Volume LXXII, Part 1 No. 137

PROCEEDINGS

May 8, 9, 10, 11, 1985

ON STEIN ESTIMATORS: "INADMISSIBILITY" OF ADMISSIBILITY AS A CRITERION FOR SELECTING ESTIMATORS

JAMES E. BUCK

Abstract

Stein estimators are an alternative (non-Bayesian) explanation for credibility. Until this year, the syllabus for Part 4 of the Society's examinations contained an article discussing Stein estimators, or James-Stein estimators, as part of the credibility readings for the exam [2]. The article focuses on some examples where Stein estimators are applied to baseball players' batting averages, among other things. In the examples, Stein estimators seem much like Bayesian credibility estimators and, in fact, credibility estimators derived from Stein's theory have been used by the Insurance Services Office for products liability classification rate-making.

Alike as Stein estimators and Bayesian credibility estimators are in practice, the theory behind Stein estimators is very much different and does not make much sense from the author's point of view. This paper consists of a discussion of the theory that underlies Stein estimators, including an example which illustrates the flaw in logic behind this alternative explanation of credibility.

INTRODUCTION

The literature of the Casualty Actuarial Society has been replete for years with papers on the theory of credibility (for instance, [3], [7], [8]). Practice, at least for most direct lines of business, has lagged far behind. In 1980, the Insurance Services Office (ISO) Credibility Subcommittee [5] produced a comprehensive report on credibility which recommended adoption of an empirical Bayes credibility procedure for products liability classification ratemaking. Normally, one would rejoice at this attempt of life to imitate art. However, the method chosen for use was adapted from the method of Morris and Van Slyke [9], which in turn is based on Stein estimation. Stein estimation is derived from the work of Charles Stein [10] (also, James and Stein [6]), and herein lies the reason for the author's less-than-jubilant reaction to the method of estimation chosen: the theory underlying Stein estimators does not make sense.

From a practical point of view, the adapted Morris-Van Slyke procedure worked better than the Buhlmann-Straub empirical Bayesian procedure in the testing done by the ISO. This is not all that surprising, given that the Morris-Van Slyke procedure is biased upwards and the testing included groups where the expected class loss ratios trended up or down over time. One of the assumptions underlying the Buhlmann-Straub credibility procedure is that the expected loss ratio of a class remains fixed over time. If the expected loss ratio changes, then the credibility to be applied to the most recent experience should be higher, since this recent experience is more related to the expected future experience of the class than the rate based on past class data.

While the Morris-Van Slyke procedure seems to work well in the simulations performed by the ISO, its theoretical flaws make the application of the technique to other problems dangerous. For example, the degree of upward bias in the class credibilities is directly related to the number of classes in the group: in its report, the ISO Subcommittee notes (p. I-19), "An interesting observation is that this process [the adapted Morris-Van Slyke method] effectively produces a minimum credibility of 3/k [where k is the number of classes] for each class in the group." Interesting, indeed. The ISO testing procedure involved groups with between 9 and 24 classes, and so the minimum credibility for each class was between 1/3 and 1/8. However, if the Morris-Van Slyke credibility procedure were applied without adjustment to private passenger auto territorial ratemaking for Rhode Island, the experience of each of the three territories would be given full credibility, regardless of the amount of experience! In this instance, it is clear that the Morris-Van Slyke procedure would not work well. What we have, then, is a procedure which works well in some instances, and yet produces poor results in other instances. Why? In the author's view, it is because the Morris-Van Slyke procedure used by the ISO is based, among other things, on Stein estimators, and Stein estimators are theoretically unsound. To understand the flaw in the theory, it is necessary to review the underlying statistical assumptions that form the basis of the development of Stein estimators.

THE THEORETICAL BASIS FOR STEIN ESTIMATORS

The focus of Bayesian estimation and Bayesian credibility is on modifying an estimate based on additional data. That is, the Bayes approach assumes that we already know something about the parameter to be estimated (the prior distribution). Bayes theorem and Bayesian credibility give us a way to combine that prior knowledge with additional information to produce a revised estimator of the parameter.

