17

p=2. Prior information: $(\mu_i,\beta_i)=(4,2)$, i=1,2. Probabilities of coverage of C^B.

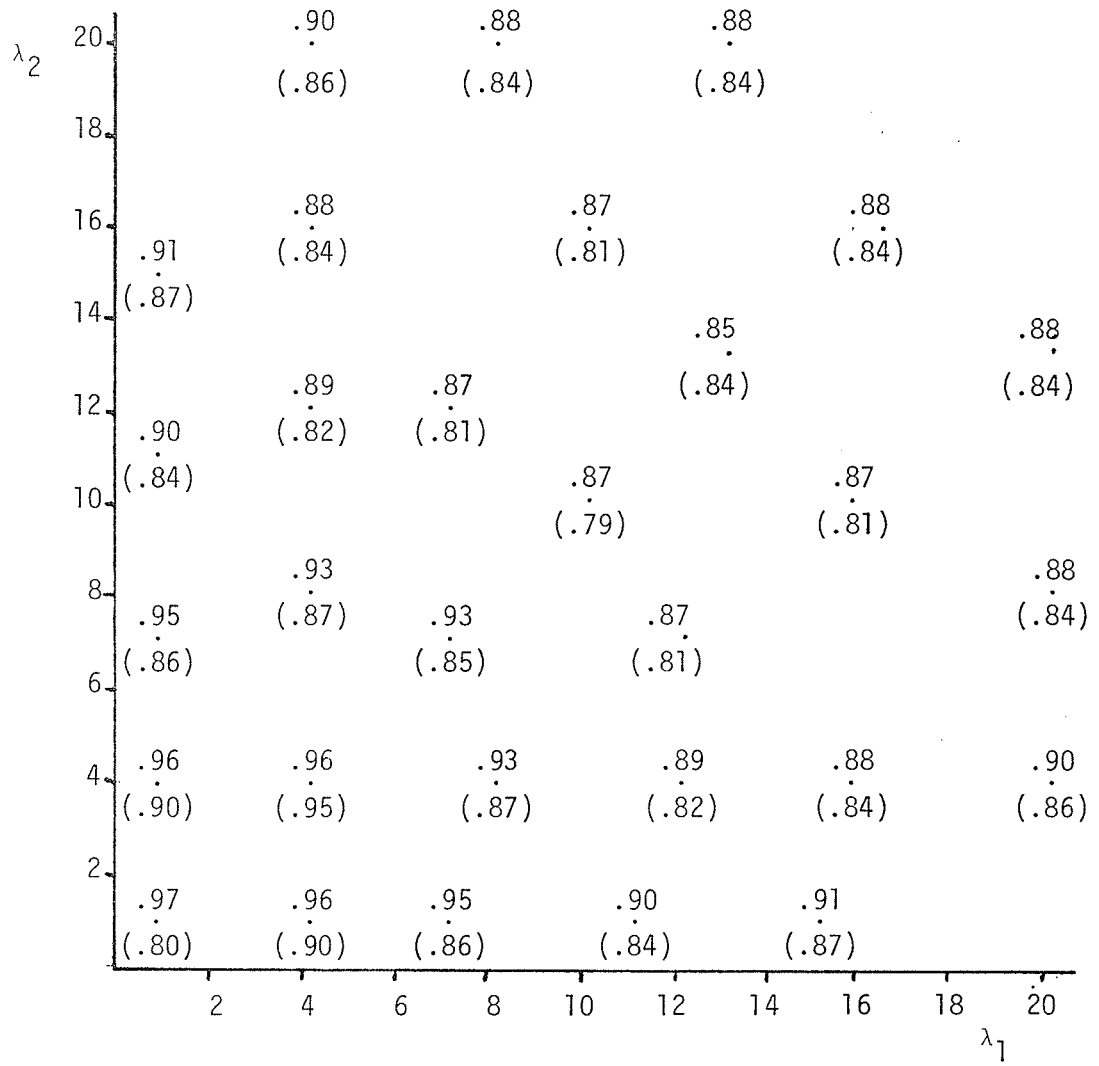


Figure 18

p=2. Prior information: $(\mu_i,\beta_i)=(4,2)$, i=1,2. Probabilities of coverage of C* (and C^R).

the region a few prior standard deviations from (4,4). In this region, C^R has fairly low probabilities of coverage, as low as .79 at the point (10,10) and .80 at the point (1,1). Note that C*, which was developed by adjusting the intervals in C^R, appears to have significantly higher probabilities of coverage than C^R, especially near the origin and in the region one to two standard deviations away from the prior mean. Thus C* appears to never have a probability of coverage much worse than .9, while having larger probabilities of coverage than C^0 in a region about the prior mean. From a classical viewpoint, C* appears to be close to a legitimate 90% confidence rectangle. From a Bayesian viewpoint, C* appears to be extremely robust to errors made in the prior specification.

Figure 17 shows the probabilities of coverage of the Bayes confidence region C^B in the same situation. Although the probabilities of coverage are high close to the prior mean, the probabilities rapidly decrease as one moves away from μ. This confidence region thus appears to be appropriate only when the unknown parameters λ_1 and λ_2 are very strongly believed to be close to the prior mean. The robust region C* seems to be more appropriate than C^B when values of λ at least two standard deviations from the prior mean are possible.

Let us consider a second example where there is asymmetry in the prior information used. Consider the case p = 3, where the prior information is $(\mu_1,\beta_1) = (2,1)$, $(\mu_2,\beta_2) = (4,3)$, and $(\mu_3,\beta_3) = (6,1)$. Again set $\gamma = .05$, so the nominal level probability of coverage is $(.95)^3 = .857$. In Figures 19, 20 and 21, probabilities of

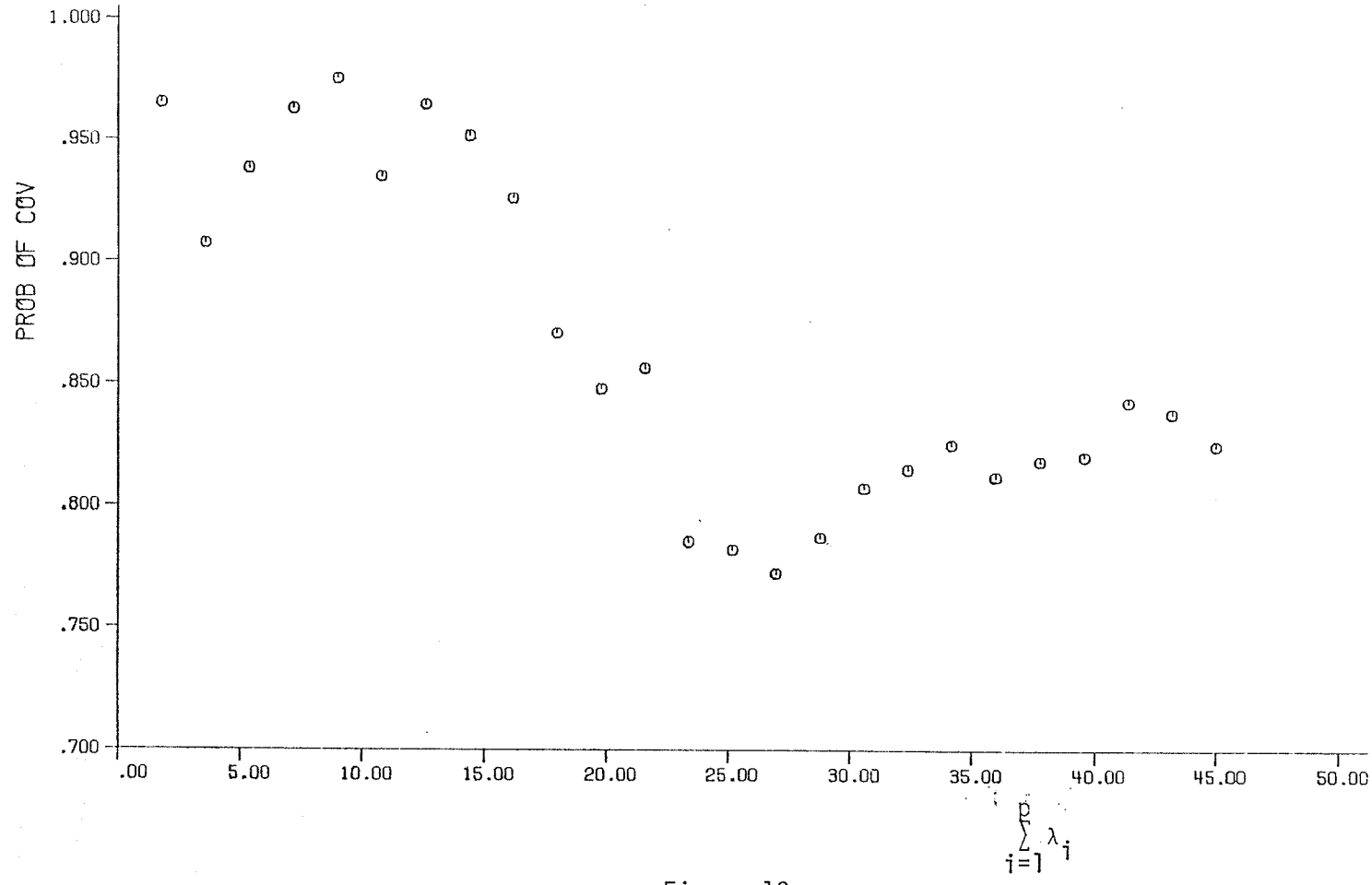


Figure 19

p=3. Prior information: $(\mu_1,\beta_1)=(2,1)$, $(\mu_2,\beta_2)=(4,3)$ and $(\mu_3,\beta_3)=(6,1)$. Probabilities of coverage of C* along line L_1: $\lambda_1=\eta$, $\lambda_2=2\eta$, $\lambda_3=3\eta$.

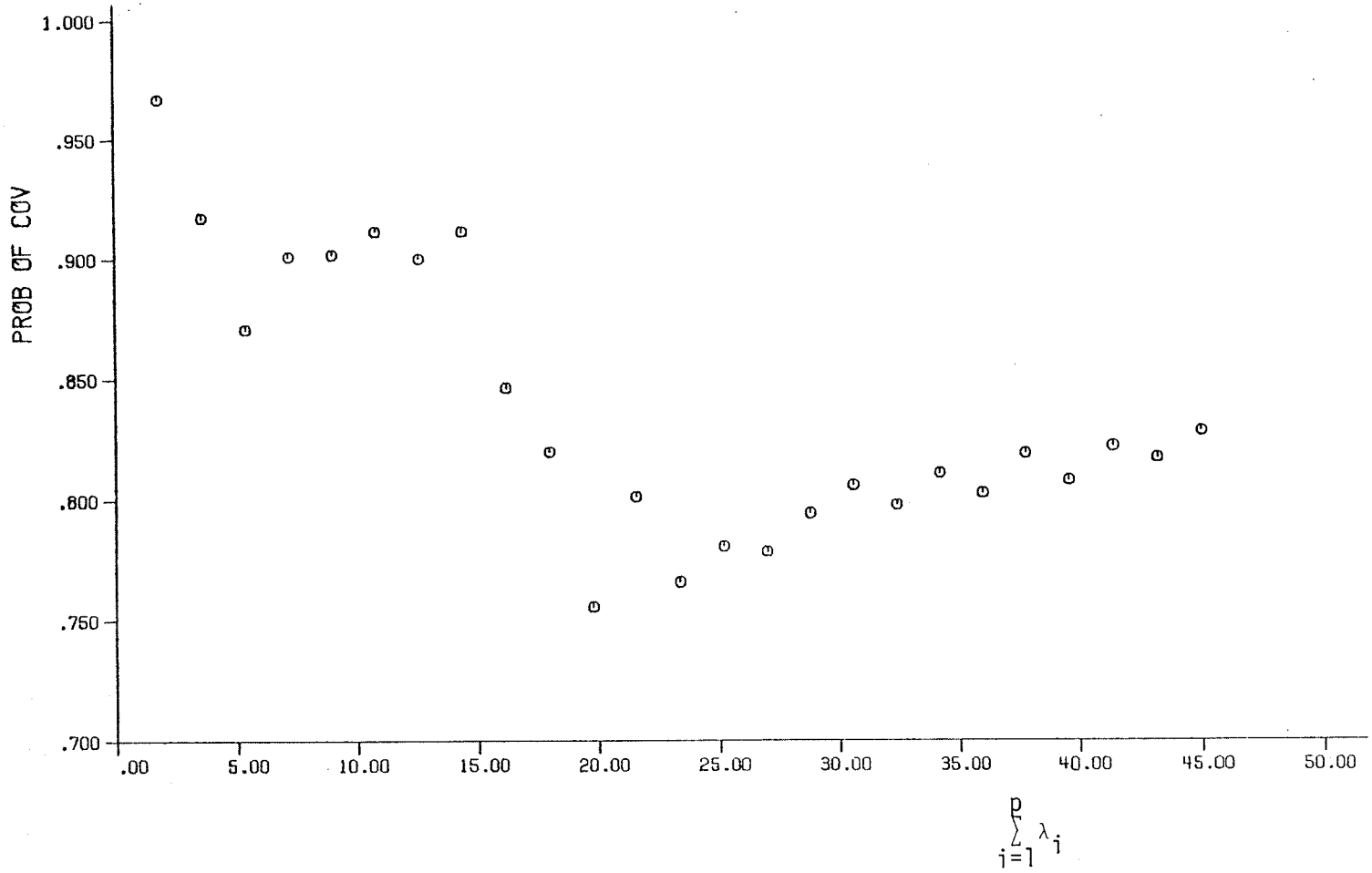


Figure 20

p=3. Prior information: $(\mu_1,\beta_1)=(2,1)$, $(\mu_2,\beta_2)=(4,3)$ and $(\mu_3,\beta_3)=(6,1)$. Probabilities of coverage of C* along line L_2: $\lambda_1 = \lambda_2 = \lambda_3$.

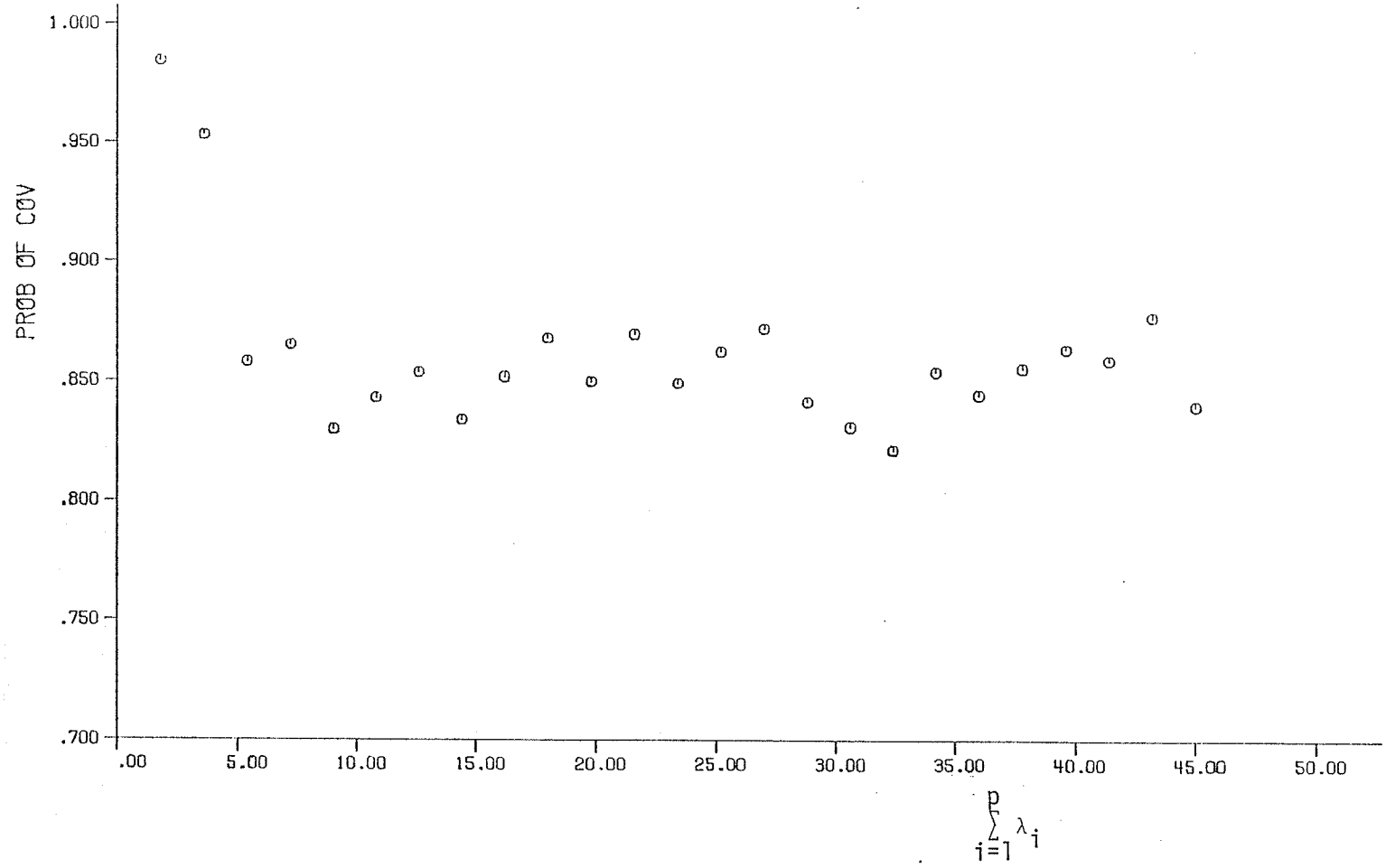


Figure 21

p=3. Prior information: $(\mu_1,\beta_1)=(2,1)$, $(\mu_2,\beta_2)=(4,3)$ and $(\mu_3,\beta_3)=(6,1)$. Probabilities of coverage of C* along line L_3: $\lambda_1=5\eta$, $\lambda_2=3\eta$, $\lambda_3=\eta$.

coverage of C* are plotted along three lines. Line L_1 goes through the prior mean (2,4,6), while lines L_2 and L_3 do not. (In this example, the calculated probabilities of coverage have a standard error of less than .004.) Looking along line L_1, one sees that C* has higher than nominal level probability of coverage near μ, and for large λ, the probability of coverage appears to approach the nominal level. Along line L_3, which is far from the prior mean, C* appears to have approximately the nominal level probability of coverage. Thus, as in the previous example, C* appears to use prior information and be robust with respect to errors in the specification of the prior information.

In the above examples, the probability of coverage of C* appears to approach the nominal level for values of $\lambda_1,\ldots,\lambda_p$ that are large and outside of the prior region. A theorem will now be given which states that C* has an asymptotic probability of coverage of $(1-\gamma)^p$ as $\lambda_1,\ldots,\lambda_p$ go to infinity along a line.

Theorem 3

Let

$$C^*(X) = \{\lambda: |\delta_i^*(X)+(1-c_i^*(X))(z_{\gamma/2}^2-1)|3-\lambda_i| \leq z_{\gamma/2}(1-c_i^*(X))^{1/2}(X_i+z_{\gamma/2}^2/4)^{1/2} \ \forall i\}.$$

Let $\lambda_i = k_i\eta$, $i = 1,\ldots,p$. Then

$$P_\lambda(\lambda \in C^*(X)) \xrightarrow[\eta\to\infty]{} (1-\gamma)^p.$$

Proof: See Appendix.

If the prior information has been incorrectly specified and the values of $\lambda_1,\ldots,\lambda_p$ are very large, then by Theorem 3 , the confidence rectangle C* will be approximately equivalent to the classical rectangle C^0 with respect to probability of coverage.

3.3. Size

In this section the classical region C^0 and the robust Bayes region C* will be compared with respect to the volume of the confidence rectangle. First, consider the confidence interval for a component of λ, say λ_i. The widths of the intervals for C^0 and C* are respectively

$$2\, z_{\gamma/2}(X_i+z^2_{\gamma/2}/4)^{1/2}$$

and

$$2\, z_{\gamma/2}(X_i+z^2_{\gamma/2}/4)^{1/2}(1-c_i^*(X))^{1/2}.$$

The ratio of the width of the robust Bayes interval to the width of the classical interval is $(1-c_i^*(X))^{1/2}$ and the ratio of the volume of C* to the volume of C^0 is

$$\prod_{i=1}^{p}(1-c_i^*(X))^{1/2}.$$

Recall that $c_i^*(X) \tilde{=} 1/(\beta_i+1)$ when p is large, moderate values of $\beta_1,\ldots,\beta_p$ are chosen, and the prior information is correct. On the other hand, for values of X far from the prior mean, $c_i^*(X) \tilde{=} 0$. Thus C* will have a substantially smaller volume than C^0 for values of

$X_1,\ldots,X_p$ near the prior mean and have a volume approaching the volume of the classical rectangle for observations away from the prior mean.

In the first example given earlier, $p = 2$, and $\beta_1 = \beta_2 = 2$. Thus for correct prior information, the volume ratio is

$$\prod_{i=1}^{2} (1-c_i^*(X))^{1/2} \tilde{=} 2/3.$$

One would expect the volume ratio to be approximately 2/3 for values of (X_1,X_2) near the prior mean (4,4). For values of (X_1,X_2) away from (4,4), one would expect the volume ratio to approach one. In this example, the expected sizes of the two regions have been found through simulation, and Figure 22 gives values of

$$\frac{\text{Expected size of } C^*}{\text{Expected size of } C^0}.$$

This figure shows that, on the average, the volumes of C^* are significantly smaller than the volumes of C^0 in the region within several standard deviations of the prior mean.

