Bayesian vs. Frequentist Shrinkage in Multivariate Normal Problems

7.1 Preliminaries

This chapter is dedicated to the comparison of Bayes and frequentist estimators of the mean $\theta$ of a multivariate normal distribution in high dimensions. For dimension $k \geq 3$, the James–Stein estimator specified in (2.15) (and its more general form to be specified below) is usually the frequentist estimator of choice. The estimator is known to improve uniformly upon the sample mean vector $\bar{X}$ as an estimator of $\theta$ when $k \geq 3$, and while it is also known that it is not itself admissible, extant alternative estimators with smaller risk functions are known to offer only very slight improvement. For this and other reasons, the James–Stein estimator is widely used among estimators which exploit the notion of shrinkage. In the results described in this chapter, I will use the form of the James–Stein estimator which shrinks $\bar{X}$ toward a (possibly nonzero) distinguished point. This serves the purpose of placing the James–Stein estimator and the Bayes estimator of $\theta$ with respect to a standard conjugate prior distribution in comparable frameworks, since the latter also shrinks $\bar{X}$ toward a distinguished point. It is interesting to note that the James–Stein estimator has a certain Bayesian flavor that goes beyond the empirical Bayes character highlighted in the writings of Efron and Morris (1973, etc.) in that the act of shrinking toward a particular parameter vector suggests that the statistician using this estimator is exercising some form of introspection in determining a good “prior guess” at $\theta$. The Bayesian of course goes further, specifying, a priori, the weight he wishes to place on the prior guess. What results in the latter case is an alternative form of shrinkage, one that leads to a linear combination of $\bar{X}$ and the prior guess, with weights influenced by the prior distribution rather than by the observed data. Since $\bar{X}$ is a sufficient statistic for the mean of a multivariate normal distribution with known variance-covariance matrix $\Sigma$, I will henceforth, without loss of generality, take the sample size $n$ to be 1.

A general treatment of a comparison between Bayesian and frequentist shrinkage in estimating the mean vector of the distribution $\mathcal{N}(\theta, \Sigma)$ remains, as of the present exposition, intractable. I will begin by defining the threshold problem in the general case. I will then turn to the special case from which considerable insight and intuition
can be gleaned and on which the work presented in this chapter is focused. The core of this chapter is based on two papers in which the solution of the version of the threshold problem treated here is given. The reader is referred to Vestrup and Samaniego (2004a, 2004b) for detailed proofs and further discussion.

As in Chapters 4 and 5, we will posit the existence of a “true prior distribution” $G_0$ representing the true state of nature in a given estimation problem. As before, the special case in which $G_0$ is degenerate at a point $\theta_0$ — the true but unknown value of the target parameter $\theta$ — will play a prominent role in the analysis pursued here, as this assumption is appropriate in most applications of interest, where $\theta$ is simply an unknown $k$-dimensional vector. In a general treatment of the threshold problem, one would typically make the following assumptions regarding the true prior $G_0$, the operational prior $G$, the sampling distribution of $X$ and the loss function $L$:

$$G_0 : \theta \sim \mathcal{N}_k (\theta_0, \Sigma_0) \quad (7.1)$$

$$G : \theta \sim \mathcal{N}_k (\theta_G, \Sigma_G) \quad (7.2)$$

$$F_{X|\theta} : X|\theta \sim \mathcal{N}_k (\theta, \Sigma) \quad (7.3)$$

where $\Sigma$ is a known positive definite matrix, and

$$L(\theta, a) = (\theta - a)' \Sigma^{-1} (\theta - a) . \quad (7.4)$$

The Bayes risk $E_{G_0} E_{X|\theta} L(\theta, \hat{\theta}_G)$, relative to the true prior $G_0$, of the Bayes estimator $\hat{\theta}_G$ wrt the operational prior $G$ is shown in Vestrup and Samaniego (2003) to be

$$r(G_0, \hat{\theta}_G) = \text{tr} \left( \Sigma^{-1/2} A \Sigma A' \Sigma^{-1/2} \right) + \text{tr} \left( \Sigma^{-1/2} B \Sigma_0 B' \Sigma^{-1/2} \right) + \| \Sigma^{-1/2} B (\theta_G - \theta_0) \|^2 , \quad (7.5)$$

where $A = \Sigma_G (\Sigma_G + \Sigma)^{-1}$ and $B = I - A$. The James–Stein estimator which shrinks $X$ toward the constant vector $\theta^*$ will be denoted by $\hat{\theta}_{JS, \theta^*}$. The Bayes risk of the $\hat{\theta}_{JS, \theta^*}$ (in its general form, i.e., applicable to estimating the mean $\theta$ in the model (7.3)) relative to the conjugate prior $G_0$ in (7.1) is also derived in Vestrup and Samaniego (2004b), and is shown to be given by an infinite series involving expectations of rather complex functions of an infinite collection of Poisson random variables. Inspection of these two expressions for Bayes risk makes clear that the determination of the class of operational priors $G$ for which the Bayes estimator $\hat{\theta}_G$ is superior to the James–Stein estimator $\hat{\theta}_{JS, \theta^*}$ which shrinks $X$ toward the vector $\theta_G$, that is, for which

$$r \left( G_0, \hat{\theta}_G \right) \leq r \left( G_0, \hat{\theta}_{JS, \theta^*} \right) \quad (7.6)$$

is not a tractable exercise. (I have nonetheless listed this version of the threshold problem as Exercise 7.1, and I invite the highly motivated, long-suffering reader to mail me his or her solution!)

The following simplifications do lead to a definitive solution which yields substantial insight. Consider the following special case of the framework in (7.1)–(7.4):