Stein estimators are based on a different (sometimes called frequentist or classical) view of estimation. According to this view, it is meaningless to discuss prior distributions of parameters; the parameters of a distribution are fixed values, even though the values may be unknown. Frequentists study the distribution of estimators about parameters in order to make inferences about the quality of different estimators. One of the properties of estimators used for comparison is expected squared error. To use a more specific example, let's take the normal distribution of mean μ and variance 1, or $N(\mu,1)$. If we select a sample point x from the distribution and use it as an estimator of μ, we know from the definition of variance that the estimator has an expected squared error of 1. Are there better estimators of μ? That is a very tough question to answer directly, if you believe that talking about the distribution of μ is meaningless. Since μ is fixed but unknown, there may well be better estimators, depending on the particular value of μ. For instance, if μ happens to be between 1 and 3, the fixed estimator $f(x)=2$ has smaller expected squared error than 1, the expected squared error of the estimator x.

Thus, we need an additional requirement besides low expected squared error if we are to choose among estimators in the frequentist framework. One such requirement is that an estimator be unbiased. An estimator is said to be unbiased if its expected value is always equal to the parameter to be estimated. In terms of the example, an estimator is unbiased if the expected value of the estimator is equal to μ for all values of μ. The sample point, x, is an unbiased estimator of μ and has been shown to be the unbiased estimator of minimum expected squared error (see, for example, [4], pp.362–365).

The requirement that an estimator be unbiased is one way to help define what is meant by best estimate, but in some cases it is felt to be too stringent. After all, an estimator that is biased but with low expected squared error may well be more desirable than an unbiased estimator of high expected squared error. This led to the alternate standard of admissibility for estimators. An estimator is said to be admissible with respect to a loss function (e.g. expected squared error) for a class of distributions if there is no other estimator which has expected squared error less than or equal to the expected squared error of the estimator for all distributions in the class, with the strict inequality holding for at least one distribution. Admissibility certainly sounds like an admirable quality for an estimator to have, but using it produces some disturbing results. In fact, the theoretical basis for Stein estimates is a proof by Stein [10] that the sample mean is not an admissible estimator of the mean of the n-variate normal distribution, $n \geq 3$. (This result is sometimes referred to as Stein's paradox.)

In order to discuss Stein's results, let's review briefly the multivariate normal distribution and its notation. Conceptually an n-variate normal distribution can be thought of as a collection of n separate variables, each normally distributed. Using vector notation, any particular multivariate normal distribution can be specified as $N(\vec{\mu},\Sigma)$, where $\vec{\mu}$ is a mean vector $\vec{\mu}(\mu_1,\ldots,\mu_n)$, with μ_i representing the mean of the i-th variable, and Σ is a symmetrical n-by-n covariance matrix, with each element of the matrix, σ_{ij}^2, representing the covariance between the i-th and j-th variables. If the n-variate distribution is independent, then the covariances between variables are equal to zero, and Σ is a diagonal matrix.

Stein considers the task of estimating $\vec{\mu}(\mu_1,\ldots,\mu_n)$ given a single sample point $\vec{x}(x_1,\ldots,x_n)$ selected from the multivariate independent normal distribution of variance 1, i.e., $\vec{x} \sim N(\mu, I)$, where I is the identity matrix. The usual estimator, $\vec{x}$, has expected squared error of n, the number of parameters to be estimated. James and Stein [6] developed an estimator with smaller squared error. The development of the estimator is based on the following property of the multivariate normal distribution: for any point $\vec{p}$,

$$P(|\vec{x} - \vec{p}| > |\vec{\mu} - \vec{p}|) > .50 \ .$$

In words, there is always a better than even chance that a point chosen at random from the multivariate normal distribution is farther away from $\vec{p}$ than $\vec{\mu}$, the mean of the distribution, is from $\vec{p}$, no matter what $\vec{p}$ is chosen to be. Stein estimators which shrink $\vec{x}$ to an arbitrary $\vec{p}$ by a factor of