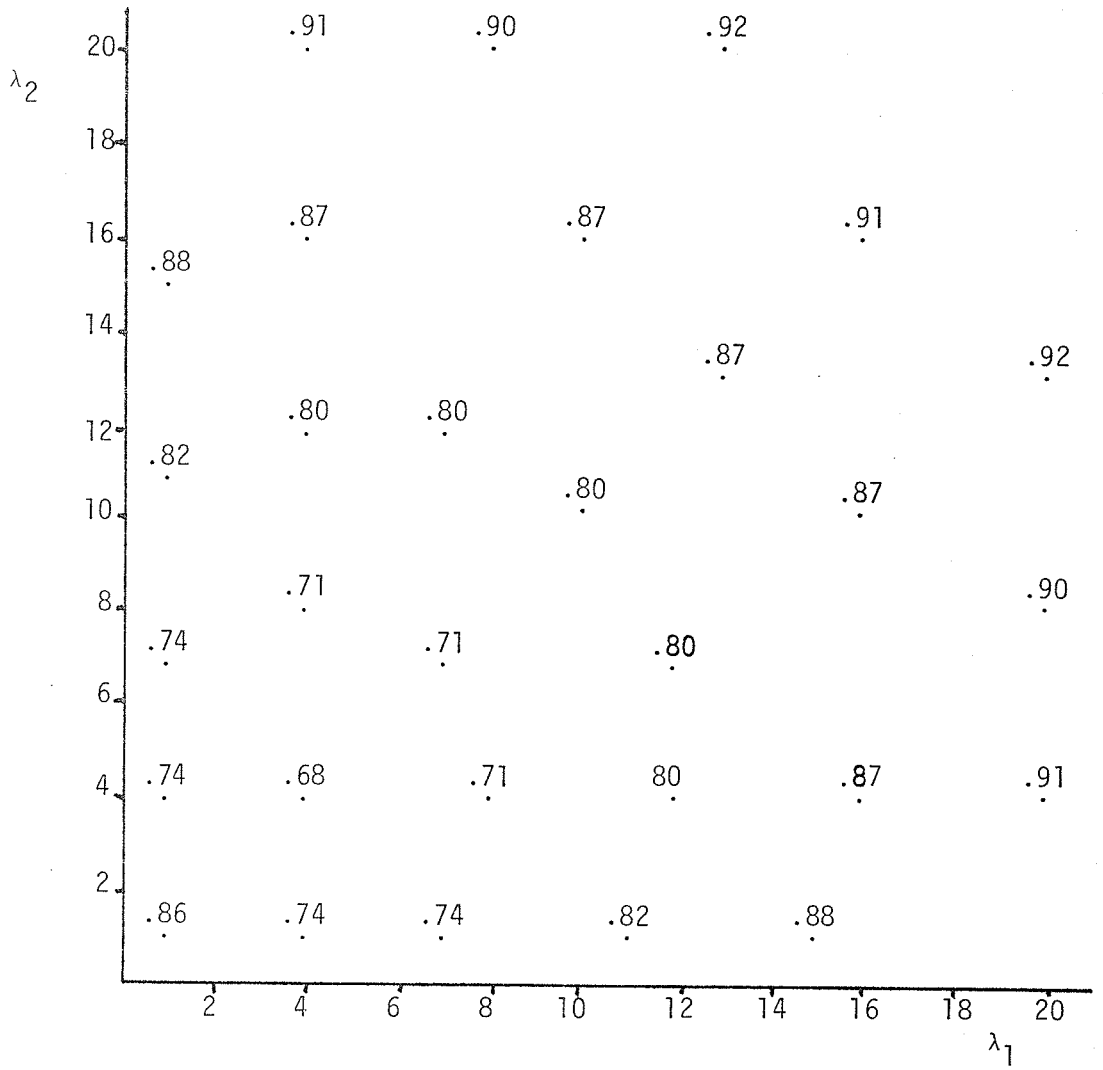


Figure 22

p=2. Prior information: $(\mu_i,\beta_i)=(4,2)$, i=1,2. Values of (expected size of C*)/(expected size of C^0).

CHAPTER 4

ROBUST BAYES ESTIMATION OF MULTINOMIAL PROPORTIONS

1. Introduction

1.1. History

The usual estimator of the multinomial parameter θ is $\delta^0(X) = X/N$, which is the MLE and the MVUE. Johnson (1971) has shown δ^0 to be admissible under loss $L_1(\delta,\theta) = N\sum_{i=1}^{p}(\delta_i-\theta_i)^2$, so unlike the Poisson problem, estimators can not be found which improve uniformly in risk over δ^0. Note that $R(\delta^0,\theta) = 1-\sum_{i=1}^{p}\theta_i^2$, and the reason why δ^0 is admissible appears to be its small risk at the extreme points of the parameter space, say when $\theta_k = 1$ and $\theta_j = 0$ for $j \neq k$. Nevertheless we can hope to improve upon δ^0 in a robust Bayesian sense in the interior of the parameter space.

Since we will want to input prior information into an estimator, some Bayesian procedures for estimating θ will now be discussed. The conjugate prior for θ is the Dirichlet distribution, and this prior and the corresponding Bayes estimator (under loss L_1) will be discussed in the next section. Other Bayesian procedures have been proposed which are based on transforming the parameter θ and placing a normal prior on the transformed parameter. For example, Leonard (1972) performs the transformation

$$\xi_i = \ell n\ \theta_i - \ell n\,(1-\theta_i),\quad i = 1,\ldots,p.$$

He then assumes $\xi_1,\ldots,\xi_p$ are independent and identically distributed with $\xi_i \sim N(\mu,\sigma^2)$, and then puts priors on the parameters μ and σ^2. Using this two stage prior, he proposes estimators which are approximately modes of the posterior distribution. Novick, Lewis and Jackson (1973) perform a similar analysis by first using an arc sin transformation on θ.

Fienberg and Holland (1973) develop estimators using prior information which are similar to estimators discussed in this chapter. As will be shown later, they first consider the conjugate Bayes estimator, and from this they derive estimators with shrinking constants that are functions of the observations $X_1,\ldots,X_p$. These estimators are viewed as a Bayesian technique of smoothing contingency tables, that is, eliminating zeros from the table. They discuss estimators which shrink toward a priori selected means or means that are determined from the data. To evaluate different estimators, they mainly consider the asymptotic situation where N and θ simultaneously approach infinity such that N/θ is a constant. In this situation, called the asymptotics of sparce multinomials, risk functions of estimators within a certain class are compared.

1.2. Need for a robust Bayes estimator

In this section the conjugate Bayes estimator of the multinomial parameter θ will be evaluated with respect to prior robustness. If $X \sim$ multinomial (N,θ), then the conjugate prior for θ is the Dirichlet distribution, whose density is

$$\pi(\theta) = \Gamma(K) \prod_{i=1}^{p} \frac{\theta_i^{K\gamma_i - 1}}{\Gamma(K\gamma_i)}, \quad 0 \leq \theta_i \leq 1 \quad \text{for all } i, \quad \sum_{i=1}^{p} \theta_i = 1,$$

with parameters K and $\gamma = (\gamma_1, \ldots, \gamma_p)$. The posterior distribution for θ is Dirichlet with parameters K+N and $(X+K\gamma)/(N+K) = ((X_1+K\gamma_1)/(N+K), \ldots, (X_p+K\gamma_p)/(N+K))$, and the Bayes estimator of a component of θ, say θ_i, under loss L_1 is

$$\delta_i^B(X) = \frac{X_i + K\gamma_i}{N+K}.$$

It is of interest to observe how robust the Dirichlet prior is with respect to uncertain prior information concerning θ. As in the Poisson estimation problem, robustness of a Bayes estimator with respect to the prior tail is desired, and one way to evaluate this robustness is to inspect the estimator's risk function away from the prior region. The risk function of δ^B can be evaluated; it is

$$R(\delta^B, \theta) = \left(\frac{N}{N+K}\right)^2 \left(1 - \sum_{i=1}^{p} \theta_i^2\right) + N\left(\frac{K}{N+K}\right)^2 \sum_{i=1}^{p} (\theta_i - \gamma_i)^2.$$

(For comparison purposes note that the risk function of δ^0 is $R(\delta^0, \theta) = 1 - \sum_{i=1}^{p} \theta_i^2$.) Here γ is the prior mean of θ, and as θ moves away from the prior mean, $N(K/(N+K))^2 \sum_{i=1}^{p} (\theta_i - \gamma_i)^2$ is the dominant term in the risk of δ^B. At an extreme point of the parameter space, where $\theta_i = 1$ and $\theta_j = 0$ for $j \neq i$,

$$R(\delta^B,\theta) = N(\frac{K}{N+K})^2[\sum_{j\neq i} \gamma_j^2 + (1-\gamma_i)^2],$$

while $R(\delta^0,\theta) = 0$. Thus δ^B can have significantly larger risk than δ^0 away from the prior mean and will perform the worst in risk compared to δ^0 at an extreme point.

How poorly δ^B performs in risk away from the prior mean depends to a good extent on the prior parameters K and γ and the number of trials N. Note that

$$N(\frac{K}{N+K})^2 \sum_{i=1}^{p}(\theta_i-\gamma_i)^2 \leq N \sum_{i=1}^{p}(\theta_i-\gamma_i)^2,$$

and the right term is approximately achieved when K is chosen a priori much larger than N. Large values of K correspond to strong prior information about θ (we will show this later), and in this case δ^B will be most sensitive (in terms of risk) to parameter values away from the prior region.

Finally, it should be noted the size of the risk decrement of δ^B compared to δ^0 away from the prior mean is bounded, since the parameter space is bounded. Therefore in some cases, δ^B will not have large risk for parameter values that are important to the user. For example, consider the case p = 2, where the prior mean γ_1 is equal to .1. Here the risk of δ^B can be expressed as

$$R(\delta^B,\theta) = (\frac{N}{N+K})^2\theta_1(1-\theta_1) + 2(\frac{K}{N+K})^2(\theta_1-.1)^2.$$

Note that δ^B will only have a much larger risk than δ^0 for values of

θ_1 much larger than .1, and the Bayes estimator is somewhat robust against the error of specifying the prior mean of θ_1 larger than its true value.

We have assumed in the above that the absolute risk $R(\delta,\theta)$ is important. If instead one is concerned with the risk of δ^B relative to the risk of δ^0, then the proportional risk, $R(\delta^B,\theta)/R(\delta^0,\theta)$, may be of interest. In the above example, δ^B has large values of proportional risk for small θ_1, and therefore is not robust using this criteria. In addition, L_1 has been assumed to be the appropriate loss for this problem. If errors in estimating small parameter values are most important, then a loss function such as $N\sum_{i=1}^{p}\theta_i^{-1}(\delta_i-\theta_i)^2$ may be appropriate. Again, in the above example, δ^B will no longer be robust under this loss to values of θ_1 much smaller than .1.

Generally the Bayes estimator from the Dirichlet prior appears sensitive to prior uncertainty in the tail. Estimators will be developed which incorporate prior information, but are insensitive to extreme parameter values far from the prior mean.

2. Development of two robust Bayes estimators

Two estimators will now be developed which accept a certain type of prior information and are robust when the prior information is wrong. The derivation of the first estimator uses techniques similar to those used in Chapter 2. The second estimator is of a similar form to the first, but appears to perform better than the first when the prior information is not chosen symmetrically.

To develop the first robust Bayes estimator, consider the conjugate Bayes estimator δ^B. We would like a robust Bayes estimator to perform like δ^B in a prior region of the parameter space. The conjugate Bayes estimator of θ_i, δ_i^B, may be expressed by

$$\delta_i^B(X) = \gamma_i + (1 - \frac{K}{N+K})(\frac{X_i}{N} - \gamma_i),$$

where γ_i is the prior mean of θ_i. One would like the robust Bayes estimator to shrink towards γ_i like δ_i^B, but somehow control the amount of shrinkage when the observations contradict the prior information. The estimator

$$\hat{\delta}_i(X) = \gamma_i + (1 - \frac{K}{N+K}\, c(X))(\frac{X_i}{N} - \gamma_i)$$

is thus considered.

As in the Poisson derivation of δ^*, assume c is a constant and minimize the risk of $\hat{\delta}$ over all values of c. The risk of $\hat{\delta}$ is

$$R(\hat{\delta},\theta) = (1 - \frac{cK}{N+K})^2(1- \sum_{i=1}^{p} \theta_i^2) + N(\frac{cK}{N+K})^2 \sum_{i=1}^{p} (\theta_i-\gamma_i)^2,$$

and minimizing $R(\hat{\delta},\theta)$ with respect to c gives the optimal c value to be

$$c' = (1 + \frac{N}{K}) \frac{1- \sum_{j=1}^{p} \theta_i^2}{1- \sum_{i=1}^{p} \theta_i^2+N \sum_{i=1}^{p} (\theta_i-\gamma_i)^2}.$$

Finally c' is estimated by its MLE and the resulting expression is

substituted into $\hat{\delta}$. The estimator that is obtained is defined componentwise by

$$\delta_i'(X) = \gamma_i + \left(1 - \frac{1 - \sum_{j=1}^{p} \hat{\theta}_j^2}{1 - \sum_{j=1}^{p} \hat{\theta}_j^2 + N \sum_{j=1}^{p} (\hat{\theta}_j - \gamma_j)^2}\right)(\hat{\theta}_i - \gamma_i), \quad i = 1,\ldots,p,$$

where $\hat{\theta}_i = X_i/N$, $i = 1,\ldots,p$. Fienberg and Holland (1973) used this derivation to find δ'.

A robust Bayes estimator should accept some type of prior information and show improvement in risk over the MVUE δ^0 in a region of the parameter space. In the Poisson case δ^* accepts two types of prior information, prior means or guesses of the parameters $\lambda_1,\ldots,\lambda_p$, and prior parameters $\beta_1,\ldots,\beta_p$, which reflect the accuracy of the guesses. One would like to input a similar type of prior information into the robust Bayes multinomial estimator. First note that the prior mean and variance of θ_i, from the Dirichlet prior, are $E(\theta_i) = \gamma_i$, and $Var(\theta_i) = \gamma_i(1-\gamma_i)/(K+1)$. Thus $\gamma = (\gamma_1,\ldots,\gamma_p)$ is the prior guess at θ and larger values of K reflect more precise information about γ. (How to choose γ and K will be discussed in Section 5). The estimator δ_i' shrinks $\hat{\theta}_i$ towards γ_i, but K does not appear in the estimator. One can put K into the estimator by insuring that δ_i' shrinks towards γ_i no more than the shrinkage amount of δ_i^B. Then δ_i' is modified to

$$\delta_i^*(X) = \gamma_i + (1 - \min\{\frac{K}{N+K}, \frac{1 - \sum_{j=1}^{p} \hat{\theta}_j^2}{1 - \sum_{j=1}^{p} \hat{\theta}_j^2 + N \sum_{j=1}^{p} (\hat{\theta}_j - \gamma_j)^2}\})(\hat{\theta}_i - \gamma_i).$$

Note that a similar truncation of the shrinkage constant was made in the derivation of the Poisson estimator δ^*. Note also that the estimator proposed by Fienberg and Holland, δ', is equivalent to δ^* with K set equal to infinity (corresponding to precise prior information.

To gain some insight into the performance of δ^* in the prior region, assume p is large and that we can approximate the shrinking constant of δ^* in terms of the prior parameters. Assuming the Dirichlet prior model, the mean and variance of $\hat{\theta}_i = X_i/N$ under the marginal distribution of X_i can be found to be

$$E(\hat{\theta}_i) = \gamma_i$$

and

$$Var(\hat{\theta}_i) = \frac{\gamma_i(1-\gamma_i)}{K+1} (\frac{N+K}{K}).$$

Now consider the shrinking constant of δ^*,

$$1 - c^*(X) = 1 - \min\{\frac{K}{N+K}, \frac{1 - \sum_{j=1}^{p} \hat{\theta}_j^2}{1 - \sum_{j=1}^{p} \hat{\theta}_j^2 + N \sum_{j=1}^{p} (\hat{\theta}_j - \gamma_j)^2}\}.$$

When p is large,

$$(4.1)\qquad \frac{1-\sum_{j=1}^{p}\hat{\theta}_j^2}{1-\sum_{j=1}^{p}\hat{\theta}_j^2+N\sum_{j=1}^{p}(\hat{\theta}_j-\gamma_j)^2} \cong \frac{E[1-\sum_{j=1}^{p}\hat{\theta}_j^2]}{E[1-\sum_{j=1}^{p}\hat{\theta}_j^2]+E[N\sum_{j=1}^{p}(\hat{\theta}_j-\gamma_j)^2]},$$

and one can show that

$$(4.2)\qquad E[1-\sum_{j=1}^{p}\hat{\theta}_j^2] = \frac{K(N+1)}{N(K+1)}\,(1-\sum_{j=1}^{p}\gamma_j^2)$$

and

$$(4.3)\qquad E[N\sum_{j=1}^{p}(\hat{\theta}_j-\gamma_j)^2] = \frac{K+N}{K+1}\,(1-\sum_{j=1}^{p}\gamma_j^2).$$

Thus for large p and correct prior information, the shrinking constant of δ^*, $1-c^*(X)$, is approximately a function of $1-\sum_{j=1}^{p}\gamma_j^2$. This suggests that the risk of δ^* relative to the risk of δ^0 in the prior region depends on the selection of $\gamma_1,\ldots,\gamma_p$. If one of the γ_i's is chosen near one, then $1-\sum_{i=1}^{p}\gamma_i^2$ is very small; if $\gamma_1,\ldots,\gamma_p$ are chosen symmetrically, that is, if $\gamma_1 =\ldots= \gamma_p = p^{-1}$, then $1-\sum_{j=1}^{p}\gamma_j^2 = 1-p^{-1}$. Therefore the prior mean γ may determine not only the location of the improvement region (over δ^0) but the amount of improvement in risk realized near the prior mean. This problem will be illustrated in Section 3. Therefore we now consider an estimator of a similar form to δ^* that appears to possess a shrinking constant independent of $\gamma_1,\ldots,\gamma_p$.

Let us again consider estimators of the form

$$\delta_i(X) = \gamma_i + (1-c(X))(\hat{\theta}_i-\gamma_i), \quad i = 1,\ldots,p,$$

but consider a different choice for the shrinking constant. Leonard (1977) and Good (1965) both considered shrinking constants that are functions of

$$\sum_{j=1}^{p} \gamma_j^{-1}(\hat{\theta}_j-\gamma_j)^2.$$

When p is large and the prior information is correct, one can show as in (4.1), (4.2), and (4.3), that

$$\sum_{j=1}^{p} \gamma_j^{-1}(\hat{\theta}_j-\gamma_j)^2 \cong \frac{1}{K+1}\frac{N+K}{N}\sum_{j=1}^{p}(1-\gamma_j)$$

$$= \frac{1}{K+1}\frac{N+K}{N}(p-1).$$

Thus for large p, this expression is approximately a constant, not depending on γ. An estimator having a shrinking constant that is a function of this expression should (intuitively) perform equally well (in terms of risk) in the prior region for all selections of $\gamma_1,\ldots,\gamma_p$.

To construct an alternative estimator, recall that it is desirable for an estimator to approximate some natural Bayes estimator when the prior information is correct. The second estimator considered is

$$\tilde{\delta}_i(X) = \gamma_i+(1-\min\{\frac{K}{N+K}, \frac{p-1}{N\sum_{j=1}^{p}\gamma_j^{-1}(\hat{\theta}_j-\gamma_j)}\})(\hat{\theta}_i-\gamma_i),$$

$i = 1,\ldots,p$. Note that for large p, under the Dirichlet prior,

$$\frac{p-1}{N\sum_{j=1}^{p}\gamma_j^{-1}(\hat{\theta}_j-\gamma_j)^2} \cong \frac{K+1}{N+K}.$$

Thus for large p, $\tilde{\delta}$ is similar to δ^B, the conjugate Bayes estimator, when the prior information is correct. The shrinking constant is truncated to the conjugate Bayes shrinking constant for two reasons. First, as mentioned in the derivation of δ^*, it allows us to input the prior parameter K. Second, the term $(p-1)/[N\sum_{j=1}^{p}\gamma_j^{-1}(\hat{\theta}_j-\gamma_j)^2]$ is unstable for values of $\hat{\theta}_1,\ldots,\hat{\theta}_p$ near their respective prior means $\gamma_1,\ldots,\gamma_p$, and the truncation removes that instability.

We have now developed two possible robust Bayes estimators of θ, δ^* and $\tilde{\delta}$. The shrinking constants of the two estimators are respectively

$$1-c^*(X) = 1-\min\{\frac{K}{N+K}, \frac{1-\sum_{j=1}^{p}\hat{\theta}_j^2}{1-\sum_{j=1}^{p}\hat{\theta}_j^2+N\sum_{j=1}^{p}(\hat{\theta}_j-\gamma_j)^2}\}$$

and

$$1-\tilde{c}(X) = 1-\min\{\frac{K}{N+K}, \frac{p-1}{N\sum_{j=1}^{p}\gamma_j^{-1}(\hat{\theta}_j-\gamma_j)^2}\}.$$

When the prior information has been specified correctly, both shrinking constants are estimates of K/(N+K), and both estimators hopefully will be similar to the Bayes estimator δ^B. When the true value of θ is inconsistent with the prior information, most likely one observation

$\hat{\theta}_i$ will be far from its prior mean γ_i, and both of the shrinking constants will be close to one. In this situation, both δ^* and $\tilde{\delta}$ will approximately be equal to $\delta^0(X)$, the MVUE. In the next section we will evaluate how well the robust estimators emulate δ^B in the prior region and how the robust estimators compare with δ^0 outside of the prior region.