$$\frac{n - 2}{|\vec{x} - \vec{p}|^2}$$

have smaller expected squared error than $\vec{x}$ for all $\vec{\mu}$. That is,

$$E|\hat{\mu} - \vec{\mu}|^2 < n,$$

$$\text{for } \hat{\mu} = \left[1 - \frac{n-2}{|\vec{x} - \vec{p}|^2}\right]\vec{x} + \frac{n-2}{|\vec{x} - \vec{p}|^2}\vec{p},\ n \geq 3$$

When Stein estimators are applied to problems, $\vec{p}$ is usually chosen to be the average result for the group—in the notation above, the average of the x_i—and the resulting formula looks a lot like a Bayesian credibility estimate.

It's *important to note*, however, that there is *no* requirement that $\vec{p}$ be chosen as the average of the group in the theoretical work by Stein. And this flexibility with regard to $\vec{p}$ produces unusual results, particularly if we change the frame of reference. For instance, consider the three-dimensional case, where we select $\vec{x}(x_1,x_2,x_3)$ from a multivariate normal distribution of mean $\vec{\mu}(\mu_1,\mu_2,\mu_3)$ and covariance matrix I, the identity matrix. To make the presentation simpler, let $\vec{x} = (0,0,0)$, the origin. According to Stein, $\vec{x}$ can be combined with any arbitrary $\vec{p}$ (shrunk toward $\vec{p}$) to produce a better estimate of $\vec{\mu}$. For example, if we select $\vec{p} = (1,0,0)$, the Stein estimate combining $\vec{p}$ and $\vec{x}$ is $0\vec{x} + 1\vec{p}$, or $\vec{p}$ itself. In fact, for any point chosen from the sphere of radius 1 centered at origin, the estimate is the point itself. Thus, every point on the sphere of radius 1 centered at the origin is a "better" estimate of $\vec{\mu}$ than $\vec{x}$, the origin.

If that were not unusual enough, we can go further and show that any point $\vec{a}$ is a Stein estimate of $\vec{\mu}$, if we select an appropriate $\vec{p}$. The $\vec{p}$ to choose, for any given $\vec{a}$, is determined from the formula $\vec{a}/|\vec{a}|^2$. So, to show that $\vec{a} = (100,0,0)$ is a Stein estimate, we need only choose $\vec{p} = (.01,0,0)$. Therefore, based on the theory underlying Stein estimators, even a point as far away as $(100,0,0)$ is a better estimate of $\vec{\mu}$ than $\vec{x} = (0,0,0)$, the sample point!

THE CIRCLE DISTRIBUTION

To understand what's wrong with Stein estimators, it helps to go through the development of a Stein-like estimator for a simpler distribution. The chosen distribution is the one-dimensional distribution defined on a plane by the function

$$\begin{aligned} f(x_1,x_2) &= \frac{1}{2\pi},\ x_1^2 + x_2^2 = 1 \\ &= 0, \text{ elsewhere.} \end{aligned}$$

This distribution represents the chance of randomly picking a point on the circle of radius 1 centered at the origin. The mean of the distribution is also the origin.

The circle distribution was chosen because from any point $\vec{p}$ on the plane, there is a better than 50% chance that the distance between a randomly selected point on the circle and $\vec{p}$ will be greater than the distance between the origin and $\vec{p}$. Geometrically, we can see this by noting that, for any point $\vec{p}$, the arc around $\vec{p}$ through the origin contains less than half the circle. Because the circle distribution shares this property with the multivariate normal distribution, we should be able to shrink the values of the circle distribution to an arbitrary $\vec{p}$ and get an estimate that is, on average, closer to the mean. Indeed, if we notice that, for any $\vec{p} = (p_1, p_2)$, the average squared distance between the circle distribution and $\vec{p}$ is

$$\frac{1}{2\pi}\int_0^{2\pi} (p_1 - \sin\theta)^2 + (p_2 - \cos\theta)^2 \, d\theta$$

$$= \frac{1}{2\pi}\int_0^{2\pi} p_1^2 - 2p_1 \sin\theta + \sin^2\theta + p_2^2 - 2p_2\cos\theta + \cos^2\theta \, d\theta$$

$$= p_1^2 + p_2^2 + 1,$$

we might consider estimators of the form

$$\hat{\mu} = \frac{1}{|\vec{x} - \vec{p}|^2 + c}\,\vec{p} + \frac{|\vec{x} - \vec{p}|^2 + c - 1}{|\vec{x} - \vec{p}|^2 + c}\,\vec{x}.$$

And, in fact, Appendix I shows that if c is greater than $(|\vec{x} - \vec{p}| + 1)^2$, the expected squared error of this estimator, $\hat{\mu}$, is always less than 1, the expected squared error of the usual estimator, $\vec{x}$.

If one were to take the classical viewpoint, and the viewpoint that underlies the standard of admissibility of estimators, we should use this form of estimator in determining $\vec{\mu}$, given a particular $\vec{x}$. The fallacy in this approach can be seen by taking a Bayesian point of view. Let's again use the circle distribution of radius 1 and choose at random a point $\vec{x}$ from a circle of radius 1 with an unknown center. Without loss of generality, we can set $\vec{x} = (0,0)$. Now, we want to estimate the center of the circle, given that $\vec{x}$ is a point on the circle. If we consider all possible circles of radius 1 equally likely, then a good candidate for the distribution of $f(\vec{\mu}|\vec{x} = (0,0))$ would be

$$f(\vec{\mu}) = \frac{1}{2\pi}, \quad \mu_1^2 + \mu_2^2 = 1$$

$$= 0, \text{ elsewhere.}$$

In fact, if we represent equally likely (or no prior knowledge) as the prior distribution

$$h(\vec{\mu}) \lim_{n\to\infty} g_n(\mu_1,\mu_2) = \frac{1}{4n^2}, \ -n \le \mu_1 \le n, \ -n \le \mu_2 \le n$$

$$= 0, \text{ elsewhere},$$

among others, then the candidate distribution shown as $f(\vec{\mu})$ above can be derived through the use of Bayes Theorem for continuous functions (see Appendix II).

Now, from a Bayesian point of view, we have determined the distribution of $f(\vec{\mu}|\vec{x})$. The next step is to determine the best point etimate of the $\vec{\mu}$ distribution (uniform distribution on the unit circle centered at the origin). The squared error function between the $\vec{\mu}$ distribution and any estimate $\vec{e} = (e_1,e_2)$ is given by

$$\frac{1}{2\pi}\int_0^{2\pi} (\sin\theta - e_1)^2 + (\cos\theta - e_2)^2 \, d\theta$$

$$= 1 + e_1^2 + e_2^2$$

which obviously is at a minimum at (0,0), or $\vec{x}$.

Stein estimation takes another approach. Stein's argument in this case would be, let us select an arbitrary point $\vec{p}$, say $\vec{p} = (2,0)$. It was previously shown that, if we shrink the $\vec{\mu}$'s to $\vec{p}$ by a factor of

$$1 - \frac{1}{|\vec{\mu} - \vec{p}|^2 + c}, \ c \ge (|\vec{\mu} - \vec{p}| + 1)^2,$$

the transposed $\vec{\mu}$'s are closer to $\vec{x}$. Based on this, it is therefore appropriate to shift $\vec{x}$ towards $\vec{p}$ by a factor of

$$1 - \frac{1}{|\vec{x} - \vec{p}|^2 + c}, \ c = (|\vec{x} - \vec{p}| + 1)^2$$

(or, equivalently, choose an estimate of (2/13, 0)) to give a better estimate of $\vec{\mu}$.