3. Evaluation

3.1. Introduction

In this section the robust Bayes estimators δ^* and $\tilde{\delta}$ are first shown to have smaller risk than δ^0 in a region of the parameter space and not much larger risk elsewhere. Through computer simulation, risks are found in the case p = 2 (binomial) for several different sets of prior means. It is shown how the prior information γ and K is reflected in the risks of δ^* and $\tilde{\delta}$ in a prior region. The risk of δ^* is compared with the risk of Fienberg and Holland's estimator δ', which does not allow for a prior variance input. Finally the robust Bayes estimators are shown to be more robust (with regard to risk) than δ^B with respect to parameter values outside the prior region.

Next, the performance of δ^* and $\tilde{\delta}$ is analyzed with respect to the choice of the prior mean γ. It was shown in the previous section that the shrinking constant of δ^*, as $p \to \infty$, is a function of $1 - \sum_{i=1}^{p} \gamma_i^2$ when the prior information has been selected correctly. In this situation it is possible for one component's prior mean to

have the greatest influence on the behavior of δ^*. It is shown in an example how the selection of the prior mean affects the proportional improvement in risk of δ^* over δ^0 at γ. The estimator δ^* appears to perform best in the prior region when $\gamma_1,\ldots,\gamma_p$ are chosen similar in size. In contrast, the alternative robust Bayes estimator $\tilde{\delta}$ appears to have a proportional improvement in risk (over δ^0) in the prior region that is generally insensitive to the choice of γ. It will be indicated through examples that $\tilde{\delta}$ is a preferable estimator to δ^* when $p > 2$ and the prior means $\gamma_1,\ldots,\gamma_p$ are dissimilar.

Finally a correspondence is made between the multinomial estimator δ^* and the Poisson estimator δ^* discussed in Chapter 2. We consider the limiting situation where the multinomial distribution approaches the distribution of independent Poisson random variables. Under a particular loss, it is shown that the risk of the multinomial δ^* approaches the risk of an estimator similar to the Poisson δ^*. The correspondence between the two estimators will be used in constructing a robust Bayesian confidence region for the multinomial parameter θ and understanding the role of the prior parameter K in the robust Bayes multinomial estimators.

3.2. Risk

In this section, comparisons are made between the risk functions of the robust Bayes estimators δ^* and $\tilde{\delta}$, Fienberg and Holland's estimator δ', the conjugate Bayes estimator δ^B, and the MVUE δ^0. It has already been noted that the risk functions of δ^B and δ^0 are respectively

$$R(\delta^B,\theta) = \left(\frac{N}{N+K}\right)^2\left(1-\sum_{i=1}^{p}\theta_i^2\right) + N\left(\frac{K}{N+K}\right)^2 \sum_{i=1}^{p}(\theta_i-\gamma_i)^2,$$

and

$$R(\delta^0,\theta) = 1-\sum_{i=1}^{p}\theta_i^2.$$

The risk functions of δ^*, $\tilde{\delta}$ and δ' are cumbersome to calculate directly, so they are calculated through computer simulation. In each example that is presented, at least 5000 occurrences of X are simulated, and the risks found have a standard error of less than 5 per cent. In this section, we consider the case N = 15, p = 2. Note that the loss L_1 can be written as

$$L_1(\delta,\theta) = [\delta_1-\theta_1]^2 + [(1-\delta_1)-(1-\theta_1)]^2$$

$$= 2(\delta_1-\theta_1)^2.$$

(For all of the estimators considered, $\sum_{j=1}^{p}\delta_j = 1.$) Therefore, all of the risks in this case are plotted as functions of θ_1.

Figures 23 and 24 show the risks of δ^*, δ', δ^B and δ^0 for different sets of prior means. The parameter γ_1, the prior mean of θ_1, is .5 for Figure 23 and .1 in Figure 24. Additionally, the prior parameter K = 5 is inputted into the estimators δ^B and δ^*.

Let us first compare the risk of the robust Bayes estimator δ^* with the risk of MVUE δ^0. In both figures, δ^* has a significantly smaller risk than δ^0 in an interval about the prior mean. Outside of this improvement region, δ^* can have a larger risk than δ^0, but the risk decrement is bounded and as θ_1 approaches 0 and 1, the two risk functions become identical.

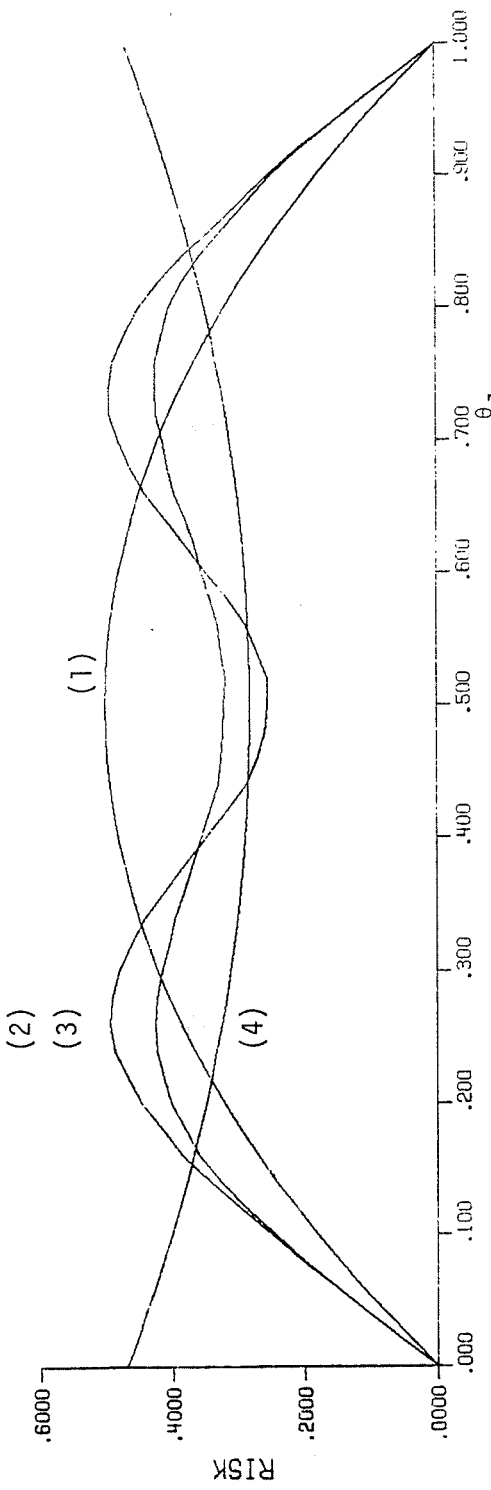


Figure 23

p=2, N=15. Risks of (1) δ^0, (2) δ', γ_1 =.5, (3) δ^*, γ_1=.5, K=5, and (4) δ^B, γ_1=.5, K=5.

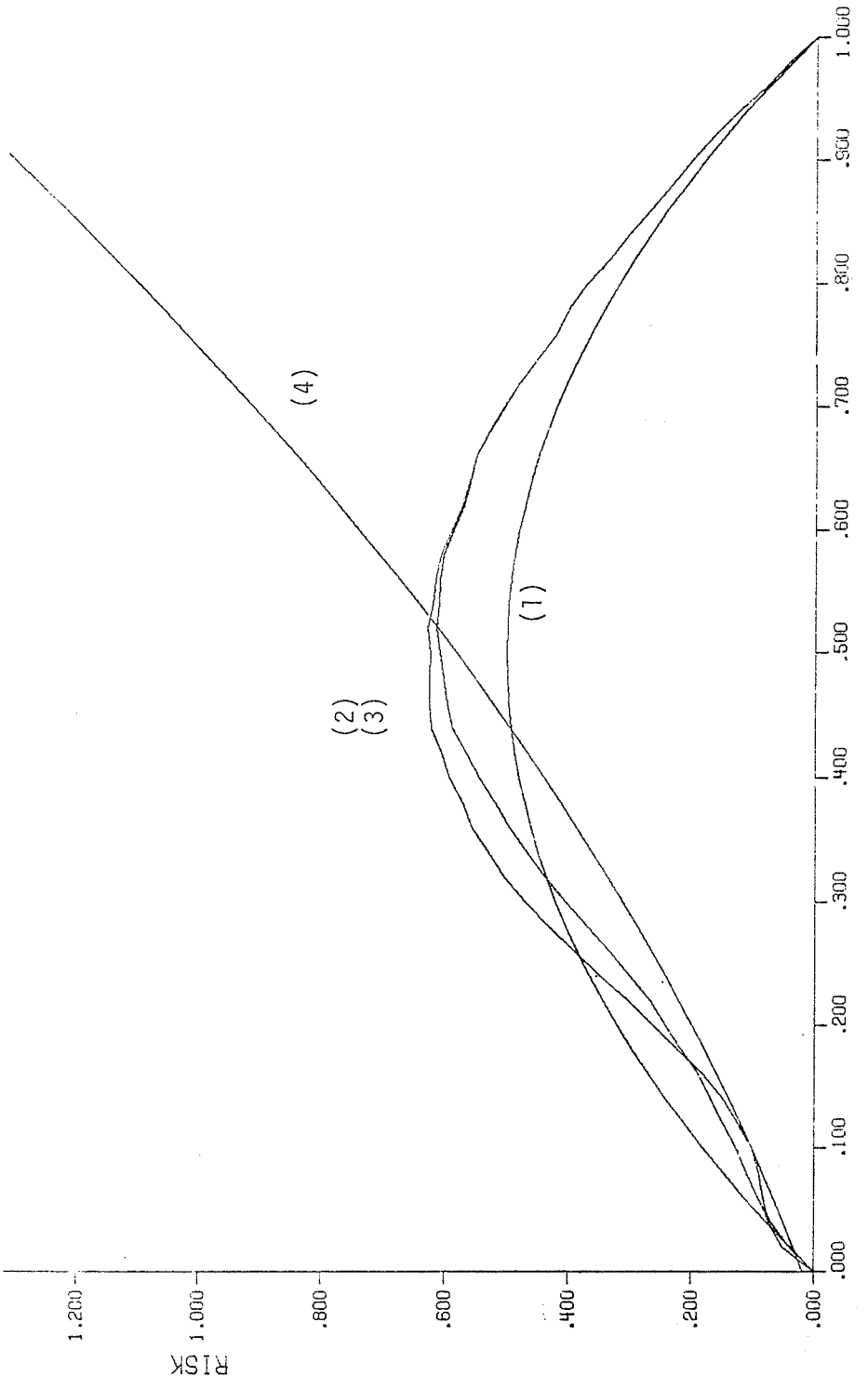


Figure 24

p=2, N=15. Risks of (1) δ^0, (2) δ', γ_1=.1, (3) δ^*, γ_1=.1, K=5, and (4) δ^B, γ_1=.1, K=5.

The conjugate Bayes estimator δ^B, using the same prior information as δ^*, also shows risk improvement over δ^0 about the prior mean, and the amount of improvement appears to be more significant than the amount of improvement of δ^*. But the risk of δ^B increases quadratically away from γ_1, showing that δ^B is very sensitive to parameter values far from the prior mean. Thus δ^* appears to be more robust than δ^B with respect to uncertainty in the extreme portions of the parameter space.

Recall that δ' (the estimator proposed by Fienberg and Holland) is identical to δ^* except that the shrinking constant of δ' is not truncated. Figures 23 and 24 show the effect of truncating the shrinking constant, i.e. inputting a prior value of K into the estimator δ^*. In Section 5, we will describe ways of obtaining prior values of K, but smaller values of K correspond to less precise prior information about θ. The estimator δ', in effect, chooses a prior value of $K = \infty$, and the result of choosing the smaller value of $K = 5$ in δ^* is to flatten the risk of δ' towards the risk of δ^0. If strong prior information exists about θ, then large values of K will be used in δ^*, and the risk functions of δ' and δ^* will be almost identical. But if less precise prior information exists, one may desire to use a smaller value of K in δ^* and obtain a relatively flat risk function.

Figure 25 compares the risk functions of the two robust Bayes estimators δ^* and $\tilde{\delta}$, and δ^0. The setting is the same as in Figures 23 and 24, except that the prior mean γ_1 is equal to .3.

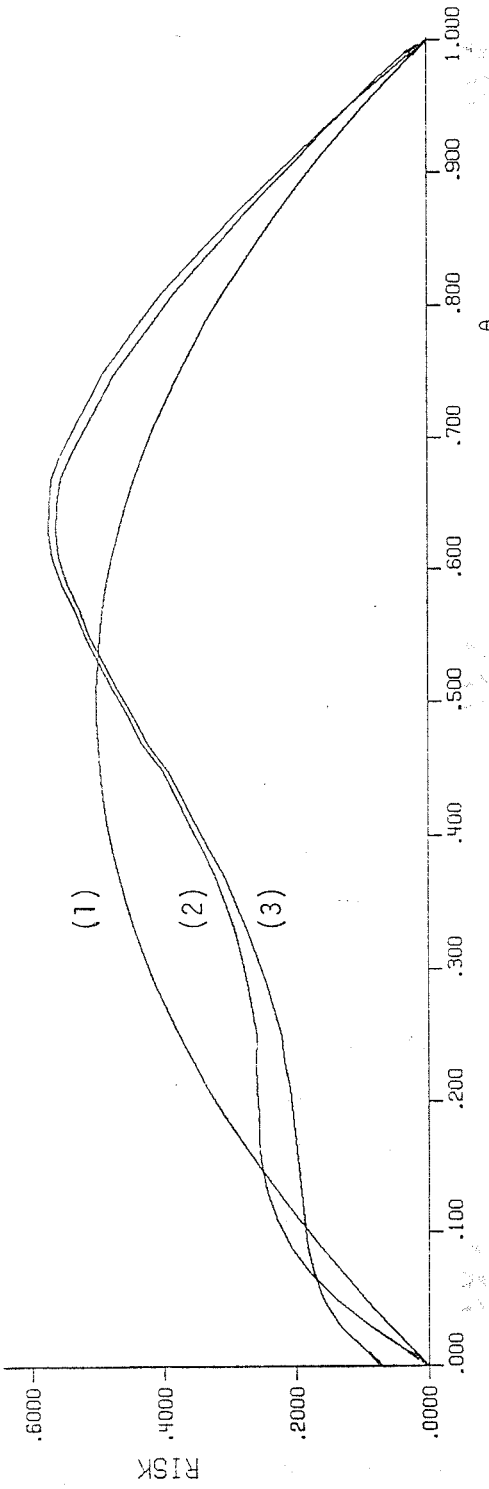


Figure 25

p=2, N=15. Risks of (1) δ^0, (2) δ^*, γ_1=.3, K=5, and (3) $\tilde{\delta}$, γ_1=.3, K=5.

It appears that $\tilde{\delta}$ has a slightly larger region of improvement than δ^*, but $\tilde{\delta}$ has larger risk than δ^* for small values of θ_1. In fact, it can be shown in this example that

$$\lim_{\theta_1 \to 0} R(\tilde{\delta},\theta) \doteq .065,$$

while

$$\lim_{\theta_1 \to 0} R(\delta^*,\theta) = 0.$$

If the proportional risk, $R(\delta,\theta)/R(\delta^0,\theta)$, is used as a criterion, then the estimator δ^* will perform much better than $\tilde{\delta}$ for small values of θ_1. (In general δ^* will perform better than $\tilde{\delta}$ with respect to proportional risk at an extreme point of the parameter space.) However, in this example and other examples that we have studied, both robust estimators appear to be reasonable alternatives to δ^0 (under loss L_1) when prior information is available.

3.3. Performance in the case of asymmetric prior information

In Section 3.2, it has been indicated that the robust Bayes estimators δ^* and $\tilde{\delta}$ have smaller risk than δ^0 in a prior region of the parameter space. Here the risks of δ^* and $\tilde{\delta}$ in the prior region are analyzed further; we are especially interested in the risk improvement of δ^* and $\tilde{\delta}$ when there exist differences in the prior means $\gamma_1,\ldots,\gamma_p$.

In Section 2, it was shown that, for large p and correct prior information, the shrinking constant of δ^* is approximately a function of $1-\sum_{i=1}^{p}\gamma_i^2$. In contrast, the shrinking constant of $\tilde{\delta}$ was shown to be

approximately a constant (not dependent on γ) in the same large p situation. Intuitively, the amount of risk improvement of $\tilde{\delta}$ near the prior mean should be less sensitive than δ^* to the choice of γ.

Let us illustrate the difference between the two estimators in an example. In Table 3, values of the proportional risk, $R(\delta,\theta)/R(\delta^0,\theta)$, have been calculated when $\theta = \gamma$ for different selections of the prior mean γ and different dimensions p. We have set N = 15, and set K = 10 in the two estimators. Looking at the p = 4 case, one sees that the proportional risk of δ^*, and therefore the proportional improvement of δ^*, depends to some extent on the selection of the prior mean. For example, if γ = (.25, .25, .25, .25), the value of the proportional risk is .44, while if γ = (.49, .49, .01, .01), the proportional risk has increased to .56. One notes that if $\gamma = (p^{-1},\ldots,p^{-1})$, then values of the proportional risk at γ decrease as p increases. Thus as more θ_i's are estimated simultaneously, δ^* appears to show more proportional risk improvement over δ^0 about this prior mean. But for general γ, it appears that the true dimensionality of the improvement of δ^* near the prior mean is just the number of components of γ significantly larger than zero. For example, the proportional risk of δ^* at the point (.49, .49, .01, .01) (p = 4) is approximately the same as the proportional risk at the point (.5,.5) (p = 2).

Table 3 also shows values of the proportional risk for the estimator $\tilde{\delta}$ at the same points with the same prior information. The

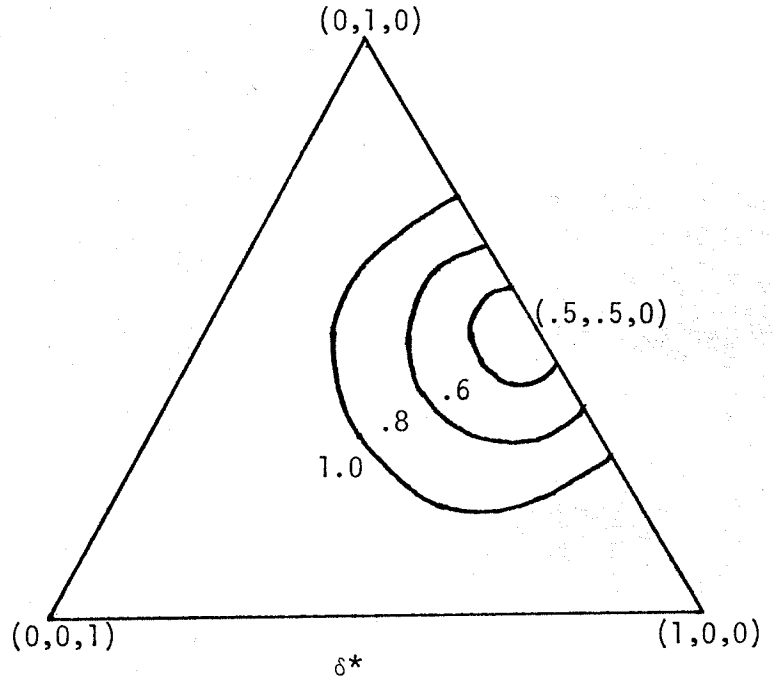
Table 3

Values of proportional risk at the prior mean for the estimators δ^* and $\tilde{\delta}$. N=15, K=10.

		$R(\delta,\lambda)/R(\delta^0,\lambda)$ at γ	
p	γ	δ^*	$\tilde{\delta}$
2	(.5,.5)	.56	.46
3	(.49,.49,.02)	.55	.41
	(.33,.33,.34)	.46	.39
4	(.25,.25,.25,.25)	.44	.38
	(.32,.32,.32,.04)	.48	.39
	(.49,.49,.01,.01)	.56	.40

proportional risk values for $\tilde{\delta}$ appear to be more robust or insensitive to the choice of prior mean than δ^*. Generally it appears that for p = 4, $\tilde{\delta}$ will show approximately a 60-62% improvement in risk over the risk of δ^0 at the prior mean, regardless of the value of γ.

As another example of this asymmetry problem consider the situation where p = 3, N = 15 and the prior information is K = 10, $(\gamma_1,\gamma_2,\gamma_3) = (.49, .49, .02)$. Figure 26 shows contours of constant values of proportional risk over the simplex of parameter values for the estimators δ^* and $\tilde{\delta}$. In this example, θ_1 and θ_2 have a prior standard deviation of .15, and θ_3 has a prior standard deviation of .04. Therefore the chosen values of γ and K imply that one has much stronger prior information about θ_3 than about θ_1 and θ_2. Intuitively the region of substantial risk improvement of a Bayesian estimator over δ^0 should be more concentrated in the dimension of the coordinate with the smaller prior variance. The region of significant improvement should correspond to the region where the user thinks the unknown parameter lies, and the size of this latter region in different dimensions corresponds to prior standard deviations. Thus in this example, the region of risk improvement should be smallest in the dimension of θ_3. Now compare, for example, the contour of proportional risk equal to .6 for the two estimators. Looked at from the prior mean, the size of the improvement region for $\tilde{\delta}$ is smallest in the θ_3 dimension (towards the point (0,0,1)). In contrast, the corresponding improvement



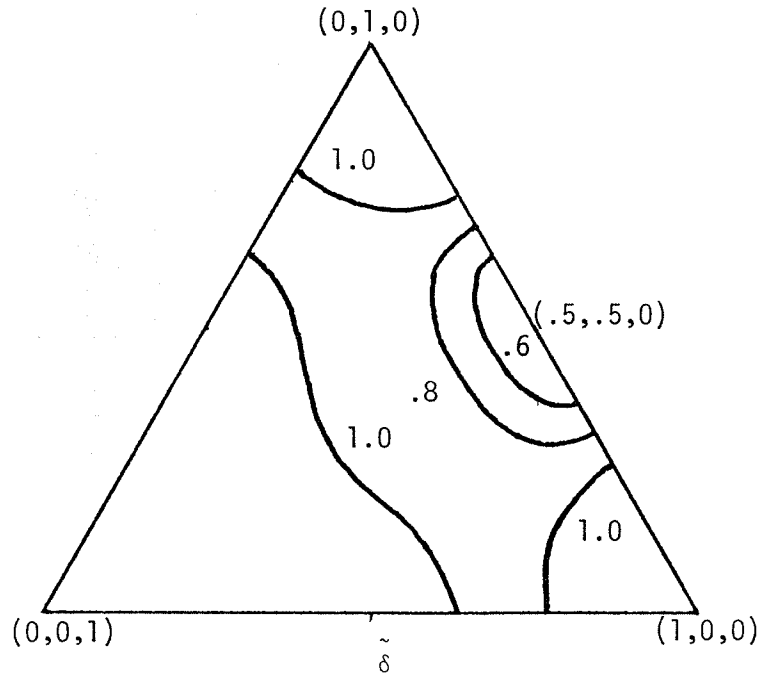


Figure 26

p=3, N=15. Prior information: γ=(.49,.49,.02), K=10. Contours of constant values of proportional risk for δ^* and $\tilde{\delta}$.

region for δ^* actually is smallest in the θ_1 and θ_2 dimensions (towards the points (0,1,0) and (1,0,0)). The estimator $\tilde{\delta}$ appears to perform much better than δ^* when θ_3 is close to its prior mean $\gamma_3 = .02$ and θ_1 and θ_2 are a moderate distance from their prior means. Thus the prior information that is inputted appears to be best reflected in the improvement region of $\tilde{\delta}$.

In summary, both robust Bayes estimators δ^* and $\tilde{\delta}$ show substantial risk improvement over δ^0 in a region of the parameter space. When the prior means $\gamma_1,\ldots,\gamma_p$ are chosen to be about the same size, then both estimators appear to be equally attractive alternatives to δ^0. It appears, however, that $\tilde{\delta}$ is better than δ^* when there exists considerable asymmetry among $\gamma_1,\ldots,\gamma_p$. Unlike δ^*, the proportional improvement in risk of $\tilde{\delta}$ in the prior region appears to be approximately the same for symmetric and nonsymmetric selections of prior means.

3.4. Equivalence of two estimators

It is well known that if N approaches infinity and $\theta_1,\ldots,\theta_{p-1}$ all approach zero at a particular rate, the distribution of a multinomial random variable approaches the distribution of p-1 independent Poisson random variables. Thus there is a close connection between the Poisson estimation problem described in Chapter 2 and the multinomial estimation problem described here. Also note that the derivation of δ^* in the multinomial problem is similar to the derivation of δ^* in the Poisson problem. In both problems, an estimator which shrinks the MVUE towards the prior

mean (like the conjugate Bayes estimator) is considered, and a shrinking constant that is a function of X is found by taking the MLE of an "optimal" shrinking constant that is a function of the unknown parameters.

In order to establish a correspondence between the multinomial estimator δ^* and the Poisson estimator, define the new multinomial loss

$$L_2(\delta,\theta) = N^2 \sum_{i=1}^{p-1} (\delta_i - \theta_i)^2.$$

In the limiting situation to be described, the multinomial loss L_2 will be equivalent to the loss L_1 in estimating p-1 Poisson parameters (see Chapter 2). Theorem 4 shows that in this limiting situation, the risk (under loss L_2) of the multinomial estimator δ^* is equivalent to the risk of a Poisson estimator of a form similar to the robust Poisson estimator δ^*. Note that in the statement of the theorem, we let K approach infinity such that $N/K \to D$, where D is a constant. It will be explained in Section 5 that the quantity N/K is an indication of the strength of the prior information, so it is natural to keep this quantity a constant in this asymptotic situation.

Theorem 4: Let X be multinomial with parameters N, $\gamma_1,\ldots,\gamma_p$ and consider the estimator

$$\delta_i^*(X) = \gamma_i + (1 - \min\{\frac{K}{N+K}, \frac{1 - \sum_{j=1}^{p} \hat{\theta}_j^2}{1 - \sum_{j=1}^{p} \hat{\theta}_j^2 + N \sum_{j=1}^{p} (\hat{\theta}_j - \gamma_j)^2}\})(\hat{\theta}_i - \gamma_i), \quad i = 1,\ldots,p,$$

where $\hat{\theta}_i = X_i/N$, $i = 1,\ldots,p$. Consider the situation where

(4.4) $\quad \theta_1,\ldots,\theta_{p-1} \to 0$, $\gamma_1,\ldots,\gamma_{p-1} \to 0$, $N \to \infty$, $K \to \infty$,

such that $N\theta_i \to \lambda_i$, $N\gamma_i \to \mu_i$, $0 < \lambda_i, \mu_i < \infty$, $i = 1,\ldots,p-1$, and $N/K \to D$. Then

$$N^2 \sum_{i=1}^{p-1} E(\delta_i^*(X) - \theta_i)^2 \to \sum_{i=1}^{p-1} E(\delta_i^{**}(X) - \lambda_i)^2,$$

where

$$\delta_i^{**}(X) = \mu_i + (1 - \min\{\frac{1}{D+1}, \frac{2\sum_{j=1}^{p-1} X_j}{2\sum_{j=1}^{p-1} X_j + \sum_{j=1}^{p-1} (X_j - \mu_j)^2 + (\sum_{j=1}^{p-1} (X_j - \mu_j))^2}\})(X_i - \mu_i),$$

and the latter expectation is taken over the distribution of $p-1$ independent Poisson random variables with means $\lambda_1,\ldots,\lambda_{p-1}$.

Proof: See Appendix.

Let us analyze the Poisson estimator δ^{**} further by considering the term

$$(4.5) \qquad \min\{\frac{1}{D+1}, \frac{2\sum_{j=1}^{p-1} X_j}{2\sum_{j=1}^{p-1} X_j + \sum_{j=1}^{p-1}(X_j-\mu_j)^2 + (\sum_{j=1}^{p-1}(X_j-\mu_j))^2}\}.$$

Assume that λ_i has prior mean μ_i and prior variance τ_i^2, so marginally X_i has mean μ_i and variance $\mu_i + \tau_i^2$. Consider the situation where p approaches infinity. By the law of large numbers,

$$\sum_{j=1}^{p-1} X_j/(p-1) \to \lim_{p\to\infty} \sum_{j=1}^{p-1} \mu_j/(p-1) = \bar{\mu} \quad \text{a.s.},$$

and

$$\sum_{j=1}^{p-1}(X_j-\mu_j)^2/(p-1) \to \lim_{p\to\infty} \sum_{j=1}^{p-1}(\mu_j+\tau_j^2)/(p-1) = \overline{\mu+\tau^2} \quad \text{a.s.},$$

assuming the above limits exist. Now one can write

$$\frac{[\sum_{j=1}^{p-1}(X_j-\mu_j)]^2}{p-1} = \frac{\sum_{j=1}^{p-1}(\mu_j+\tau_j^2)}{p-1}\left[\frac{\sum_{j=1}^{p-1}(X_j-\mu_j)}{(\sum_{j=1}^{p-1}(\mu_j+\tau_j^2))^{1/2}}\right]^2.$$

If the assumptions in Liapunov's Theorem are satisfied (see Rao (1973), p. 127), then

$$\left[\frac{\sum_{j=1}^{p-1}(X_j-\mu_j)}{(\sum_{j=1}^{p-1}(\mu_j+\tau_j^2))^{1/2}}\right]^2 \xrightarrow{\mathcal{L}} U,$$

where U is distributed chi squared with one degree of freedom. Using this result, the expression (4.5) converges in distribution as p

approaches infinity to

$$\min\{\frac{1}{D+1}, \frac{2\bar{\mu}}{2\bar{\mu} + \mu+\tau^2 (1+U)}\}.$$

Since $E(U) = 1$, this term is very similar to

$$\min\{\frac{1}{D+1}, \frac{\bar{\mu}}{\bar{\mu} + \mu+\tau^2}\},$$

and this latter quantity is the asymptotic shrinking term of the robust Bayes estimator

$$\delta^*(X) = \mu+(1-\min\{\frac{1}{D+1}, \frac{\sum_{j=1}^{p-1} X_j}{\sum_{j=1}^{p-1} X_j + \sum_{j=1}^{p-1} (X_j-\mu_j)^2}\})(X-\mu).$$

Thus an equivalence has been shown between the multinomial estimator δ^* and an estimator similar to the Poisson estimator δ^*.

This equivalence will be used in two ways. In Section 4, the problem of constructing a confidence region for θ will be discussed and the work that has been done in constructing regions for Poisson means (Chapter 2) will be used in developing robust Bayesian regions for the multinomial parameter. Second, in the discussion of using the robust estimator (Section 6), the correspondence with the Poisson estimator will help in understanding the relationship of the prior parameter K with the sample size N.

4. Robust Bayes confidence regions

4.1. Introduction

As in the Poisson estimation problem, it is desired to develop a confidence region for θ which will allow the input of prior information and, in some sense, improve upon the classical confidence region. In this section, some methods of constructing confidence regions for θ will be discussed, and in Section 4.2, two robust Bayes confidence regions will be developed.

We will consider simultaneous confidence intervals for $\theta_1,\ldots,\theta_p$, or equivalently a confidence rectangle for θ. For a given j, a method will be described to calculate a confidence interval (π_j^L,π_j^U) for θ_j, $j = 1,\ldots,p$. Then one confidence rectangle for θ is defined by

$$C^1 = \{\theta\colon \pi_j^L \le \theta_j \le \pi_j^U,\ \ j = 1,\ldots,p\}.$$

Note that C^1 ignores the restriction $\sum_{i=1}^{p} \theta_i = 1$. To calculate an alternative confidence rectangle for θ using this restriction, simultaneous confidence intervals need only be found for p-1 components of θ, say $\{\theta_j\colon j \neq k\}$. Then the corresponding confidence rectangle is defined by

$$C_k^2 = \{\theta\colon \pi_j^L \le \theta_j \le \pi_j^U,\ j \neq k,\ 1 - \sum_{j\neq k} \pi_j^U \le \theta_k \le 1 - \sum_{j\neq k} \pi_j^L\}.$$

The confidence rectangle C_k^2 will in general depend on the value of k, or equivalently the set of p-1 components of θ for which simultaneous confidence intervals are first found. To decide on a

value of k, we will always choose the value of k that corresponds to the confidence rectangle C_k^2 with the maximum volume. This confidence rectangle is defined by

$$(4.6) \qquad C^3 = \{\theta: \pi_j^L \leq \theta_j \leq \pi_j^U,\ j \neq k,\ 1 - \sum_{j \neq k} \pi_j^U \leq \theta_k \leq 1 - \sum_{j \neq k} \pi_j^L,$$

$$\text{where } \pi_k^U - \pi_k^L = \min_j \{\pi_j^U - \pi_j^L\}\}.$$

To find C^3, simultaneous confidence intervals are first calculated for the p-1 components of θ which correspond to the maximum widths. Since the width of the remaining confidence interval is

$$\sum_{j \neq k} (\pi_j^U - \pi_j^L),$$

this procedure will maximize the width of all p intervals, and therefore maximize the volume of the confidence rectangle for θ. Note that k in C^3 is defined to be the index of the interval with the minimum width. If k is not unique, it can be chosen to be any one of the indices of the intervals with the minimum width. All of the confidence rectangles discussed in this section will be of the form of C^3.

We now discuss one classical method of constructing a confidence interval for a component of θ, say θ_i. It is well known that the statistic.

$$\frac{N^{1/2}(\hat{\theta}_i - \theta_i)}{(\hat{\theta}_i(1-\hat{\theta}_i))^{1/2}},$$

as N approaches infinity, has an asymptotic standard normal distribution. By solving the equations

$$\frac{N^{1/2}(\hat{\theta}_i-\theta_i)}{(\hat{\theta}_i(1-\hat{\theta}_i))^{1/2}} = \pm z_{\alpha/2}$$

for θ_i, where $z_{\alpha/2}$ is the $100(1-\alpha/2)$ percentile of the standard normal distribution, the following large sample confidence interval is obtained for θ_i:

$$\hat{\theta}_i \pm z_{\alpha/2}[\hat{\theta}_i(1-\hat{\theta}_i)/N]^{1/2}.$$

This interval has asymptotically a probability of $1-\alpha$ of covering θ_i. The corresponding confidence rectangle for θ as defined by (4.6) is

$$C^0(X) = \{\theta: |\hat{\theta}_j-\theta_j| \leq z_{\alpha/2}[\hat{\theta}_j(1-\hat{\theta}_j)/N]^{1/2}, \quad j\neq k,$$

$$|\hat{\theta}_k-\theta_k| \leq z_{\alpha/2} \sum_{j\neq k}[\hat{\theta}_j(1-\hat{\theta}_j)/N]^{1/2}, \text{ where } \hat{\theta}_k(1-\hat{\theta}_k) = \min_j\{\hat{\theta}_j(1-\hat{\theta}_j)\}\}.$$

As with the Poisson situation, we begin the development of a robust Bayes confidence region by finding an approximate credible region using the conjugate prior. Thus assume $\theta \sim$ Dirichlet $(K,\gamma_1,\ldots,\gamma_p)$, so that the posterior distribution of θ is Dirichlet $(N+K, (K\gamma_1+X_1)/(N+K),\ldots,(K\gamma_p+X_p)/(N+K))$. For a given η, a region which bounds approximately $(1-\eta)100\%$ of the posterior distribution will now be found.

First, if K is large, then the Dirichlet $(K,\gamma_1,\ldots,\gamma_p)$ distribution may be approximated by a multivariate normal distribution with mean vector $\gamma = (\gamma_1,\ldots,\gamma_p)$ and covariance matrix $(K+1)^{-1}(D_\gamma - \gamma'\gamma)$, where D_γ is a diagonal matrix with diagonal elements $\gamma_1,\ldots,\gamma_p$. This approximation is motivated first by the fact that the distribution of

$$Y_1 = X_1/(K+1),\ldots,Y_p = X_p/(K+1),$$

where $X = (X_1,\ldots,X_p)$ is multinomial $(K+1,\gamma_1,\ldots,\gamma_p)$, has the same moments up to second order as the Dirichlet distribution. Then a multivariate normal distribution is commonly used to approximate the distribution of $Y_1,\ldots,Y_p$, when K is large.

Using this normal approximation, an ellipsoid can be found which covers $(1-\eta)100\%$ of the posterior distribution. Instead, we will find a rectangle, since the classical region C^0 is a rectangle. If $W_1,\ldots,W_p \sim$ Dirichlet $(K,\gamma_1,\ldots,\gamma_p)$, then by using this approximation, the interval

$$\gamma_i \pm z_{\alpha/2}[\gamma_i(1-\gamma_i)/(K+1)]^{1/2}$$

will cover approximately the middle $(1-\alpha)100\%$ of the distribution of W_i, $i = 1,\ldots,p$. If we temporarily assume that $W_1,\ldots,W_p$ are independent, then the rectangle formed by p-1 of these intervals will cover approximately $(1-\alpha)^{p-1}100\%$ of the joint distribution of $W_1,\ldots,W_p$. (Note that $W_1,\ldots,W_p$ are approximately independent when p is large and the parameters $\gamma_1,\ldots,\gamma_p$ are nearly equal.) Therefore, by choosing α such that $(1-\alpha)^{p-1} = 1-\eta$, the rectangle that covers

approximately $(1-\eta)100\%$ of the posterior distribution of θ will have intervals

$$\frac{X_i+K\gamma_i}{N+K} \pm z_{\alpha/2}(N+K+1)^{-1/2}\left[\frac{X_i+K\gamma_i}{N+K}\left(1-\frac{X_i+K\gamma_i}{N+K}\right)\right]^{1/2}, \quad i=1,\ldots,p-1.$$

(Since the multivariate normal distribution is symmetric about the mean, this rectangle is an approximate HPD region for θ.) This approximate credible region will be used in the next section in developing the robust Bayes confidence regions.

4.2. Development of robust Bayes confidence regions

Consider again the approximate credible region for θ_i given in the last section. It may be expressed as

$$\frac{X_i+K\gamma_i}{N+K} \pm z_{\alpha/2}\left(1-\frac{K+1}{N+K+1}\right)^{1/2}\left[\frac{X_i+K\gamma_i}{N+K}\left(1-\frac{X_i+K\gamma_i}{N+K}\right)/N\right]^{1/2}.$$

There are two major terms in this interval, $(X_i+K\gamma_i)/(N+K)$, the conjugate Bayes estimator of θ_i, and $1-(K+1)/(N+K+1)$, approximately the shrinking constant of this Bayes estimator. To develop a robust Bayes confidence interval, it is natural to substitute one of our recommended robust Bayes estimators (δ_i^* or $\tilde{\delta}_i$) for $(X_i+K\gamma_i)/(N+K)$, and substitute the corresponding shrinking constant from the robust Bayes estimator ($1-c^*(X)$ or $1-\tilde{c}(X)$) for $1-(K+1)/(N+K+1)$. The two resulting confidence intervals are

$$\delta_i^*(X) \pm z_{\alpha/2}(1-c^*(X))^{1/2}[\delta_i^*(X)(1-\delta_i^*(X))/N]^{1/2}$$

and

$$\tilde{\delta}_i(X) \pm z_{\alpha/2}(1-\tilde{c}(X))^{1/2}[\tilde{\delta}_i(X)(1-\tilde{\delta}_i(X))/N]^{1/2}.$$

One final adjustment will be made to the above intervals to obtain the robust Bayes confidence intervals for θ_i that will be recommended. That adjustment is to replace the terms $[\delta_i^*(X)(1-\delta_i^*(X))/N]^{1/2}$ and $[\tilde{\delta}_i(X)(1-\tilde{\delta}_i(X))/N]^{1/2}$ in the width terms of the above intervals by $[\hat{\theta}_i(1-\hat{\theta}_i)/N]^{1/2}$. Recall that the classical confidence interval for θ_i was defined by

$$\hat{\theta}_i \pm z_{\alpha/2}[\hat{\theta}_i(1-\hat{\theta}_i)/N]^{1/2}.$$

This adjustment ensures that the robust Bayes interval will only shrink the width of the classical interval by a factor of $(1-c(X))^{1/2}$, where $1-c(X)$ is the shrinking constant of the robust Bayes estimator. Without this adjustment, it is possible for the above two intervals to shrink the classical interval width an amount larger than $(1-c(X))^{1/2}$. From our experience with robust Bayes intervals for Poisson means (see Chapter 3), it appears that a robust Bayes interval cannot shrink the width of the classical interval by a larger factor than $(1-c(X))^{1/2}$ and still retain a good uniform probability of coverage. Therefore, the recommended robust Bayes confidence intervals for θ_i are

$$\delta_i^*(X) \pm z_{\alpha/2}(1-c^*(X))^{1/2}[\hat{\theta}_i(1-\hat{\theta}_i)/N]^{1/2}$$

and

$$\tilde{\delta}_i(X) \pm z_{\alpha/2}(1-\tilde{c}(X))^{1/2}[\hat{\theta}_i(1-\hat{\theta}_i)/N]^{1/2}.$$

Let C^* and $\tilde{C}$ denote the corresponding confidence rectangles for θ, defined by

$$C^*(X) = \{\theta: |\delta_j^*(X)-\theta_j| \leq z_{\alpha/2}(1-c^*(X))^{1/2}[\hat{\theta}_j(1-\hat{\theta}_j)/N]^{1/2}, j \neq k,$$

$$|\delta_k^*(X)-\theta_k| \leq z_{\alpha/2}(1-c^*(X))^{1/2} \sum_{j \neq k} [\hat{\theta}_j(1-\hat{\theta}_j)/N]^{1/2}, \text{ where}$$

$$\hat{\theta}_k(1-\hat{\theta}_k) = \min_j\{\hat{\theta}_j(1-\hat{\theta}_j)\}\},$$

and

$$\tilde{C}(X) = \{\theta: |\tilde{\delta}_j(X)-\theta_j| \leq z_{\alpha/2}(1-\tilde{c}(X))^{1/2}[\hat{\theta}_j(1-\hat{\theta}_j)/N]^{1/2}, j \neq k,$$

$$|\tilde{\delta}_k(X)-\theta_k| \leq z_{\alpha/2}(1-\tilde{c}(X))^{1/2} \sum_{j \neq k} [\hat{\theta}_j(1-\hat{\theta}_j)/N]^{1/2}, \text{ where}$$

$$\hat{\theta}_k(1-\hat{\theta}_k) = \min_j\{\hat{\theta}_j(1-\hat{\theta}_j)\}\}.$$

The robust Bayes regions C^* and $\tilde{C}$ will only differ significantly from the classical region C^0 in an area about the prior mean γ. In this prior region, it will be indicated that C^* and $\tilde{C}$ both obtain higher probabilities of coverage than C^0, and have rectangles of a smaller size than C^0. Outside of this prior region, both robust Bayes regions appear to possess probabilities of coverage and size approaching that of C^0.

4.3. Evaluation

In this section we briefly evaluate the goodness of C^* and $\tilde{C}$ as alternative confidence regions to C^0. As in the evaluation of the

Poisson confidence regions in Chapter 2, our criteria for evaluation are probability of coverage and volume of the confidence rectangle.