A geometric analogy may be of some help in understanding this point. Figure 1 shows the problem in graphical terms, from a Bayesian standpoint. Imagine that $\vec{\mu}$ represents the rim of a dartboard attached to the back of a door, and $\vec{p}$ represents the doorknob. The problem is to place a dart on the wall that is closest, on average, to the points on the rim of the dartboard ($f(\vec{\mu}|\vec{x})$). From the calculations above, and from common sense, we can see that the dart should be placed at the center of the dartboard ($\vec{x}$).

Figures 2 and 3 represent the Stein estimator approach. Figure 2 shows that if one squishes the rim of the dartboard a bit towards the doorknob (shifts the $\vec{\mu}$'s), there is a smaller average distance between the rim of the dartboard and the center of the dartboard ($\vec{x}$). This is then used to justify aiming the dart at a point closer to the doorknob, even though the problem is to get as close as possible, on average, to the rim of the original (unshifted) dartboard (Figure 3).

APPLICABILITY TO MULTIVARIATE NORMAL DISTRIBUTION

While it is easier to see the fallacy of admissibility and Stein estimators with respect to the circle distribution, Stein estimators are equally invalid for the multivariate normal distribution. Let's again take the problem of estimating $\vec{\mu}$ given $\vec{x}$, $\vec{x} \sim N(\vec{\mu}, I)$.

Using a variety of "flat" prior distributions, including $N(0,\infty)$ and the rectangular distribution used above, we can derive $\vec{\mu} \sim N(\vec{x}, I)$. Here, also, the standard of admissibility asks the wrong question from the Bayesian viewpoint. The proper question to ask is not what function $f(\vec{\mu})$ minimizes $E|f(\vec{\mu}) - \vec{x}|^2$, but rather, what value $\vec{p}$ minimizes $E|\vec{\mu} - \vec{p}|^2$. Because the multivariate normal distribution is independent and can be expressed as the product of one-dimensional normal distributions, the minimum of $\vec{p}$ at $\vec{x}$ follows from the fact that the squared distance function is minimized at the mean in the one-dimensional case.

From a theoretical point of view, it would seem that the major accomplishment of Stein estimators is to show that admissibility as applied by Stein isn't a very good criterion for choosing estimators and that the Bayesian theory of estimation, when properly applied, gives consistent and reasonable results. In fact, Stein's paradox is not a paradox at all when viewed from a Bayesian standpoint. From a practical point of view, biased estimators are still appropriate to use in many cases, but not those derived from this particular theory of estimation.

ACKNOWLEDGEMENTS

The author would like to thank Messrs. Chris Svendsgaard and Steven Yaset of the ISO for their review and comments on earlier drafts of this paper. The Committee on Review of Papers was also very helpful in this regard.