To evaluate the probabilities of coverage of the regions C^*, $\tilde{C}$ and C^0, we need to compare them against some nominal level. Typically a confidence level of $1-\alpha$ is assigned to each interval covering a component of θ. Then by using a Bonferroni inequality, the probability that the confidence rectangle

$$C_k^2 = \{\theta: \pi_j^L \le \theta_j \le \pi_j^U, j\neq k, 1-\sum_{j\neq k} \pi_j^U \le \theta_k \le 1-\sum_{j\neq k} \pi_j^L\}$$

covers θ is at least $1-(p-1)\alpha$. (Note that $\pi_j^L \le \theta_j \le \pi_j^U$, $j\neq k$ implies $1-\sum_{j\neq k} \pi_j^U \le \theta_k \le 1-\sum_{j\neq k} \pi_j^L$.) We are interested in obtaining a nominal level probability of coverage for a region of the form (4.6). Note that for any k,

$$C_k^0 = \{\theta: |\hat{\theta}_j-\theta_j| \le z_{\alpha/2}[\hat{\theta}_j(1-\hat{\theta}_j)/N]^{1/2}, j\neq k,$$
$$|\hat{\theta}_k-\theta_k| \le z_{\alpha/2} \sum_{j\neq k}[\hat{\theta}_j(1-\hat{\theta}_j)/N]^{1/2}\} \subset C^0,$$

$$C_k^* = \{\theta: |\delta_j^*(X)-\theta_j| \le z_{\alpha/2}(1-c^*(X))^{1/2}[\hat{\theta}_j(1-\hat{\theta}_j)/N]^{1/2}, j\neq k,$$
$$|\delta_k^*(X)-\theta_k| \le z_{\alpha/2}(1-c^*(X))^{1/2} \sum_{j\neq k}[\hat{\theta}_j(1-\hat{\theta}_j)/N]^{1/2}\} \subset C^*,$$

and

$$\tilde{C}_k = \{\theta: |\tilde{\delta}_j(X)-\theta_j| \le z_{\alpha/2}(1-\tilde{c}(X))^{1/2}[\hat{\theta}_j(1-\hat{\theta}_j)/N]^{1/2}, j\neq k,$$
$$|\tilde{\delta}_k(X)-\theta_k| \le z_{\alpha/2}(1-\tilde{c}(X))^{1/2} \sum_{j\neq k}[\hat{\theta}_j(1-\hat{\theta}_j)/N]^{1/2}\} \subset \tilde{C}.$$

The regions C_k^0, C_k^* and $\tilde{C}_k$, from above, have a probability of coverage of at least $1-(p-1)\alpha$. Therefore, the regions C^0, C^* and $\tilde{C}$ will also have a probability of coverage of at least $1-(p-1)\alpha$, and this value will be used as the nominal level.

Let us now compare the probabilities of coverage of C^*, $\tilde{C}$ and C^0 in the case where $p = 3$ and $N = 30$. In this example, $\alpha = .05$, and therefore the nominal level probability of coverage is $1-2(.05) = .90$, Figures 27, 28 and 29 show probabilities of coverage of the above regions plotted over the space of θ_1 and θ_2 values. The prior information used in the robust Bayes regions is $K = 15$, $\gamma_1 = \gamma_2 = .33$, and because of the symmetry of this example, probabilities of coverage are given only for $\theta_1 \geq \theta_2$. These probabilities of coverage are found through simulation, and the standard error of the values presented is approximately .006. Figure 27 shows that with only a few exceptions, the probabilities of coverage of C^0 are uniformly .90 or greater. The lowest probabilities of coverage of C^0 appear to occur in the extremities of the parameter space where one of the θ_i's is close to one. From looking at Figures 28 and 29, both robust Bayes regions appear to show higher probabilities of coverage than C^0 in a region about the prior mean (.33, .33). Note that C^* and $\tilde{C}$ can display smaller probabilities of coverage than C^0 outside of the prior region, but the probabilities of coverage appear to approach the nominal level as one moves farther away from the prior mean. In this example, C^* and $\tilde{C}$ appear to improve upon the classical region C^0 in much the

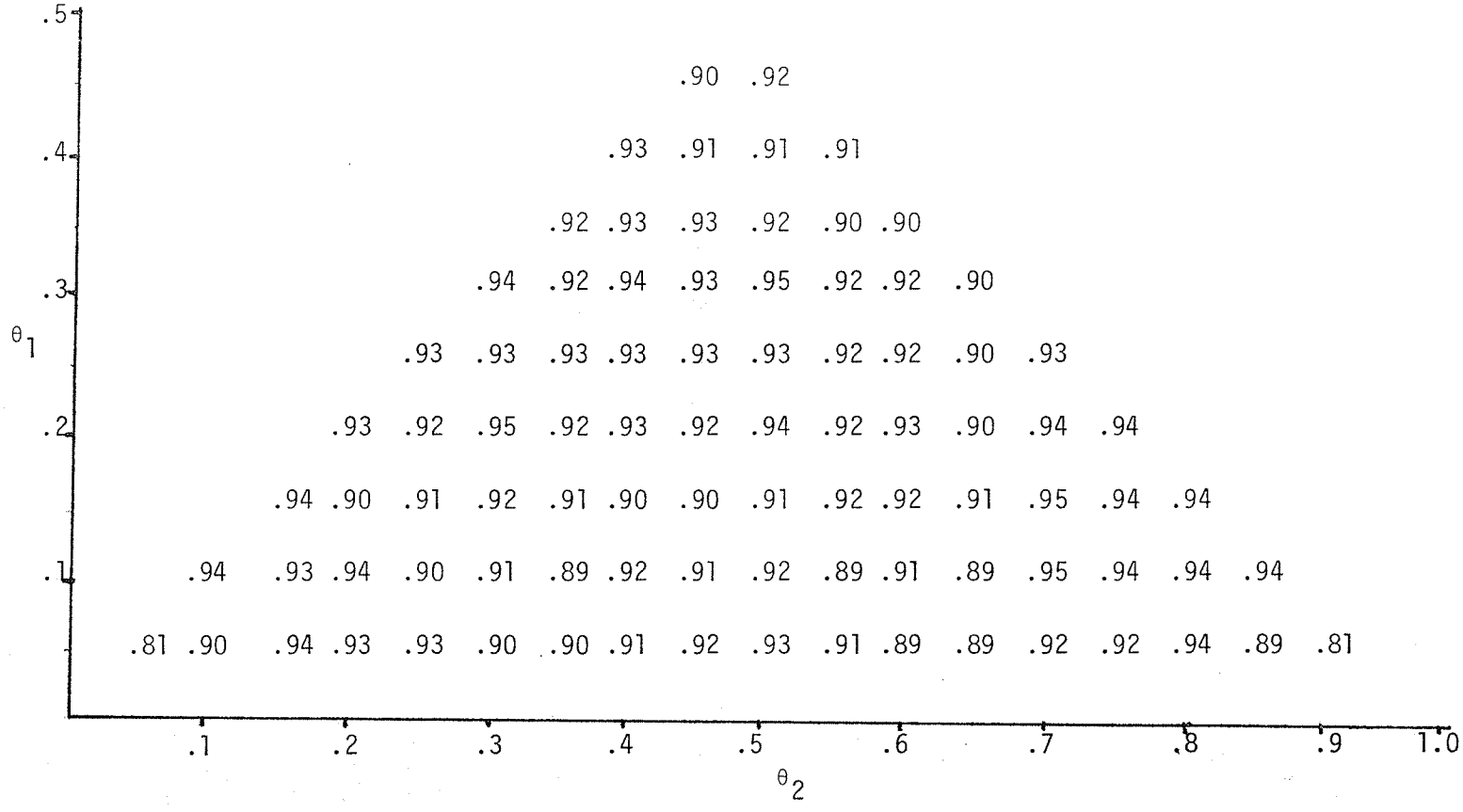


Figure 27

N=30, p=3. Probabilities of coverage of C^0.

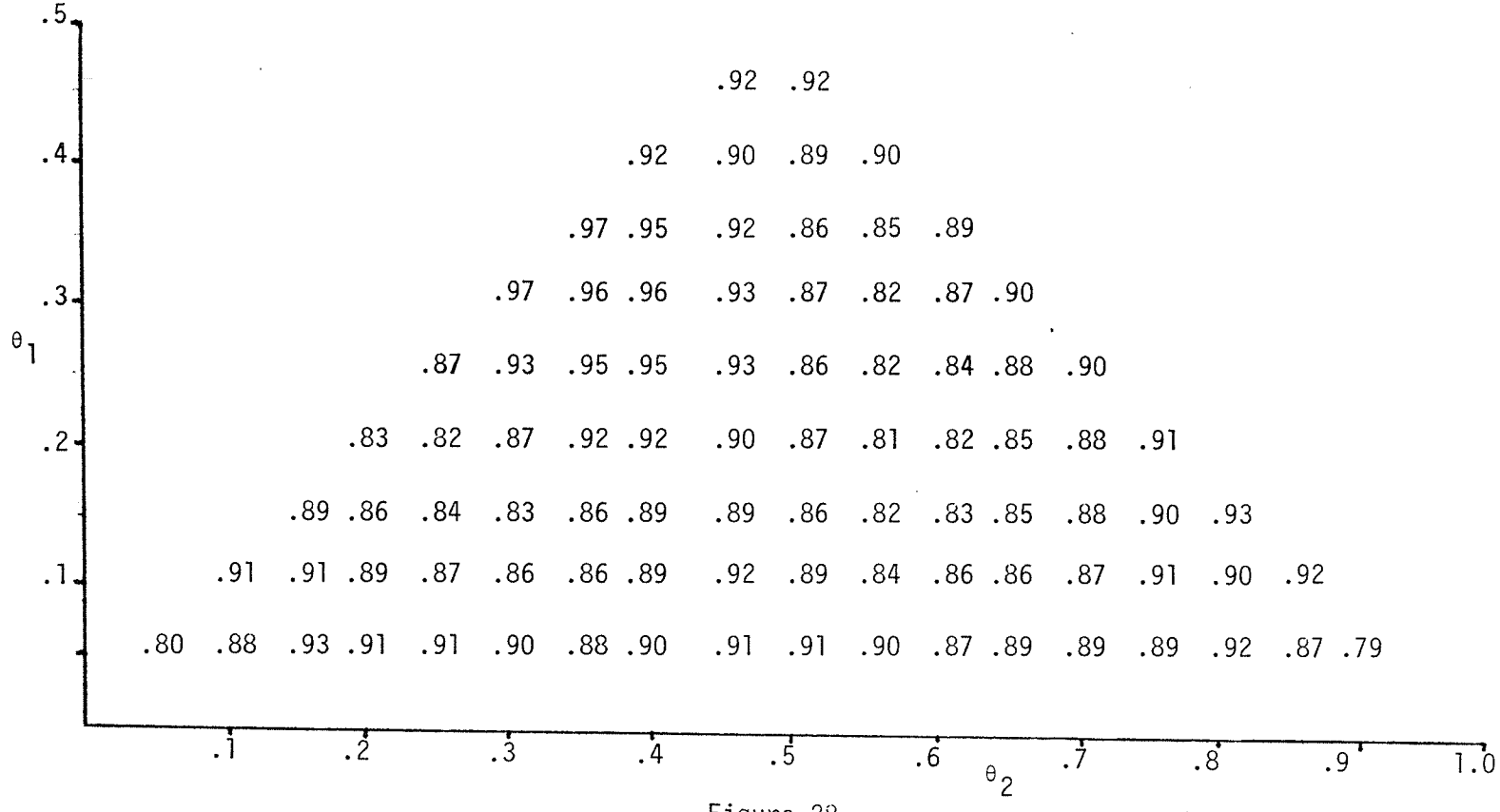


Figure 28

Prior information: γ=(.33,.33,.34), K=15. Probabilities of coverage of C*. N=30, p=3.

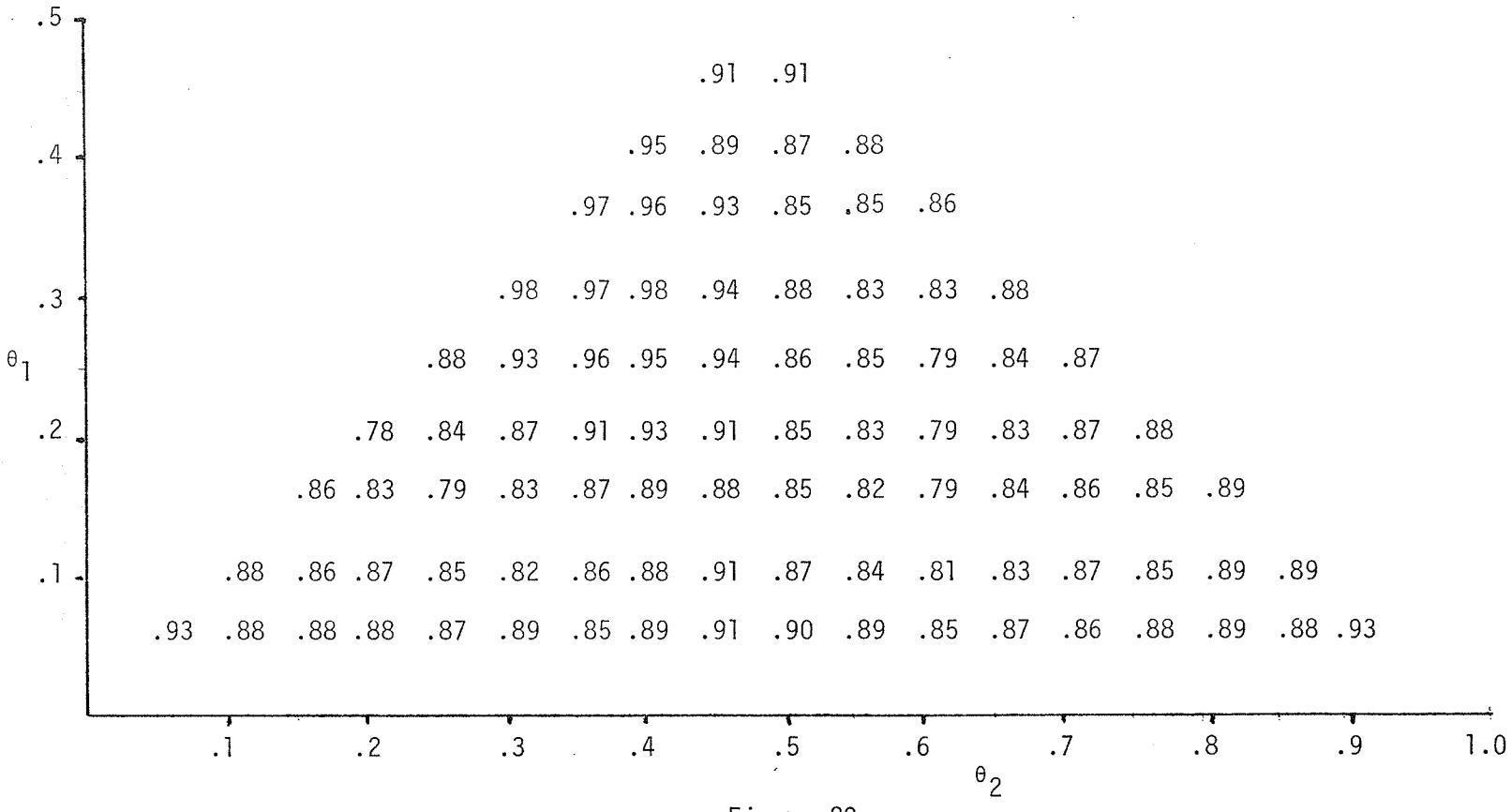


Figure 29

Prior information: γ=(.33,.33,.34), K=15. Probabilities of coverage of $\tilde{C}$. N=30, p=3.

same way as the robust Bayes region improves upon the classical region in the Poisson estimation problem.

Recall that, in Section 3, the robust Bayes estimator $\tilde{\delta}$ appears to perform better than δ^* in the prior region when there exist significant differences in the prior means $\gamma_1,\ldots,\gamma_p$. It is reasonable to expect a similar type of behavior for the corresponding confidence regions $\tilde{C}$ and C^*. More examples need to be considered where there exist differences between the prior means chosen.

The robust Bayes confidence regions C^* and $\tilde{C}$ will also have a smaller size than the classical region C^0 in the prior region. The volumes of the regions C^0, C^*, and $\tilde{C}$ are respectively

$$z_{\alpha/2}^p \prod_{j\neq k} [\hat{\theta}_j(1-\hat{\theta}_j)/N]^{1/2} \Big(\sum_{j\neq k} [\hat{\theta}_j(1-\hat{\theta}_j)/N]^{1/2} \Big),$$

$$(1-c^*(X))^{p/2} \prod_{j\neq k} [\hat{\theta}_j(1-\hat{\theta}_j)/N]^{1/2} \Big(\sum_{j\neq k} [\hat{\theta}_j(1-\hat{\theta}_j)/N]^{1/2} \Big),$$

and

$$(1-\tilde{c}(X))^{p/2} \prod_{j\neq k} [\hat{\theta}_j(1-\hat{\theta}_j)/N]^{1/2} \Big(\sum_{j\neq k} [\hat{\theta}_j(1-\hat{\theta}_j)/N]^{1/2} \Big).$$

The ratio of the volume of a robust Bayesian region to the volume of C^0 is

$$(1-c(X))^{p/2},$$

where $1-c(X)$ is the shrinking constant of the corresponding robust Bayes estimator. Now $0 \leq c(X) \leq K/(N+K)$, and when X is observed close to the prior mean, $c(X)$ will be significantly larger than zero. In this case both robust Bayes regions C^* and $\tilde{C}$ will have a

significantly smaller volume than C^0.

5. Using the estimator

In the robust Bayes estimators δ^* and $\tilde{\delta}$, two prior parameters, γ and K, are inputted. In this section, we discuss how to choose values for these parameters and then make some general comments concerning the use of the estimators.

It has already been mentioned that γ is the prior mean of θ, and K reflects the preciseness of the prior knowledge about θ. Specifically, under the Dirichlet prior,

$$\mathrm{Var}(\theta_i) = \gamma_i(1-\gamma_i)/(K+1), \quad i = 1,\ldots,p,$$

and

$$\mathrm{Cov}(\theta_i,\theta_j) = -\gamma_i\gamma_j/(K+1), \quad i \neq j.$$

Note that when p is moderately large and $\gamma_1,\ldots,\gamma_p$ are nearly equal, then the covariance terms are small relative to the variance terms. In this situation, the Dirichlet prior implies weak association between the unknown parameters. Thus one way to obtain values of γ and K is to assume $\theta_1,\ldots,\theta_p$ are not strongly related and make guesses at the prior mean and variance of each θ_i. Specifically, if $\tau_1^2,\ldots,\tau_p^2$ are the guesses for the prior variances of $\theta_1,\ldots,\theta_p$, then through the relations

$$\tau_i^2 = \frac{\gamma_i(1-\gamma_i)}{K_i}, \quad i = 1,\ldots,p,$$

one can obtain the values $K_1,\ldots,K_p$. If these values are not

too disparate, then K can be chosen to be some central value among $K_1,\ldots,K_p$. If great differences do exist among $K_1,\ldots,K_p$, then a central value K may not be appropriate to use. For example, let $p = 4$, and say $K_1 = 5$, $K_2 = 7$, $K_3 = 20$, and $K_4 = 22$. A single value of K would not reflect the prior variances well. One way to use more than one value of K in the estimator is to first separate $\{K_1,K_2,K_3,K_4\}$ into the similar groups $\{K_1,K_2\}$ and $\{K_3,K_4\}$. Then the parameters $\theta_1' = (\theta_1,\theta_2)$ and $\theta_2' = (\theta_3,\theta_4)$ could be estimated separately (by robust Bayes estimators) using the respective prior information $K_1' = 6$, and $K_2' = 21$. A disadvantage of the above method is that the resulting estimator would not necessarily satisfy the condition that $\sum_{i=1}^{p} \delta_i = 1$.

In the above discussion, we assumed that prior variances are known to the user. Unfortunately prior information usually consists of probabilities assigned to regions of the parameter space. Such prior information can also be used to find γ and K however. For example, assume $p = 2$ (one independent parameter) and that θ_1 is known to lie between a_1 and a_2 with probability .8. The middle section of the prior density of θ_1 may be well approximated by a beta density, and by using tables of beta fractiles, values of γ_1 and K can be obtained. If $p \geq 3$, then the above method can be used by first pretending $\theta_1,\ldots,\theta_p$ are independent, and then working with each θ_i separately to find values of γ_i and K_i. Then, as above, K is chosen to be some central value among $K_1,\ldots,K_p$, or the coordinates are grouped in some fashion.

It appears to be hard in general to specify values of the parameter K. One good way to understand the role of K in the estimators δ^* and $\tilde{\delta}$ is by means of an "equivalent sample size" argument, introduced by I. J. Good. To illustrate this argument, consider the conjugate Bayes estimator δ^B, which is defined componentwise as

$$\delta_i^B(X) = \frac{X_i + K\gamma_i}{N+K}, \quad i = 1,\ldots,p.$$

Here X_i is the count in cell i and N is the total count in the sample. If $K\gamma_i$ and K represent the count in cell i and total count in a preliminary sample (before the data is taken), then δ_i^B combines the information from the two samples to estimate θ_i. Thus a value of K reflects the amount of information that is known a priori about θ as measured in observations taken in a preliminary sample. The relative sizes of K and N will indicate how much the prior information will influence the posterior distribution and the Bayes estimator. For example, if N = 20, then K = 5 would mean that the prior information represents only five observations or about 1/4 of the information in the sample. If prior information can be expressed in terms of prior observations and related to the sample size N, then a value of K is easy to obtain.

In the estimators δ^* and $\tilde{\delta}$, the shrinking terms are truncated at $K/(N+K) = (1+N/K)^{-1}$. Thus the prior parameter K enters into the estimators only through the ratio N/K. If K is much larger than N, then the prior information is very strong, and the robust Bayes

estimators will shrink towards γ virtually as far as possible. On the other hand, if N/K is large, weak prior information has been inputted and the robust Bayes estimators will not shrink the MVUE $\hat{\theta}$ significantly.

In the Poisson robust Bayes estimator δ^* discussed in Chapter 2, prior parameters $\beta_1,\ldots,\beta_p$ are chosen which reflect the certainty of the prior information concerning $\lambda_1,\ldots,\lambda_p$. If $\beta_1 = \ldots = \beta_p = \beta^*$, then δ_i^* may be expressed as

$$\delta_i^*(X) = \mu_i + \left(1 - \min\left\{\frac{1}{\beta^*+1}, \frac{\sum_{j=1}^{p} X_j}{\sum_{j=1}^{p} X_j + \sum_{j=1}^{p} (X_j-\mu_j)^2}\right\}\right)(X_i-\mu_i), \quad i=1,\ldots,p.$$

Consider now the limiting situation where the multinomial distribution approaches the distribution of independent Poisson random variables. As in Theorem 4, let N and K simultaneously go to infinity such that $N/K \to D$, where D is a constant. By letting $N/K \to D$, one is inputting the same amount of prior information relative to N as $N \to \infty$. In this limiting situation, we saw in Theorem 4 that the truncation term of the multinomial estimator, $(1+N/K)^{-1}$, approaches $(1+D)^{-1}$, the truncation term of a Poisson estimator similar to δ^*. Thus the role of the quantity N/K in the multinomial estimator is similar to the role of β^* above in the Poisson estimator δ^*.

In general, robust Bayes estimators such as δ^* and $\tilde{\delta}$ are good alternatives to the MVUE when vague prior information is available. Both estimators will accept information concerning the central part

of the prior and they are robust with respect to uncertainty in the tail of the prior. It has already been noted that δ^0 has small risk at the extreme points of the parameter space. Thus, the robust Bayes estimators will show the greatest improvement upon δ^0 when γ has been chosen away from these extreme points. Finally, the proportion of risk improvement of $\tilde{\delta}$ near the prior mean appears to increase as p increases, and generally it appears that, as in the Poisson case, the robust Bayes estimators perform better in risk relative to the MVUE for larger values of p.

BIBLIOGRAPHY

Anscombe, F. J. (1963). Bayesian inference concerning many parameters with reference to supersaturated designs. Bull. Inst. Stat. Inst., 40, 721-733.

Berger, J. O. (1976). Admissible minimax estimation of a multivariate normal mean with arbitrary quadratic loss. Ann. Statist., 4, 223-226.

Berger, J. (1977a). A robust generalized Bayes estimator and confidence region for a multivariate normal mean. Mimeograph Series No. 480, Department of Statistics, Purdue University.

Berger, J. (1977b). Multivariate estimation with nonsymmetric loss functions. Mimeograph Series No. 517, Department of Statistics, Purdue University.

Berger, J. (1978). Improving on inadmissible estimators in continuous exponential families with applications to simultaneous estimation of gamma scale parameters. Mimeograph Series No. 78-6, Department of Statistics, Purdue University.

Berger, J. (1979). Statistical Decision Theory. Unpublished manuscript.

Berger, J., Bock, M. E., Brown, L. D., Casella, G. and Gleser, L. (1977). Minimax estimation of a normal mean vector for arbitrary quadratic loss and unknown covariance matrix. Ann. Statist., 5, 763-771.

Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1975). Discrete Multivariate Analysis. Cambridge, Mass.: MIT Press.

Bock, M. E. (1975). Minimax estimators of the mean of a multivariate normal distribution. Ann. Statist., 3, 209-218.

Brown, L. D. (1966). Admissibility of invariant estimators. Ann. Math. Statist., 37, 1087-1136.

Clevenson, M. L. and Zidek, J. V. (1975). Simultaneous estimation of the means of independent Poisson laws. J. Amer. Statist. Assoc., 70, 698-705.

Dawid, A. P. (1973). Posterior expectations for large observations. Biometrika, 60, 664-666.

Dickey, J. (1974). Bayesian alternatives to the F-test and least-squares estimate in the normal linear model. In: S. E. Fienberg and A. Zellner (eds.). Studies in Bayesian Econometrics and Statistics. Amsterdam: North Holland.

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

Efron, B. and Morris, C. (1973). Stein's estimation rule and its competitors - an empirical Bayes approach. J. Amer. Statist. Assoc., 68, 117-130.

Fienberg, S. E. and Holland, P. W. (1973). Simultaneous estimation of multinomial cell probabilities. J. Amer. Statist. Assoc., 68, 683-691.

Good, I. J. (1965). The Estimation of Probabilities; An Essay on Modern Bayesian Methods. Cambridge, Mass.: MIT Press.

Goodman, L. (1965). On simultaneous confidence intervals for multinomial proportions. Technometrics, 7, 247-254.

Hill, B. M. (1969). Foundations for the theory of least squares. J. R. Statist. Soc. B, 31, 89-97.

Hill, B. M. (1974). On coherence, inadmissibility, and inference about many parameters in the theory of least squares. In: S. E. Fienberg and A. Zellner (eds.). Studies in Bayesian Econometrics and Statistics. Amsterdam: North Holland.

Huber, P. J. (1977). Robust Statistical Procedures. Philadelphia: Society for Industrial and Applied Mathematics.

Hudson, H. M. (1974). Empirical Bayes estimation. Technical Report No. 58, Department of Statistics, Stanford University.

Hudson, H. M. (1978). A natural identity for exponential families with applications in multiparameter estimation. Ann. Statist., 6, 473-484.

Jackson, D. A., Donovan, T. M., Zimmer, W. J. and Deely, J. J. (1970). Γ_2-minimax estimators in the exponential family. Biometrika, 57, 439-443.

James, W. and Stein, C. (1961). Estimation with quadratic loss. Proc. Fourth Berkeley Symp. Math. Statist. Prob., 1, 361-379.

Johnson, B. M. (1971). On the admissible estimators for certain fixed sample binomial problems. Ann. Math. Statist., 42, 1579-1587.

Johnson, N. L. and Kotz, S. (1969). Discrete Distributions. Boston: Houghton Mifflin.

Johnson, N. L. and Kotz, S. (1970). Continuous Univariate Distributions. Boston: Houghton Mifflin.

Joshi, V. M. (1967). Inadmissibility of the usual confidence sets for the mean of a multivariate normal population. Ann. Math. Statist., 38, 1868-1875.

Leonard, T. (1972). Bayesian methods for discrete data. American College Testing Technical Bulletin No. 10.

Leonard, T. (1976). Some alternative approaches to multiparameter estimation. Biometrika, 63, 69-75.

Leonard, T. (1977). A Bayesian approach to some multinomial estimation and pretesting problems. J. Amer. Statist. Assoc., 72, 869-874.

Makabe and Morimura (1955). A normal approximation to the Poisson distribution. Reports on Statistical Application Research, JUSE, 4, 37-46.

Maritz, J. S. (1969). Empirical Bayes estimators for the Poisson distribution. Biometrika, 56, 349-359.

Novick, M. R., Lewis, C., and Jackson, P. H. (1973). Estimation of proportions in m groups. Psychometrika, 38, 19-46.

Peng, J. C. (1975). Simultaneous estimation of the parameters of independent Poisson distributions. Technical report No. 78, Department of Statistics, Stanford University.

Rao, C. R. (1973). Linear Statistical Inference and Its Applications. Second edition, John Wiley and Sons.

Robbins, H. (1955). An empirical Bayes approach to statistics. Proc. Third Berkeley Symp. Math. Statist. Prob., 1, 157-163.

Robbins, H. (1964). The empirical Bayes approach to statistical decision problems. Ann. Math. Statist., 35, 1-20.

Rubin, H. (1977). Robust Bayes estimation. In: S. Gupta and D. Moore (eds.). Statistical Decision Theory and Related Topics II. New York: Academic Press.

Stein, C. S. (1962). Confidence sets for the mean of a multivariate normal distribution. J. R. Statist. Soc. B, 24, 265-296.

Stein, C. (1974). Estimation of the parameters of a multivariate normal distribution, part I. Technical Report No. 63, Department of Statistics, Stanford University.

Tsui, K. W. (1978). Multiparameter estimation of discrete exponential distributions. Technical Report No. 37, University of California, Riverside.

Tsui, K. W. and Press, S. J. (1978a). Simultaneous estimation of several Poisson parameters under K-normalized squared error loss. Technical Report No. 38, Department of Statistics, University of California, Riverside.

Tsui, K. W. and Press, S. J. (1978b). Simultaneous Bayesian estimation of the parameters of independent Poisson distributions. Revision of Technical Report No. 33, Department of Statistics, University of California, Riverside.

APPENDIX

Proof of Theorem 1

It will become clear from the proof that, without loss of generality, we can set $\mu_1 = \ldots = \mu_p = 0$. Also it will be clear from the proof that we can ignore the truncation in the shrinking constant of δ^*. Then δ^* is defined by

$$\delta_i^*(X) = X_i - \frac{(X_i/(\beta_i+1)) \sum_{j=1}^{p} X_j/(\beta_j+1)}{\sum_{j=1}^{p} X_j/(\beta_j+1)^2 + \sum_{j=1}^{p} (X_j/(\beta_j+1))^2} \quad \text{if not all } X_i=0$$

$$= 0 \qquad \text{if } X_1=\ldots=X_p=0,$$

$$i = 1,\ldots,p.$$

Let

$$\phi_i(X) = \frac{(X_i/(\beta_i+1)) \sum_{j=1}^{p} X_j/(\beta_j+1)}{\sum_{j=1}^{p} X_j/(\beta_j+1)^2 + \sum_{j=1}^{p} (X_j/(\beta_j+1))^2}.$$

Clearly

$$\begin{aligned} I &= R(\delta^0,\lambda) - R(\delta^*,\lambda) \\ &= \sum_{i=1}^{p} E(X_i-\lambda_i)^2 - \sum_{i=1}^{p} E(\delta_i^*(X)-\lambda_i)^2 \\ &= \sum_{i=1}^{p} E(X_i-\lambda_i)^2 - \sum_{i=1}^{p} E(X_i-\phi_i(X)-\lambda_i)^2 \\ &= \sum_{i=1}^{p} E[2(X_i-\lambda_i)\phi_i(X)-(\phi_i(X))^2] \end{aligned}$$

$$= \sum_{\mathcal{X}} \sum_{i=1}^{p} [2(x_i - k_i\eta)\phi_i(x) - (\phi_i(x))^2] \prod_{j=1}^{p} \frac{e^{-k_j\eta}(k_j\eta)^{x_j}}{x_j!}$$

$$= I_V + I_{V^c},$$

where I_V is the sum over the region

$$V = \{x: |x_i - k_i\eta| > \eta^{9/16} \text{ for at least one } i\},$$

and I_{V^c} is the sum over V^c. Note that

$$|\phi_i(x)| = \frac{(x_i/(\beta_i+1)) \sum_{j=1}^{p} x_j/(\beta_j+1)}{\sum_{j=1}^{p} x_j/(\beta_j+1)^2 + \sum_{j=1}^{p} (x_j/(\beta_j+1))^2}$$

$$\leq \frac{p(x_i/(\beta_i+1))^2/2 + \sum_{j=1}^{p} (x_j/(\beta_j+1))^2/2}{\sum_{j=1}^{p} (x_j/(\beta_j+1))^2}$$

$$\leq K,$$

where K is some constant. Hence

$$I_V \leq \left| \sum_{V} \sum_{i=1}^{p} [2(x_i - k_i\eta)\phi_i(x) - (\phi_i(x))^2] \prod_{j=1}^{p} \frac{e^{-k_j\eta}(k_j\eta)^{x_j}}{x_j!} \right|$$

$$\leq K_1 \sum_{V} \sum_{i=1}^{p} |x_i - k_i\eta| \prod_{j=1}^{p} \frac{e^{-k_j\eta}(k_j\eta)^{x_j}}{x_j!} + K_2 \sum_{V} \prod_{j=1}^{p} \frac{e^{-k_j\eta}(k_j\eta)^{x_j}}{x_j!}$$

for constants K_1 and K_2. Now say $|x_m - k_m\eta| > \eta^{9/16}$. Then for large enough chosen integer ℓ,

$$I_V \le K_1 \sum_{\mathcal{X}} \sum_{i=1}^{p} |x_i - k_i\eta||x_m - k_m\eta|^{\ell} \eta^{-9\ell/16} \prod_{j=1}^{p} \frac{e^{-k_j\eta}(k_j\eta)^{x_j}}{x_j!}$$

(A.1)

$$+ K_2 \sum_{\mathcal{X}} |x_m - k_m\eta|^{\ell} \eta^{-9\ell/16} \prod_{j=1}^{p} \frac{e^{-k_j\eta}(k_j\eta)^{x_j}}{x_j!}$$

$$= o(1)$$

(since $E|X_m - k_m\eta|^c = O(\eta^{c/2})$ for nonnegative integer c).

Next consider I_{V^c}, and note that $V^c = \{x: |x_i - k_i\eta| \le \eta^{9/16}$ for all i$\}$. We first obtain a Taylor's expansion for $\phi_i(x)$ in V^c. Note that

$$T_1 = \frac{1}{\sum_{j=1}^{p} x_j/(\beta_j+1)^2 + \sum_{j=1}^{p} (x_j/(\beta_j+1))^2}$$

$$= \frac{1}{\sum_{j=1}^{p} (x_j/(\beta_j+1))^2 \left[1 + \frac{\sum_{j=1}^{p} x_j/(\beta_j+1)^2}{\sum_{j=1}^{p} (x_j/(\beta_j+1))^2}\right]},$$

and that

$$\frac{\sum_{j=1}^{p} x_j/(\beta_j+1)^2}{\sum_{j=1}^{p} (x_j/(\beta_j+1))^2} \le \frac{\sum_{j=1}^{p} (k_j\eta+\eta^{9/16})/(\beta_j+1)^2}{\sum_{j=1}^{p} ((k_j\eta-\eta^{9/16})/(\beta_j+1))^2}$$

$$= \frac{\sum_{j=1}^{p} (k_j+\eta^{-7/16})/(\beta_j+1)^2}{\eta \sum_{j=1}^{p} ((k_j-\eta^{-7/16})/(\beta_j+1))^2}$$

$$= O(\eta^{-1}).$$

Thus

$$(A.2) \quad T_1 = \frac{1}{\sum_{j=1}^{p} (x_j/(\beta_j+1))^2[1+O(\eta^{-1})]}$$

$$= \frac{1}{\sum_{j=1}^{p} (x_j/(\beta_j+1))^2} (1+O(\eta^{-1})).$$

Also

$$(A.3) \quad \frac{x_i/(\beta_i+1) \sum_{j=1}^{p} x_j/(\beta_j+1)}{\sum_{j=1}^{p} (x_j/(\beta_j+1))^2}$$

$$\leq \frac{(k_i\eta+\eta^{9/16})/(\beta_i+1) \sum_{j=1}^{p} (k_j\eta+\eta^{9/16})/(\beta_j+1)}{\sum_{j=1}^{p} ((k_j\eta-\eta^{9/16})/(\beta_j+1))^2}$$

$$= O(1).$$

Hence

$$\phi_i(x) = \frac{x_i/(\beta_i+1) \sum_{j=1}^{p} x_j/(\beta_j+1)}{\sum_{j=1}^{p} (x_j/(\beta_j+1))^2} [1+O(\eta^{-1})] \quad \text{by} \quad (A.2)$$

$$= \frac{x_i/(\beta_i+1)\sum_{j=1}^{p} x_j/(\beta_j+1)}{\sum_{j=1}^{p}(x_j/(\beta_j+1))^2} + O(n^{-1}) \quad \text{by} \quad (A.3).$$

Thus in the region V^c,

$$T_2 = \sum_{j=1}^{p}[2(x_i-k_i n)\phi_i(x)-\phi_i^2(x)]$$

$$= \sum_{i=1}^{p}(2(x_i-k_i n)\frac{x_i/(\beta_i+1)\sum_{j=1}^{p} x_j/(\beta_j+1)}{\sum_{j=1}^{p}(x_j/(\beta_j+1))^2} - [\frac{x_i/(\beta_i+1)\sum_{j=1}^{p} x_j/(\beta_j+1)}{\sum_{j=1}^{p}(x_j/(\beta_j+1))^2}]^2)$$

$$+ 2\sum_{i=1}^{p}(x_i-k_i n)O(n^{-1}) - 2\frac{[\sum_{j=1}^{p} x_j/(\beta_j+1)]^2}{\sum_{j=1}^{p}(x_j/(\beta_j+1))^2} O(n^{-1}) + O(n^{-2})$$

$$= \sum_{i=1}^{p}(2(x_i-k_i n)\frac{x_i/(\beta_i+1)\sum_{j=1}^{p} x_j/(\beta_j+1)}{\sum_{j=1}^{p}(x_j/(\beta_j+1))^2} - [\frac{x_i/(\beta_i+1)\sum_{j=1}^{p} x_j/(\beta_j+1)}{\sum_{j=1}^{p}(x_j/(\beta_j+1))^2}]^2)$$

$$+ o(1).$$

Let $v_i = x_i - k_i n$, $i = 1,\ldots,p$. Then

$$T_2 = \sum_{i=1}^{p}(\frac{2v_i(v_i+k_i n)/(\beta_i+1)(n\sum_{j=1}^{p} k_j/(\beta_j+1)+\sum_{j=1}^{p} v_j/(\beta_j+1))}{\sum_{j=1}^{p}((v_j+k_j n)/(\beta_j+1))^2}$$

$$- (\frac{(v_i+k_i n)/(\beta_i+1)(n\sum_{j=1}^{p} k_j/(\beta_j+1)+\sum_{j=1}^{p} v_j/(\beta_j+1)}{\sum_{j=1}^{p}((v_j+k_j n)/(\beta_j+1))^2})^2) + o(1).$$

In terms of the v_j, $V^c = \{v: |v_i| \leq \eta^{9/16} \text{ for all } i\}$, and in this region,

$$\frac{1}{\sum_{j=1}^{p}((v_j+k_j\eta)/(\beta_j+1))^2}$$

$$= \frac{1}{\eta^2\sum_{j=1}^{p}(k_j/(\beta_j+1))^2+2\eta\sum_{j=1}^{p}k_jv_j/(\beta_j+1)^2+\sum_{j=1}^{p}(v_j/(\beta_j+1))^2}$$

$$= \frac{1}{\eta^2\sum_{j=1}^{p}(k_j/(\beta_j+1))^2}\left[1+\frac{2\sum_{j=1}^{p}k_jv_j/(\beta_j+1)^2}{\eta\sum_{j=1}^{p}(k_j/(\beta_j+1))^2}+\frac{\sum_{j=1}^{p}(v_j/(\beta_j+1))^2}{\eta^2\sum_{j=1}^{p}(k_j/(\beta_j+1))^2}\right]^{-1}$$

$$= \frac{1}{\eta^2\sum_{j=1}^{p}(k_j/(\beta_j+1))^2}\left[1-\frac{2\sum_{j=1}^{p}k_jv_j/(\beta_j+1)^2}{\eta\sum_{j=1}^{p}(k_j/(\beta_j+1))^2}+O(\eta^{-14/16})\right]$$

and

$$\frac{1}{[\sum_{j=1}^{p}((v_j+k_j\eta)/(\beta_j+1))^2]^2} = \frac{1}{\eta^4[\sum_{j=1}^{p}(k_j/(\beta_j+1))^2]^2}[1+O(\eta^{-7/16})].$$

Therefore in the region V^c,

$$\sum_{i=1}^{p}\left(\frac{2v_i(v_i+k_i\eta)/(\beta_i+1)(\eta\sum_{j=1}^{p}k_j/(\beta_j+1)+\sum_{j=1}^{p}v_j/(\beta_j+1))}{\sum_{j=1}^{p}((v_j+k_j\eta)/(\beta_j+1))^2}\right)$$

$$= \frac{2}{n^2 \sum_{j=1}^{p} (k_j/(\beta_j+1))^2} [n^2 \sum_{j=1}^{p} k_j/(\beta_j+1) \sum_{j=1}^{p} k_j v_j/(\beta_j+1)$$

$$+ n \sum_{j=1}^{p} k_j/(\beta_j+1) \sum_{j=1}^{p} v_j^2/(\beta_j+1) + n \sum_{j=1}^{p} k_j v_j/(\beta_j+1) \sum_{j=1}^{p} v_j/(\beta_j+1)]$$

$$- \frac{4}{n[\sum_{j=1}^{p} (k_j/(\beta_j+1))^2]^2} \sum_{j=1}^{p} k_j/(\beta_j+1) \sum_{j=1}^{p} k_j v_j/(\beta_j+1) \sum_{j=1}^{p} k_j v_j/(\beta_j+1)^2 + o(1),$$

and

$$\sum_{i=1}^{p} \left(\frac{(v_i+k_i\eta)/(\beta_i+1)(n \sum_{j=1}^{p} k_j/(\beta_j+1) + \sum_{j=1}^{p} v_j/(\beta_j+1))}{\sum_{j=1}^{p} ((v_j+k_j\eta)/(\beta_j+1))^2} \right)^2$$

$$= \frac{(\sum_{j=1}^{p} k_j/(\beta_j+1))^2 \sum_{j=1}^{p} (k_j/(\beta_j+1))^2}{(\sum_{j=1}^{p} (k_j/(\beta_j+1))^2)^2} + o(1)$$

$$= \frac{(\sum_{j=1}^{p} k_j/(\beta_j+1))^2}{\sum_{j=1}^{p} (k_j/(\beta_j+1))^2} + o(1).$$

Thus

$$I_{V^c} = \sum_{V^c} \sum_{i=1}^{p} [2(x_i-k_i\eta)\phi_i(x) - (\phi_i(x))^2] \prod_{j=1}^{p} \frac{e^{-k_j\eta}(k_j\eta)^{x_j}}{x_j!}$$