Figure 1

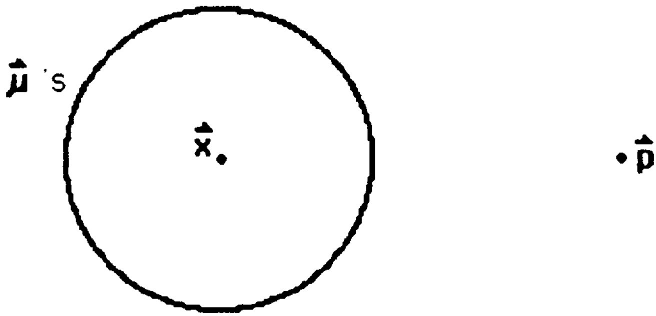


Figure 2

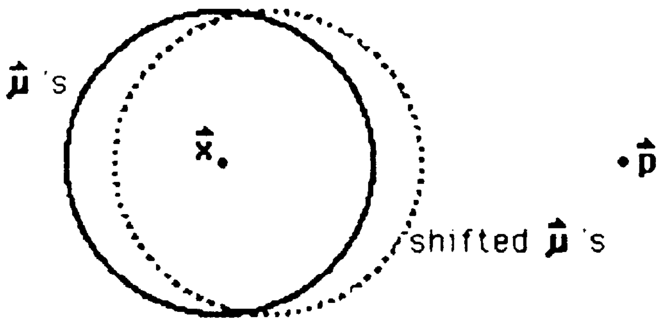
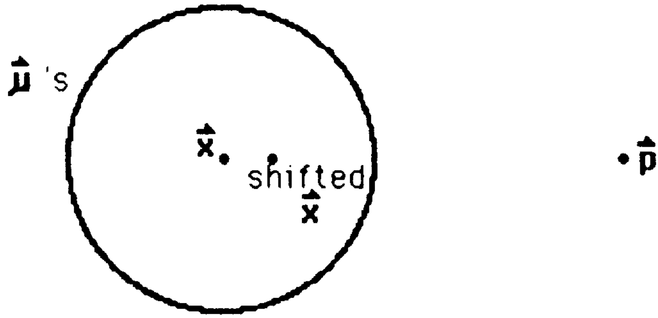


Figure 3



REFERENCES

[1] B. Efron, "Controversies in the Foundations of Statistics", *American Mathematical Monthly,* April, 1978, pp. 231–245.

[2] B. Efron and C. Morris, "Stein's Paradox in Statistics", *Scientific American,* May, 1977, pp. 119–127.

[3] C. C. Hewitt, "Credibility for Severity", *PCAS* LVII, 1970.

[4] P. G. Hoel, *Introduction to Mathematical Statistics,* 4th Edition, New York: John Wiley and Sons, Inc., 1971.

[5] Insurance Services Office, *Report of the Credibility Subcommittee: Development and Testing of Empirical Bayes Credibility Procedures for Classification Ratemaking,* New York, Insurance Services Office, 1980.

[6] W. James and C. Stein, "Estimation with Quadratic Loss", *Proceedings of the Fourth Berkeley Symposium,* Vol. 1, Berkeley, University of California Press, 1961, pp. 361–379.

[7] A. L. Mayerson, "A Bayesian View of Credibility", *PCAS* LI, 1964.

[8] A. L. Mayerson, D. A. Jones and N. L. Bowers, Jr., "On the Credibility of the Pure Premium", *PCAS* LV, 1968.

[9] C. Morris, and O. E. Van Slyke, "Empirical Bayes Methods for Pricing Insurance Classes", *Proceedings of the Business and Economics Statistics Section,* American Statistical Association, 1978.

[10] C. Stein, "Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution", *Proceedings of the Third Berkeley Symposium,* Vol. 1, Berkeley, University of California Press, 1955, pp. 197–206.

APPENDIX I

DEMONSTRATION THAT THE MEAN IS INADMISSIBLE AS AN ESTIMATOR OF THE CIRCLE DISTRIBUTION

The following shows that there is a function that combines a data point, $\vec{x}$, with any $\vec{p}$ to produce an estimate of $\vec{\mu} = (0,0)$ that has expected squared error less than 1, the squared error of $\vec{x}$. This treatment is consistent with the frame of reference discussed in the text. However, this is equivalent to showing that for a circle distribution centered at $\vec{p}$, there is a function which combines $\vec{x}$, a randomly selected point on the circle, and the origin to produce an estimate of $\vec{p}$, the mean of the circle distribution, with expected squared error of less than 1. We consider estimators of the form

$$\frac{1}{|\vec{x} - \vec{p}|^2 + c}\vec{p} + \frac{|\vec{x} - \vec{p}|^2 + c - 1}{|\vec{x} - \vec{p}|^2 + c}\vec{x},\ c \geq 1$$

$$= \vec{x} + \frac{\vec{p} - \vec{x}}{|\vec{x} - \vec{p}|^2 + c}\,.$$

For $\vec{x} = (\sin\theta, \cos\theta)$ and $\vec{p} = (p_1, p_2)$, the expected squared error is given by