(A.4)

$$= \sum_{V^c} \{ \frac{2}{\sum_{j=1}^{p} (k_j/(\beta_j+1))^2} [\sum_{j=1}^{p} k_j/(\beta_j+1) \sum_{j=1}^{p} k_j(x_j-k_j\eta)/(\beta_j+1)$$

$$+\frac{1}{n}\sum_{j=1}^{p}k_j/(\beta_j+1)\sum_{j=1}^{p}(x_j-k_j\eta)^2/(\beta_j+1)$$

$$+\frac{1}{n}\sum_{j=1}^{p}k_j(x_j-k_j\eta)/(\beta_j+1)\sum_{j=1}^{p}(x_j-k_j\eta)/(\beta_j+1)]$$

$$-\frac{4}{\eta[\sum_{j=1}^{p}(k_j/(\beta_j+1))^2]^2}\sum_{j=1}^{p}k_j/(\beta_j+1)\sum_{j=1}^{p}k_j(x_j-k_j\eta)/(\beta_j+1)\sum_{j=1}^{p}k_j(x_j-k_j\eta)/(\beta_j+1)^2$$

$$-\frac{(\sum_{j=1}^{p}k_j/(\beta_j+1))^2}{\sum_{j=1}^{p}(k_j/(\beta_j+1))^2}\}\prod_{j=1}^{p}\frac{e^{-k_j\eta}(k_j\eta)^{x_j}}{x_j!}+o(1).$$

Write $I_{V^c}=I_1-I_2$, where

$$I_1=\sum_{\mathcal{X}}\{\quad\}\prod_{j=1}^{p}\frac{e^{-k_j\eta}(k_j\eta)^{x_j}}{x_j!},$$

$$I_2=\sum_{V}\{\quad\}\prod_{j=1}^{p}\frac{e^{-k_j\eta}(k_j\eta)^{x_j}}{x_j!},$$

and $\{\quad\}$ denotes the expression in brackets in (A.4). Using Chebychev arguments as in (A.1), it can be shown that

(A.5) $$I_2=o(1).$$

Next, using the independence of $X_1,\ldots,X_p$ and the facts

$$E(X_j-k_j\eta)=0,$$

$$E(X_j-k_j\eta)^2=k_j\eta,\quad j=1,\ldots,p,$$

it is easy to see that

$$(A.6)\quad I_1 = \frac{2[(\sum_{j=1}^{p} k_j/(\beta_j+1))^2 + \sum_{j=1}^{p} (k_j/(\beta_j+1))^2]}{\sum_{j=1}^{p} (k_j/(\beta_j+1))^2}$$

$$- \frac{4 \sum_{j=1}^{p} k_j/(\beta_j+1) \sum_{j=1}^{p} (k_j/(\beta_j+1))^3}{[\sum_{j=1}^{p} (k_j/(\beta_j+1))^2]^2} - \frac{(\sum_{j=1}^{p} k_j/(\beta_j+1))^2}{\sum_{j=1}^{p} (k_j/(\beta_j+1))^2}$$

$$= \frac{(\sum_{j=1}^{p} k_j/(\beta_j+1))^2}{\sum_{j=1}^{p} (k_j/(\beta_j+1))^2} - \frac{4 \sum_{j=1}^{p} k_j/(\beta_j+1) \sum_{j=1}^{p} (k_j/(\beta_j+1))^3}{[\sum_{j=1}^{p} (k_j/(\beta_j+1))^2]^2} + 2.$$

Combining (A.1), (A.5) and (A.6) gives the desired result.

Proof of Theorem 2

Let

$$\phi_i(X) = \frac{(X_i-\mu_i) \sum_{j=1}^{p} X_j}{\sum_{j=1}^{p} X_j + \sum_{j=1}^{p} (X_j-\mu_j)^2}, \quad i = 1,\ldots,p.$$

Note that

$$R(\delta^0,\lambda) = \sum_{i=1}^{p} \lambda_i = \eta \sum_{i=1}^{p} k_i = \eta.$$

Thus

$$(A.7)\quad I = \frac{R(\delta^0,\lambda)-R(\delta^*,\lambda)}{R(\delta^0,\lambda)}$$

$$= \eta^{-1} \sum_{i=1}^{p} E[2(X_i - k_i\eta)\phi_i(X) - (\phi_i(X))^2]$$

$$= \sum_{\mathcal{X}} \sum_{i=1}^{p} [2(x_i\eta^{-1/2} - k_i\eta^{1/2})\phi_i(x)\eta^{-1/2} - (\phi_i(x)\eta^{-1/2})^2] \prod_{j=1}^{p} \frac{e^{-k_j\eta}(k_j\eta)^{x_j}}{x_j!}.$$

First, since we have assumed that

$$\lim_{n\to\infty} \frac{\theta_i}{\eta^{1/2}} = \theta_i^* = O(1) \quad \text{for all } i,$$

there exist constants K_1 and $N > 1$ such that

$$\frac{|\theta_i|}{\eta^{1/2}} < K_1 \quad \text{for all } i \text{ when } \eta \geq N.$$

Define the constant K by $K = 3K_1$ and write

$$I = I_V + I_{V^c},$$

where I_V is the sum over the region

$$V = \{x: |x_i - k_i\eta| > K\eta^{9/16} \text{ for at least one } i\}$$

and I_{V^c} is the sum over V^c.

Consider the case where $x \in V$. Note that

$$(A.8) \qquad |\phi_i(x)\eta^{-1/2}| = \left| \frac{(x_i - \mu_i)\sum_{j=1}^{p} x_j}{(\sum_{j=1}^{p} x_j + \sum_{j=1}^{p} (x_j - \mu_j)^2)\eta^{1/2}} \right|$$

$$= \left| \frac{(x_i-\mu_i)\sum_{j=1}^{p}(x_j-\mu_j)+\eta(x_i-\mu_i)-(x_i-\mu_i)\sum_{j=1}^{p}\theta_j}{(\sum_{j=1}^{p}(x_j-\mu_j)^2+\sum_{j=1}^{p}x_j)\eta^{1/2}} \right|$$

$$\leq \left| \frac{(x_i-\mu_i)\sum_{j=1}^{p}(x_j-\mu_j)}{\eta^{1/2}\sum_{j=1}^{p}(x_j-\mu_j)^2} \right| + \left| \frac{\eta^{1/2}(x_i-\mu_i)}{\sum_{j=1}^{p}(x_j-\mu_j)^2} \right| + \left| \frac{(x_i-\mu_i)\sum_{j=1}^{p}\theta_j}{\eta^{1/2}\sum_{j=1}^{p}(x_j-\mu_j)^2} \right|.$$

Now for constants K_2, K_3, K_4, and K_5, the following facts hold.

(A.9) $$\left| \frac{(x_i-\mu_i)\sum_{j=1}^{p}(x_j-\mu_j)}{\eta^{1/2}\sum_{j=1}^{p}(x_j-\mu_j)^2} \right| \leq \frac{p(x_i-\mu_i)^2/2+\sum_{j=1}^{p}(x_j-\mu_j)^2/2}{\eta^{1/2}\sum_{j=1}^{p}(x_j-\mu_j)^2}$$

$$\leq K_2.$$

If $x \in V$ and $|x_i-k_i\eta| > K\eta^{9/16}$, then

(A.10) $$\left| \frac{\eta^{1/2}(x_i-\mu_i)}{\sum_{j=1}^{p}(x_j-\mu_j)^2} \right| \leq \frac{\eta^{1/2}}{|x_i-\mu_i|}$$

$$\leq \frac{\eta^{1/2}}{K(\eta^{9/16}-\eta^{1/2}/3)}$$

$$= \frac{1}{K(\eta^{1/16}-1/3)} \leq \frac{3}{2K} = K_3.$$

If $x \in V$ and $|x_i-k_i\eta| \leq K\eta^{9/16}$, then there exists $m \in \{1,2,\ldots,p\}$ such that $|x_m-k_m\eta| > K\eta^{9/16}$ and

$$\text{(A.11)}\qquad \left|\frac{n^{1/2}(x_i-\mu_i)}{\sum_{j=1}^{p}(x_j-\mu_j)^2}\right| \le \frac{Kn^{17/16}}{(x_m-\mu_m)^2}$$

$$\le \frac{Kn^{17/16}}{K^2(n^{9/16}-n^{1/2}/3)^2}$$

$$= \frac{1}{K(n^{1/16}-2/3+n^{-1/16}/4)}$$

$$\le \frac{1}{K(n^{1/16}-2/3)} \le K_4.$$

$$\text{(A.12)}\qquad \left|\frac{(x_i-\mu_i)\sum_{j=1}^{p}\theta_j}{n^{1/2}\sum_{j=1}^{p}(x_j-\mu_j)^2}\right| \le K/3\left|\frac{(x_i-\mu_i)}{\sum_{j=1}^{p}(x_j-\mu_j)^2}\right|$$

$$\le K/3\left|\frac{n^{1/2}(x_i-\mu_i)}{\sum_{j=1}^{p}(x_j-\mu_j)^2}\right| \le K_5.$$

Therefore, using (A.8)-(A.12), for some constant K_6 and $x \in V$,

$$|\phi_i(x)n^{-1/2}| \le K_6 \quad \text{for all } i.$$

This implies that

(A.13)

$$I_V \le \Big|\sum_{V}\sum_{i=1}^{p}[2(x_i n^{-1/2}-k_i n^{1/2})\phi_i(x)n^{-1/2}-(\phi_i(x)n^{-1/2})^2]\prod_{j=1}^{p}\frac{e^{-k_j n}(k_j n)^{x_j}}{x_j!}\Big|$$

$$\le \sum_{\mathcal{X}}\sum_{i=1}^{p}[2K_6|x_i n^{-1/2}-k_i n^{1/2}|+(K_6)^2]|x_m-k_m n|^{\ell}K^{-\ell}n^{-9/16\ell}$$

$$\cdot\sum_{j=1}^{p}\frac{e^{-k_j n}(k_j n)^{x_j}}{x_j!}$$

$$= o(1)$$

for large enough chosen ℓ(since $E[|X_i-k_i n|^{\ell}] = O(n^{-\ell/2})$).

Next consider the sum I_{V^c}, and note that $V^c = \{x: |x_i-k_i n| \leq K n^{9/16}$ for all i$\}$. Expanding $\phi_i(x)n^{-1/2}$ in a Taylor's series for $x \in V^c$ gives

$$\phi_i(x)n^{-1/2} = \frac{(x_i-\mu_i)\sum_{j=1}^{p} x_j}{n^{1/2}(\sum_{j=1}^{p} x_j + \sum_{j=1}^{p}(x_j-\mu_j)^2)}$$

$$= \frac{(x_i-k_i n+\theta_i)\sum_{j=1}^{p} x_j}{n^{1/2}(\sum_{j=1}^{p} x_j + \sum_{j=1}^{p}(x_j-k_j n+\theta_j)^2)}$$

$$= \frac{(x_i-k_i n+\theta_i)n^{-1/2}\sum_{j=1}^{p}[(x_j-k_j n)n^{-1/2}+k_j n^{1/2}]}{\sum_{j=1}^{p}[(x_j-k_j n)n^{-1/2}+k_j n^{1/2}]+n^{1/2}\sum_{j=1}^{p}[(x_j-k_j n+\theta_j)n^{-1/2}]^2}$$

$$= (x_i-k_i n+\theta_i)n^{-1/2}[1 + \frac{\sum_{j=1}^{p}[(x_j-k_j n+\theta_j)n^{-1/2}]^2}{1+n^{-1/2}\sum_{j=1}^{p}(x_j-k_j n)n^{-1/2}}]^{-1}.$$

Next

$$\text{(A.14)}\quad \frac{\sum_{j=1}^{p}[(x_j-k_j n+\theta_j)n^{-1/2}]^2}{1+n^{-1/2}\sum_{j=1}^{p}(x_j-k_j n)n^{-1/2}} = \sum_{j=1}^{p}[(x_j-k_j n+\theta_j)n^{-1/2}]^2[1+O(n^{-7/16})].$$

Then using (A.14),

$$\text{(A.15)}\quad \phi_i(x)n^{-1/2} = \frac{(x_i-k_i n+\theta_i)n^{-1/2}}{1+\sum_{j=1}^{p}[(x_j-k_j n+\theta_j)n^{-1/2}]^2[1+O(n^{-7/16})]}$$

$$= \frac{(x_i - k_i\eta + \theta_i)\eta^{-1/2}}{1 + \sum_{j=1}^{p}[(x_j - k_j\eta + \theta_j)\eta^{-1/2}]^2}\,(1+O(\eta^{-7/16}))$$

$$= \frac{(x_i - k_i\eta + \theta_i)\eta^{-1/2}}{1 + \sum_{j=1}^{p}[(x_j - k_j\eta + \theta_j)\eta^{-1/2}]^2} + O(\eta^{-7/16}).$$

Let

$$\phi_i^*(x) = \frac{(x_i - k_i\eta + \theta_i)\eta^{-1/2}}{1 + \sum_{j=1}^{p}[(x_j - k_j\eta + \theta_j)\eta^{-1/2}]^2}, \quad i = 1,\ldots,p.$$

Combining (A.7), (A.13), and (A.15), we have

$$\text{(A.16)} \quad I = \sum_{V^c} \sum_{i=1}^{p}[2(x_i\eta^{-1/2} - k_i\eta^{1/2})\phi_i^*(x) - (\phi_i^*(x))^2] \prod_{j=1}^{p} \frac{e^{-k_j\eta}(k_j\eta)^{x_j}}{x_j!} + o(1).$$

Now we apply the well known fact (see Makabe and Morimura (1955)) that if $X \sim$ Poisson (λ) with density $p_X(\lambda)$, $y = \lambda^{-1/2}(x-\lambda)$, and $f(y)$ is the density of a standard normal random variable, then when $\lambda \geq 1$,

$$\text{(A.17)} \quad p_X(\lambda) = f(y)[\lambda^{-1/2} + \lambda^{-1}(y/2 - y^3/6)] + R,$$

where $|R| = L\lambda^{-3/2} + o(\lambda^{-3/2})$ and L is a constant. Using this in (A.16) gives

$$I = \sum_{V^c} \sum_{i=1}^{p}[2(x_i\eta^{-1/2} - k_i\eta^{1/2})\phi_i^*(x) - (\phi_i^*(x))^2]$$

$$\cdot \prod_{j=1}^{p}((k_j\eta)^{-1/2} f[(k_j\eta)^{-1/2}(x_j - k_j\eta)]\{1 + (k_j\eta)^{-1/2}[(x_j - k_j\eta)(k_j\eta)^{-1/2}/2$$

$$- (x_j-k_j\eta)^3(k_j\eta)^{-3/2}/6) + R_j\}) + o(1),$$

where $|R_j| = O(\eta^{-3/2})$, $j = 1,\ldots,p$. It is easy to see that I can be written as $I = I_1+I_2 + o(1)$, where

$$I_1 = \sum_{V^c} \sum_{i=1}^{p} [2(x_i\eta^{-1/2}-k_i\eta^{1/2})\phi_i^*(x) - (\phi_i^*(x))^2] \prod_{j=1}^{p} (k_j\eta)^{-1/2} f[(k_j\eta)^{-1/2}(x_j-k_j\eta)]$$

and

$$I_2 = \sum_{V^c} \sum_{i=1}^{p} [2(x_i\eta^{-1/2}-k_i\eta^{1/2})\phi_i^*(x)-(\phi_i^*(x))^2] \sum_{\ell=1}^{3^p-1} \prod_{j=1}^{p} t_{j\ell},$$

and in each product $\prod_{j=1}^{p} t_{j\ell}$, $t_{j\ell}$ is either

$$S_j = (k_j\eta)^{-1/2} f[(k_j\eta)^{-1/2}(x_j-k_j\eta)],$$

$$T_j = f[(k_j\eta)^{-1/2}(x_j-k_j\eta)](k_j\eta)^{-1}[(x_j-k_j\eta)(k_j\eta)^{-1/2}/2-(x_j-k_j\eta)^3(k_j\eta)^{-3/2}/6],$$

or R_j, with at least one of the last two types of terms occurring in each product.

First we show that $|I_2| = o(1)$. Note that for $x \in V^c$,

$$\left| \sum_{i=1}^{p} [2(x_i\eta^{-1/2}-k_i\eta^{1/2})\phi_i^*(x)-(\phi_i^*(x))^2] \right| < K_7\eta^{1/16},$$

for some constant K_7. Thus

$$|I_2| \le K_7\eta^{1/16} \left| \sum_{V^c} \sum_{\ell=1}^{3^p-1} \prod_{j=1}^{p} t_{j\ell} \right|$$

$$\leq K_7 \eta^{1/16} \sum_{\ell=1}^{3^p-1} \sum_{V^c} \Big| \prod_{j=1}^{p} t_{j\ell} \Big|.$$

Now

$$\sum_{V^c} \Big| \prod_{j=1}^{p} t_{j\ell} \Big| = \sum_{V^c} \Big| \prod_{C_1} S_j \prod_{C_2} T_j \prod_{C_3} R_j \Big|,$$

where either C_2 or $C_3 \neq \phi$, and $\{C_1, C_2, C_3\}$ is a partition of $\{1,2,\ldots,p\}$. Then

$$\sum_{V^c} \Big| \prod_{j=1}^{p} t_{j\ell} \Big| \leq \sum_{V^c} \prod_{C_1} |S_j| \prod_{C_2} |T_j| \prod_{C_3} |R_j|.$$

Let $A_j = \{x_j: |x_j - k_j\eta| \leq K\eta^{9/16}\}$, $j = 1,\ldots,p$. Note that

$$\text{(A.18)} \quad \sum_{A_j} |R_j| \leq K\eta^{9/16}[\eta^{-3/2} + o(\eta^{-3/2})]$$

$$= K\eta^{-15/16} + o(\eta^{-15/16}),$$

and

$$\sum_{A_j} |S_j| \leq (k_j\eta)^{-1/2} \sum_{B_j} f[(k_j\eta)^{-1/2}(x_j - k_j\eta)],$$

where $B_j = \{x_j: x_j = 0, \pm 1, \pm 2, \ldots\}$. Let $y_j = (k_j\eta)^{-1/2}(x_j - k_j\eta)$ and $C_j = \{y_j: y_j = 0, \pm(k_j\eta)^{-1/2}, \ldots\}$, so that

$$\sum_{A_j} |S_j| \leq (k_j\eta)^{-1/2} \sum_{C_j} f(y_j).$$

Now let $\eta \to \infty$. Since $f(y_j)$ is a bounded continuous function of y_j,

$$\lim_{\eta\to\infty} (k_j\eta)^{-1/2} \sum_{C_j} f(y_j) = \int_{-\infty}^{\infty} f(y_j)dy_j = 1.$$

Therefore

$$\text{(A.19)} \qquad \sum_{A_j} |S_j| = O(1).$$

Next

$$\sum_{A_j} |T_j| \le \sum_{A_j} f[(k_j\eta)^{-1/2}(x_j-k_j\eta)](k_j\eta)^{-1}[|x_j-k_j\eta|(k_j\eta)^{-1/2}/2 + |x_j-k_j\eta|^3(k_j\eta)^{-3/2}/6]$$

$$\le \eta^{-1}k_j^{-1/2} \sum_{C_j} f(y_j)k_j^{-1/2}[|y_j|/2 + |y_j|^3/6]$$

Note that $f(y_j)k_j^{-1/2}[|y_j|/2 + |y_j|^3/6]$ is a bounded continuous function of y_j. Therefore

$$\lim_{\eta\to\infty} (k_j\eta)^{-1/2} \sum_{C_j} f(y_j)k_j^{-1/2}[|y_j|/2 + |y_j|^3/6]$$

$$= \int_{-\infty}^{\infty} f(y_j)k_j^{-1/2}[|y_j|/2 + |y_j|^3/6]dy_j$$

$$\text{(A.20)} \qquad = O(1).$$

Hence using (A.20),

$$\text{(A.21)} \qquad \sum_{A_j} |T_j| = O(\eta^{-1/2}).$$

From (A.18), (A.19), (A.21), and the fact that for each ℓ, $\prod_{j=1}^{p} t_{j\ell}$ contains at least one of T_j or R_j,

$$\sum_{V^c} |\prod_{j=1}^{p} t_{j\ell}| = O(\eta^{-1/2}), \quad \ell = 1,\ldots,3^p-1,$$

and

$$\text{(A.23)} \qquad |I_2| = o(1).$$

It remains only to analyze I_1.

Define $w_i = n^{-1/2}(x_i - k_i n + \theta_i)$, $i = 1,\ldots,p$, and let $\tilde{\theta}_i = n^{-1/2}\theta_i$, $i = 1,\ldots,p$. Then

$$(A.23) \qquad I_1 = n^{p/2} \sum_{V^*} \{ \sum_{i=1}^{p} [2(w_i - \tilde{\theta}_i) \frac{w_i}{1+\sum_{j=1}^{p} w_j^2} - (\frac{w_i}{1+\sum_{j=1}^{p} w_j^2})^2]$$

$$\cdot \prod_{j=1}^{p} (2\pi k_j)^{-1/2} \exp\{[(w_j - \tilde{\theta}_j)k_j^{-1/2}]^2/2\}\},$$

where

$$V^* = \{w: w_i = -k_i n^{1/2+\tilde{\theta}_i},\ -k_i n^{1/2+\tilde{\theta}_i+n^{-1/2}},\ \ldots \text{ and } |w_i - \tilde{\theta}_i| \le K n^{1/16}\}.$$

Write $I_1 = I_3 - I_4$, where

$$I_3 = n^{-p/2} \sum_{A} \{\quad\},$$

$$I_4 = n^{-p/2} \sum_{A \sim V^*} \{\quad\},$$

$\{\quad\}$ is the quantity in brackets in (A.23), and

$$A = \{w: w_i = 0, \pm n^{-1/2}, \pm 2n^{-1/2}, \ldots\}$$

$$A \sim V^* = \{w: w_i = 0, \pm n^{-1/2}, \pm 2n^{-1/2}, \ldots\}$$

$$\cap \{w: w_i = -k_i n^{-1/2+\tilde{\theta}_i}, \ldots \text{ for all } i, \text{ and } |w_j - \tilde{\theta}_j| > K n^{1/16} \text{ for at least one } j\}.$$

Now

$$A \sim V^* \subset V' = \{w: |w_i - \tilde{\theta}_i| > K n^{1/16} \text{ for at least one } i\}.$$

For $w \in V'$, say $|w_m - \tilde{\theta}_m| > K n^{1/16}$. Then

$$|I_4| \le \eta^{1/2} \sum_{V'} \sum_{i=1}^{p} [2|w_i-\tilde{\theta}_i|+1]|w_m-\tilde{\theta}_m|^{\ell} K^{-\ell}\eta^{-\ell/16}$$

(A.24)

$$\cdot \prod_{j=1}^{p} (2\pi k_j)^{-1/2} \exp\{[(w_j-\tilde{\theta}_j)k_j^{-1/2}]^2/2\}$$

$$\le K^{-\ell}\eta^{-\ell/16}(\eta^{-p/2} \sum_{A} \sum_{i=1}^{p} [2|w_i-\tilde{\theta}_i|+1]|w_m-\tilde{\theta}_m|^{\ell}$$

$$\cdot \prod_{j=1}^{p} (2\pi k_j)^{-1/2} \exp\{[(w_j-\tilde{\theta}_j)k_j^{-1/2}]^2/2\}).$$

Let $\eta \to \infty$ and note that

$$\sum_{i=1}^{p} [2|w_i-\tilde{\theta}_i|+1]|w_m-\tilde{\theta}_m|^{\ell} \prod_{j=1}^{p} (2\pi k_j)^{-1/2} \exp\{[(w_j-\tilde{\theta}_j)k_j^{-1/2}]^2/2\}$$

is a continuous bounded function of $w_1,\ldots,w_p$. Therefore the right factor in (A.24) converges to

$$\int_{\mathbb{R}^p} \sum_{i=1}^{p} [2|w_i-\theta_i^*|+1]|w_m-\theta_m^*|^{\ell} \prod_{j=1}^{p} (2\pi k_j)^{-1/2} \exp\{[(w_j-\theta_j^*)k_j^{-1/2}]^2/2\} dw_j$$

$$= O(1),$$

since all absolute central moments of W_i, $i = 1,\ldots,p$, are finite. Therefore

$$I_4 = o(1). \tag{A.25}$$

Finally,

$$\lim_{\eta\to\infty} I_3 = \lim_{\eta\to\infty} \eta^{-p/2} \sum_{A} \sum_{i=1}^{p} [2(w_i-\tilde{\theta}_i) \frac{w_i}{1+\sum_{j=1}^{p} w_j^2} - (\frac{w_i}{1+\sum_{j=1}^{p} w_j^2})^2] \tag{A.26}$$

$$\cdot \prod_{j=1}^{p} (2\pi k_j)^{-1/2} \exp\{[(w_j-\tilde{\theta}_j)k_j^{-1/2}]^2/2\}$$

$$= \int_{\mathbb{R}^p} \sum_{i=1}^{p} \left[2(w_i-\theta_i^*) \frac{w_i}{1+\sum_{j=1}^{p} w_j^2} - \left(\frac{w_i}{1+\sum_{j=1}^{p} w_j^2}\right)^2\right] \prod_{j=1}^{p} (2\pi k_j)^{-1/2} \exp\{[(w_j-\theta_j^*)k_j^{-1/2}]^2/2\} dw_j .$$

Combining (A.22), (A.25) and (A.26) gives the desired result.

Proof of Theorem 3

Let

$$V^c = \{x: \ |x_i - k_i\eta| \leq \eta^{9/16} \ \text{ for all } i\}.$$

Using arguments as in the proofs of Theorems 1 and 2, it can be shown that for $x \in V^c$,

$$c_i^*(x) = \frac{1}{\beta_i+1} \min\{1, \frac{\sum_{j=1}^{p} x_j/(\beta_j+1)}{\sum_{j=1}^{p} x_j/(\beta_j+1)^2 + \sum_{j=1}^{p} ((x_j-\mu_j)/(\beta_j+1))^2}\}$$

$$= \frac{1}{\beta_i+1} \frac{\sum_{j=1}^{p} k_j/(\beta_j+1)}{\eta \sum_{j=1}^{p} (k_j/(\beta_j+1))^2} + O(\eta^{-3/2})$$

$$= O(\eta^{-1}).$$

Thus

$$\delta_i^*(X) = x_i - \frac{(x_i-\mu_i)}{\beta_i+1} \min\{1, \frac{\sum_{j=1}^{p} x_j/(\beta_j+1)}{\sum_{j=1}^{p} x_j/(\beta_j+1)^2 + \sum_{j=1}^{p} ((x_j-\mu_j)/(\beta_j+1))^2}\}$$

$$= x_i - \frac{k_i/(\beta_i+1) \sum_{j=1}^{p} k_j/(\beta_j+1)}{\sum_{j=1}^{p} (k_j/(\beta_j+1))^2} + O(\eta^{-3/2}).$$

Let

$$M_i = \frac{k_i/(\beta_i+1) \sum_{j=1}^{p} k_j/(\beta_j+1)}{\sum_{j=1}^{p} (k_j/(\beta_j+1))^2}.$$

Then for $x \in V^c$,

(A.27) $$(\delta_i^*(x) + (1-c_i^*(x))(z_{\gamma/2}^2-1)/3-\lambda_i)^2$$

$$= (x_i-M_i+(z_{\gamma/2}-1)/3-k_i\eta+t_1)^2, \text{ where } t_1 = O(\eta^{-1/2})$$

$$= (x_i-k_i\eta)^2+2(x_i-k_i\eta)[(z_{\gamma/2}-1)/3-M_i+t_1]$$

$$+ [(z_{\gamma/2}^2-1)/3-M_i+t_1]^2$$

$$= (x_i-k_i\eta)^2+2(x_i-k_i\eta)t_2+t_3,$$

where $t_2 = O(1)$, $t_3 = O(1)$. Also

$$(A.28)\qquad z^2_{\gamma/2}(1-c_i^*(x))(x_i-z^2_{\gamma/2}/4) = z^2_{\gamma/2}k_i\eta+z^2_{\gamma/2}(x_i-k_i\eta)+t_4,$$

where $t_4 = O(1)$. Let $\Omega = \{x: \lambda \in C^*(x)\}$. Then from (A.27) and (A.28),

$$\Omega \cap V^c = \{x: (\delta_i^*(x)+(1-c_i^*(x))(z^2_{\gamma/2}-1)/3-\lambda_i)^2$$

$$< z^2_{\gamma/2}(1-c_i^*(x))(x_i-z^2_{\gamma/2}/4) \text{ and } |x_i-k_i\eta|\leq\eta^{9/16} \text{ for all } i\}$$

$$= \{x: (x_i-k_i\eta)^2+2(x_i-k_i\eta)t_2+t_3$$

$$< z^2_{\gamma/2}k_i\eta+z^2_{\gamma/2}(x_i-k_i\eta)+t_4 \text{ and } |x_i-k_i\eta| \leq \eta^{9/16} \text{ for all } i\}$$

$$= \{x: \frac{(x_i-k_i\eta)^2}{k_i\eta} + 2\,\frac{(x_i-k_i\eta)}{(k_i\eta)^{1/2}}\,\frac{t_2}{(k_i\eta)^{1/2}} + \frac{t_3}{k_i\eta}$$

$$< z^2_{\gamma/2} + \frac{(x_i-k_i\eta)}{(k_i\eta)^{1/2}}\,\frac{z^2_{\gamma/2}}{(k_i\eta)^{1/2}} + \frac{t_4}{k_i\eta} \text{ and } |x_i-k_i\eta|\leq\eta^{9/16} \text{ for all } i\}$$

$$= \{x: [\frac{(x_i-k_i\eta)}{(k_i\eta)^{1/2}} + t_5]^2 < z^2_{\gamma/2}+t_6 \text{ and } |x_i-k_i\eta|\leq\eta^{9/16} \text{ for all } i\},$$

where for $x \in \Omega \cap V^c$, $t_5 = O(\eta^{-1/2})$ and $t_6 = O(\eta^{-1})$. Now

$$P_\lambda(\lambda \in C^*(X)) = \sum_{\Omega} \prod_{i=1}^{p} \frac{e^{-k_i\eta}(k_i\eta)^{x_i}}{x_i!}$$

$$= \sum_{\Omega\cap V} \prod_{i=1}^{p} \frac{e^{-k_i\eta}(k_i\eta)^{x_i}}{x_i!} + \sum_{\Omega\cap V^c} \prod_{i=1}^{p} \frac{e^{-k_i\eta}(k_i\eta)^{x_i}}{x_i!}.$$

Using a Chebychev argument as in Theorem 2, line (A.13),

$$\sum_{\Omega\cap V} \prod_{i=1}^{p} \frac{e^{-k_i\eta}(k_i\eta)^{x_i}}{x_i!} = o(1).$$

From Theorem 2, line (A.17),

$$\sum_{V^c} \prod_{i=1}^{p} \frac{e^{-k_i\eta}(k_i\eta)^{x_i}}{x_i!} = \sum_{V^c} \prod_{i=1}^{p} ((2\pi)^{-1/2}\exp\{-\frac{1}{2}[\frac{x_i-k_i\eta}{(k_i\eta)^{1/2}}]^2\}$$

$$\cdot[(k_i\eta)^{-1/2} + \frac{3(\frac{x_i-k_i\eta}{(k_i\eta)^{1/2}}) - (\frac{x_i-k_i\eta}{(k_i\eta)^{1/2}})^3}{6k_i\eta}] + R_i)$$

$$= \sum_{V^c} \prod_{i=1}^{p} (2\pi k_i\eta)^{-1/2}\exp\{-\frac{1}{2}(\frac{x_i-k_i\eta}{(k_i\eta)^{1/2}})^2\} + o(1).$$

Therefore

$$\sum_{\Omega\cap V^c} \prod_{i=1}^{p} \frac{e^{-k_i\eta}(k_i\eta)^{x_i}}{x_i!} = \sum_{\Omega\cap V^c} \prod_{i=1}^{p} (2\pi k_i\eta)^{-1/2}\exp\{-\frac{1}{2}(\frac{x_i-k_i\eta}{(k_i\eta)^{1/2}})^2\} + o(1).$$

Let

$$y_i = \frac{x_i-k_i\eta}{(k_i\eta)^{1/2}} + t_5,$$

and $A = \{y: y_i^2 < z_{\gamma/2}^2 + t_6\}$. (Recall that for $y \in A$, $t_6 = O(\eta^{-1})$.)

Then

$$(A.29)\quad \sum_{\Omega\cap V^c} \prod_{i=1}^{p} \frac{e^{-k_i\eta}(k_i\eta)^{x_i}}{x_i!} = \sum_{A} \prod_{i=1}^{p} (2\pi k_i\eta)^{-1/2}\exp\{-\tfrac{1}{2}(y_i-t_5)^2\}+o(1)$$

$$= \sum_{B} \prod_{i=1}^{p} (2\pi k_i\eta)^{-1/2}\exp\{-\tfrac{1}{2}y_i^2\} + o(1),$$

Where $B = \{y: y_i^2 < z_{\gamma/2}^2\}$. Note that the summation in (A.29) is over values $\{0,(k_i\eta)^{-1/2}, 2(k_i\eta)^{-1/2},\ldots\}$ such that $y_i^2 < z_{\gamma/2}^2$ for each i. Let $B_i = \{y_i: y_i^2 < z_{\gamma/2}^2\}$, $i = 1,\ldots,p$. Then

$$\lim_{\eta\to\infty} (k_i\eta)^{-1/2} \sum_{B_i} (2\pi)^{-1/2} \exp\{-\tfrac{1}{2}y_i^2\}$$

$$= \int_{B_i} (2\pi)^{-1/2}\exp\{-\tfrac{1}{2}y_i^2\}dy_i$$

$$= P(|Y| \le z_{\gamma/2}), \text{ where } Y \sim N(0,1)$$

$$= \gamma.$$

Therefore,

$$\lim_{\eta\to\infty} \sum_{\Omega\cap V^c} \prod_{i=1}^{p} \frac{e^{-k_i\eta}(k_i\eta)^{x_i}}{x_i!} = (1-\gamma)^p,$$

and this proves the desired result.

Proof of Theorem 4

Let $g(x)$ denote the multinomial density with parameters N and θ. Clearly

$$N^2 \sum_{i=1}^{p-1} E(\delta_i^*(X)-\theta_i)^2$$

$$= E \sum_{i=1}^{p-1} N^2[\gamma_i+(1-\min\{\frac{K}{N+K}, \frac{1-\sum_{j=1}^{p}\hat{\theta}_j^2}{1-\sum_{j=1}^{p}\hat{\theta}_j^2+N\sum_{j=1}^{p}(\hat{\theta}_j-\gamma_j)^2}\})(\hat{\theta}_i-\gamma_i)-\theta_i]^2$$

$$= \sum_{\mathcal{X}} \sum_{i=1}^{p-1} [N\gamma_i+(1-\min\{\frac{K}{N+K}, \frac{1-\sum_{j=1}^{p}\hat{\theta}_j^2}{1-\sum_{j=1}^{p}\hat{\theta}_j^2+N\sum_{j=1}^{p}(\hat{\theta}_j-\gamma_j)^2}\})(x_i-N\gamma_i)-N\theta_i]^2 g(x)$$

$$= I_V + I_{V^c},$$

where I_V is the summation over the region

$$V = \{x: |x_i| > N^{1/8} \text{ for at least one } i\}$$

and I_{V^c} is the summation over V^c. For $x \in V$, say $|x_m| > N^{1/8}$. Then

$$I_V \leq \sum_{V} \sum_{i=1}^{p-1} [(x_i-N\theta_i)^2+2|x_i-N\gamma_i||x_i-N\theta_i|+(x_i-N\gamma_i)^2]g(x)$$

$$\text{(A.30)} \leq N^{-1/8} \sum_{\mathcal{X}} \sum_{i=1}^{p-1} [(x_i-N\theta_i)^2+2|x_i-N\gamma_i||x_i-N\theta_i|+(x_i-N\gamma_i^2)]|x_m|g(x)$$

$$= o(1),$$

since in the limiting situation (4.4), all moments of the multinomial are $O(1)$.

Now consider the summation I_{V^c}. For $x \in V^c$, the following facts hold:

$$\frac{1-\sum_{j=1}^{p}\hat{\theta}_j^2}{1-\sum_{j=1}^{p}\hat{\theta}_j^2+N\sum_{j=1}^{p}(\hat{\theta}_j-\gamma_j)^2} = \frac{N(1-\sum_{j=1}^{p}\hat{\theta}_j^2)}{N(1-\sum_{j=1}^{p}\hat{\theta}_j^2)+N^2\sum_{j=1}^{p}(\hat{\theta}_j-\gamma_j)^2},$$

$$N(1-\sum_{j=1}^{p}\hat{\theta}_j^2) = N[1-\sum_{j=1}^{p-1}(x_j/N)^2-(1-\sum_{j=1}^{p-1}(x_j/N))^2]$$

$$= 2\sum_{j=1}^{p-1}x_j-\sum_{j=1}^{p-1}x_j^2/N-(\sum_{j=1}^{p-1}x_j)^2/N$$

$$= 2\sum_{j=1}^{p-1}x_j + O(N^{-3/4}),$$

$$N^2\sum_{j=1}^{p}(\hat{\theta}_j-\gamma_j)^2 = N^2\sum_{j=1}^{p-1}(x_j/N-\gamma_j)^2+N^2[(1-\sum_{j=1}^{p-1}x_j/N)-(1-\sum_{j=1}^{p-1}\gamma_j)]^2$$

$$= \sum_{j=1}^{p-1}(x_j-N\gamma_j)^2+(\sum_{j=1}^{p-1}x_j-\sum_{j=1}^{p-1}N\gamma_j)^2.$$

Hence

$$I_{V^c}=\sum_{V^c}\sum_{i=1}^{p-1}[N\gamma_i+(1-\min\{\frac{K}{N+K}, \frac{2\sum_{j=1}^{p-1}x_j+O(N^{-3/4})}{2\sum_{j=1}^{p-1}x_j+\sum_{j=1}^{p-1}(x_j-N\gamma_j)^2+(\sum_{j=1}^{p-1}(x_j-N\gamma_j))^2+O(N^{-3/4})}\})$$

$$\cdot(x_i-N\gamma_i)-N\theta_i]g(x).$$

Let

$$C_1 = 2\sum_{j=1}^{p-1}x_j,\ C_2 = \sum_{j=1}^{p-1}(x_j-N\gamma_j)^2 + (\sum_{j=1}^{p-1}(x_j-N\gamma_j))^2.$$

(A.31) We claim: there exists N_1, $K_1 > 0$ such that if $N > N_1$, then $C_1 + C_2 > K_1$. To prove this claim, note that

if $\sum_{j=1}^{p-1} x_j > 0$, then $C_1+C_2 \geq 2\sum_{j=1}^{p-1} x_j \geq 2$, and

if $\sum_{j=1}^{p-1} x_j = 0$, then $C_1+C_2 \geq \sum_{j=1}^{p-1} (N\gamma_j)^2$.

Now $N\gamma_j \to \mu_j$, $0 < \mu_j < \infty$, $j = 1,\ldots,p-1$. Therefore there exist ϵ and N_2 such that for $N > N_2$, $N\gamma_j > \mu_i - \epsilon > 0$, and hence $C_1 + C_2 > 0$. The claim is thus established. It can be concluded that for $x \in V^c$,

$$N\gamma_i + \left(1 - \frac{C_1+O(N^{-3/4})}{C_1+C_2+O(N^{-3/4})}\right)(x_i - N\gamma_i)$$

$$= x_i - \frac{(C_1+O(N^{-3/4}))(x_i-N\gamma_i)}{C_1+C_2+O(N^{-3/4})}$$

$$= x_i - \frac{(C_1+O(N^{-3/4}))(x_i-N\gamma_i)}{C_1+C_2}\left(1 + \frac{O(N^{-3/4})}{C_1+C_2}\right)$$

$$= x_i - \frac{C_1(x_i-N\gamma_i)}{C_1+C_2} + \frac{(x_i-N\gamma_i)O(N^{-3/4})}{C_1+C_2}$$

$$+ \frac{(C_1+O(N^{-3/4}))(x_i-N\gamma_i)O(N^{-3/4})}{(C_1+C_2)^2}.$$

$$= x_i - \frac{C_1(x_i-N\gamma_i)}{C_1+C_2} + o(1),$$

since from (A.31), $(C_1+C_2)^{-1} < K_1^{-1}$. Thus

$$I_{V^c} = \sum_{V^c} \sum_{i=1}^{p-1} [x_i - \min\{\frac{K}{N+K}, \frac{C_1}{C_1+C_2}\}(x_i - N\gamma_i) - N\theta_i + o(1)]^2 g(x)$$

$$= \sum_{V^c} \sum_{i=1}^{p-1} [x_i - \min\{\frac{K}{N+K}, \frac{C_1}{C_1+C_2}\}(x_i - N\gamma_i) - N\theta_i]^2 g(x) + o(1)$$

$$= \sum_{\mathcal{X}} \sum_{i=1}^{p-1} [x_i - \min\{\frac{K}{N+K}, \frac{C_1}{C_1+C_2}\}(x_i - N\gamma_i) - N\theta_i]^2 g(x)$$

$$- \sum_{V} \sum_{i=1}^{p-1} [x_i - \min\{\frac{K}{N+K}, \frac{C_1}{C_1+C_2}\}(x_i - N\gamma_i) - N\theta_i]^2 g(x) + o(1)$$

$$= \sum_{\mathcal{X}} \sum_{i=1}^{p-1} [x_i - \min\{\frac{K}{N+K}, \frac{C_1}{C_1+C_2}\}(x_i - N\gamma_i) - N\theta_i]^2 g(x) + o(1).$$

The last equality uses arguments as in (A.30). Finally let $\theta_1,\ldots,\theta_{p-1} \to 0$, $\gamma_1,\ldots,\gamma_{p-1} \to 0$, $N \to \infty$, $K \to \infty$, such that $N\theta_i \to \lambda_i$, $N\gamma_i \to \mu_i$, $0 < \lambda_i, \mu_i < \infty$, $i = 1,\ldots,p-1$, and $N/K \to D$. Then

$$\lim \sum_{\mathcal{X}} \sum_{i=1}^{p-1} [x_i - \min\{\frac{K}{N+K}, \frac{2\sum_{j=1}^{p-1} x_j}{2\sum_{j=1}^{p-1} x_j + \sum_{j=1}^{p-1}(x_j - N\gamma_j)^2 + (\sum_{j=1}^{p-1}(x_j - N\gamma_j))^2}\}$$

(A.32)

$$\cdot (x_i - N\gamma_i) - N\theta_i]^2 g(x)$$

$$= \sum_{\mathcal{X}} \sum_{i=1}^{p-1} [x_i - \min\{\frac{1}{D+1}, \frac{2\sum_{j=1}^{p-1} x_j}{2\sum_{j=1}^{p-1} x_j + \sum_{j=1}^{p-1}(x_j - \mu_j)^2 + (\sum_{j=1}^{p-1}(x_j - \mu_j))^2}\}$$

$$\cdot (x_i - \mu_i) - \lambda_i]^2 p(x),$$

where $p(x)$ is the density of $p-1$ independent Poisson random variables with means $\lambda_1,\ldots,\lambda_{p-1}$. Combining (A.30) and (A.32) gives the desired result.