$$\frac{1}{2\pi}\int_0^{2\pi}\left[\sin\theta + \frac{p_1 - \sin\theta}{(p_1 - \sin\theta)^2 + (p_2 - \cos\theta)^2 + c}\right]^2$$

$$+ \left[\cos\theta + \frac{p_2 - \cos\theta}{(p_1 - \sin\theta)^2 + (p_2 - \cos\theta)^2 + c}\right]^2 d\theta$$

$$= \frac{1}{2\pi}\int_0^{2\pi} \sin^2\theta + \frac{2(p_1 - \sin\theta)\sin\theta}{(p_1 - \sin\theta)^2 + (p_2 - \cos\theta)^2 + c}$$

$$+ \frac{(p_1 - \sin\theta)^2}{[(p_1 - \sin\theta)^2 + (p_2 - \cos\theta)^2 + c]^2} + \cos^2\theta$$

$$+ \frac{2(p_2 - \cos\theta)\cos\theta}{(p_1 - \sin\theta)^2 + (p_2 - \cos\theta)^2 + c}$$

$$+ \frac{(p_2 - \cos\theta)^2}{[(p_1 - \sin\theta)^2 + (p_2 - \cos\theta)^2 + c]^2}\, d\theta$$

$$< \frac{1}{2\pi} \int_0^{2\pi} 1 + \frac{2(p_1 - \sin\theta)\sin\theta + 2(p_2 - \cos\theta)\cos\theta + 1}{(p_1 - \sin\theta)^2 + (p_2 - \cos\theta)^2 + c}\, d\theta$$

$$= \frac{1}{2\pi} \int_0^{2\pi} 1 + \frac{2p_1 \sin\theta + 2p_2 \cos\theta - 1}{p_1^2 - 2p_1 \sin\theta + p_2^2 - 2p_2 \cos\theta + 1 + c}\, d\theta$$

$$= \frac{1}{2\pi} \int_0^{2\pi} 1 + \frac{2p_1 \sin\theta + 2p_2 \cos\theta - 1}{2p_1 \sin\theta + 2p_2 \cos\theta - 1 - (p_1^2 + p_2^2 + c)}\, d\theta$$

Using the relation $\dfrac{a}{a-b} = 1 + \dfrac{b}{a-b}$,

$$\frac{1}{2\pi} \int_0^{2\pi} 1 - \left[1 + \frac{p_1^2 + p_2^2 + c}{2p_1 \sin\theta + 2p_2 \cos\theta - 1 - (p_1^2 + p_2^2 + c)}\right] d\theta$$

$$= \frac{-(p_1^2 + p_2^2 + c)}{2\pi} \int_0^{2\pi} \frac{d\theta}{2p_1 \sin\theta + 2p_2 \cos\theta - (p_1^2 + p_2^2 + c + 1)}$$

$$= \frac{-(p_1^2 + p_2^2 + c)}{2\pi} \int_{-\pi}^{\pi} \frac{d\theta}{\left[\begin{array}{c} 2p_1 \sin(\pi + \theta) + 2p_2 \cos(\pi + \theta) \\ - (p_1^2 + p_2^2 + c + 1) \end{array}\right]}$$

$$= \frac{p_1^2 + p_2^2 + c}{2\pi} \int_{-\pi}^{\pi} \frac{d\theta}{2p_1 \sin\theta + 2p_2 \cos\theta + p_1^2 + p_2^2 + c + 1}$$

Using integral tables, we find that the above is equivalent to

$$\frac{p_1^2 + p_2^2 + c}{2\pi} \cdot \frac{2}{\sqrt{(p_1^2 + p_2^2 + c + 1)^2 - (2p_1)^2 - (2p_2)^2}}$$

$$\tan^{-1} \frac{2p_1 + (p_1^2 + p_2^2 + c + 1 - 2p_2)\tan(\theta/2)}{\sqrt{(p_1^2 + p_2^2 + c - 1)^2 - (2p_1)^2 - (2p_2)^2}} \Bigg|_{-\pi}^{\pi}$$

$$= \frac{p_1^2 + p_2^2 + c}{2\pi} \cdot \frac{2\pi}{\sqrt{(p_1^2 - p_2^2 + c + 1)^2 - (2p_1)^2 - (2p_2)^2}}$$

$$= \sqrt{\frac{(p_1^2 + p_2^2 + c)^2}{(p_1^2 - p_2^2 + c)^2 + 2(p_1^2 + p_2^2 + c) + 1 - 4p_1^2 - 4p_2^2}}$$

$$= \sqrt{\frac{(p_1^2 + p_2^2 + c)^2}{(p_1^2 + p_2^2 + c)^2 + 2c + 1 - 2p_1^2 - 2p_2^2}}$$

So, if $c > p_1^2 + p_2^2 - 1/2$, the squared error is less than 1. In particular, since $(|\vec{p} - \vec{x}| + 1)^2 > \mathrm{p}_1^2 + \mathrm{p}_2^2 - 1/2$, if we choose $c \geq (|\vec{p} - \vec{x}| + 1)^2$, the estimator has expected squared error of less than 1.

APPENDIX II

DERIVATION OF THE POSTERIOR DISTRIBUTION USING THE CIRCLE DISTRIBUTION AND A "FLAT" PRIOR DISTRIBUTION

The purpose of this appendix is to determine $f(\vec{\mu}|\vec{x})$ for

$$f(\vec{x}|\vec{\mu}) = \frac{1}{2\pi}, \quad |\vec{x} - \vec{\mu}|^2 = 1$$

$$= 0, \text{ elsewhere and}$$

$$h(\vec{\mu}) = \lim_{n\to\infty} g_n(\mu_1,\mu_2) = \frac{1}{4n^2}, \quad -n \le \mu_1 \le n, \ -n \le \mu_2 \le n,$$

$$= 0, \text{ elsewhere.}$$

For $\vec{x} = (0,0)$ and any particular $n \ge 1$, the joint distribution is given by

$$f(\vec{x}|\vec{\mu})g_n(\vec{\mu}) = \frac{1}{8\pi n^2}, \quad |\vec{x} - \vec{\mu}|^2 = 1, \ -n \le \mu_1 \le n, \ -n \le \mu_2 \le n,$$

$$= 0, \text{ elsewhere.}$$

$$= \frac{1}{8\pi n^2}, \quad |\vec{\mu}|^2 = 1,$$

$$= 0, \text{ elsewhere, and}$$

$$\int f(\vec{x}|\vec{\mu})g_n(\vec{\mu})\, d\vec{\mu} = \int_0^{2\pi} \frac{d\theta}{8\pi n^2}$$

$$= \frac{1}{4n^2}$$

From Bayes Theorem for continuous functions, we have, for all $n \ge 1$,

$$f_n(\vec{\mu}|\vec{x}) = \frac{f(\vec{x}|\vec{\mu})g_n(\vec{\mu})}{\int f(\vec{x}|\vec{\mu})g_n(\vec{\mu})\, d\vec{\mu}}$$

$$= 4n^2 \cdot \frac{1}{8\pi n^2} = \frac{1}{2\pi}, \quad |\vec{\mu}|^2 = 1,$$

$$= 0, \text{ elsewhere.}$$

and thus the distribution of $f(\vec{\mu}|\vec{x})$ is given by

$$f(\vec{\mu}|\vec{x}) = \frac{f(\vec{x}|\vec{\mu})\ h(\vec{\mu})}{f(\vec{x}|\vec{\mu})\ h(\vec{\mu})\ \mathrm{d}\vec{\mu}}$$

$$= \lim_{n\to\infty} f_n(\vec{\mu}|\vec{x})$$

$$= \frac{1}{2\pi}\quad , \quad |\vec{\mu}|^2 = 1,$$

$$= 0, \text{ elsewhere.}$$