AUTOMATED BIAS-VARIANCE TRADE-OFF: INTUITIVE INADMISSIBILITY OR INADMISSIBLE INTUITION?

XIAO-LI MENG

meng@stat.harvard.edu

ABSTRACT. Seeking an appropriate bias-variance trade-off is a common challenge for any sensible statistician, especially those at the forefront of statistical applications. Recently, addressing a class of bias-variance trade-off problems for studying gene-environment interactions, Mukherjee and Chatterjee (2008) adopted an approximate empirical partially Bayes approach to derive an estimator that amounts to using the following weighted estimator as a compromise:

$$\hat{\beta}_c = \frac{(\hat{\beta}-\hat{\beta}_0)^2}{\hat{\mathrm{V}}(\hat{\beta})+(\hat{\beta}-\hat{\beta}_0)^2}\hat{\beta} + \frac{\hat{\mathrm{V}}(\hat{\beta})}{\hat{\mathrm{V}}(\hat{\beta})+(\hat{\beta}-\hat{\beta}_0)^2}\hat{\beta}_0.$$

Here $\hat{\beta}$ and $\hat{\mathrm{V}}(\hat{\beta})$ are respectively our point estimator and its variance estimate of a parameter β under a model, and $\hat{\beta}_0$ is a more efficient estimator of β under a sub-model via fixing a nuisance parameter. The intuition here appears to be that since $\hat{B} = \hat{\beta}_0 - \hat{\beta}$ is an estimate of the bias in $\hat{\beta}_0$ when the sub-model fails, $\hat{\beta}_c$ should automatically give more weight to the robust $\hat{\beta}$ or the efficient $\hat{\beta}_0$ depending whether or not $\hat{B}^2$ is larger than $\hat{\mathrm{V}}(\hat{\beta})$. The implication here seems to be that the original $\hat{\beta}$ is inadmissible in terms of MSE because it is dominated by $\hat{\beta}_c$, which appears to possess this magic *self-adjusting* mechanism for bias-variance trade-off without needing any assumption beyond those that guarantee the validity of the original $\hat{\beta}$. But is this intuition itself admissible? This question was posed as a Ph.D. qualifying exam problem at Harvard, in the context of a bivariate normal model. This article documents this examination, and concludes with a suggestion of revisiting the classic theory of admissibility, to which Professor Jim Berger has made fundamental contributions. The investigation also reveals a partial shrinkage phenomenon of the partially Bayes method, as well as a misguided insight in the literature of gene-environment interaction studies. Parts of this article adopt an interlacing style interweaving research investigations with pedagogical probes, honoring Berger's prolific contributions in both endeavors.

Date: March 14, 2010.

Key words and phrases. Admissibility, Bias, Empirical Bayes, Mean-Squared Error, Partial Shrinkage, Self-efficiency, Stein Estimator.

Prepared for *Frontiers of Statistical Decision Making and Bayesian Analysis*, a volume in honor of J.O. Berger's 60th birthday. The author thanks Alan Agresti, Joe Blitzstein and Xianchao Xie for constructive comments, Bhramar Mukherjee and Nilanjan Chatterjee for their truly inspirational article, and NSF for partial financial support.

1. Always a Good Question ...

To many students, job candidates, and even some seasoned seminar speakers, a few faculty members are known to be "intimidating". We pose tough questions, demand intuitive explanations, challenge superficial answers, and we do so almost indiscriminately. A few students have expressed their surprise to me: "How could you guys be able to pick on almost any topic, and ask those penetrating questions even for things you apparently have never worked on?" I cannot speak for my fellow challengers, but the students are certainly correct that I have never worked on many of these topics, some of which I heard for the first time before I posed a question. If there is any secret—or bragging—here, it is the one that many senior statisticians work hard to pass on to our future generations. That is, there are only a very few fundamental principles in statistics, and the *bias-variance trade-off* is one of them. It is so deeply rooted in almost any statistical analysis, whether the investigator/speaker realizes it or not. Equipped with a few such powerful weapons, one can fire essentially in any situation, and almost surely not miss the target by too much.

The story I am about to tell is squarely a case of understanding bias-variance trade-off. In the context of a genetic study, a speaker mentioned a recent proposal by Mukherjee and Chatterjee (2008; hereafter M&C) for automatically achieving an appropriate bias-variance trade-off. The moment I saw the proposed formula, as given in the abstract, I knew silence would not be golden in this case. The method, if it has the properties it was designed for, would have profound general implications given its simplicity and the practical demand for such "automated" methods. For the very same reason, however, it could do serious damage if it is applied indiscriminately but without its advertised properties in real terms.

As it happened, shortly after the seminar I needed to submit a problem for our Ph.D. qualifying examination. What could be more appropriate to test students' understanding of bias-variance trade-off, and at the same time their ability to carry out a rigorous investigation of a seemingly intuitive idea? The resulting qualifying exam problem is reproduced in Section 4 below, and the annotated solution is given in Section 5. Before presenting these materials *in verbatim*, which document an effort of integrating research investigation with pedagogical exploration, obviously the stage needs to be set. This is accomplished by Section 2, which discusses a gene-environment interaction study that motivates M&C; and by Section 3, which illustrates M&C's partially empirical Bayes approach via a bivariate normal example. Sections 2 and Section 3 also reveal, respectively, a misguided approximation in the literature of gene-environment interactions, and a partial shrinkage phenomenon of partially Bayes methods, and therefore they may be of independent interest. Indeed, Section 6 concludes with a suggestion of revisiting the classic theory of admissibility but with *partially*

Bayes risk, which is also hoped to be a piece of admissible cake to the birthday-cake tasting (testing?) event for Jim Berger, an amazingly prolific researcher and Ph.D. adviser.

2. Gene-Environment Interaction and A Misguided Insight

2.1. **Estimating Multiplicative Interaction Parameter.** The motivation for M&C's proposal appears to be the need to address a bias-variance trade-off in studying gene-environment interactions. Following their setup, let E and G be respectively a binary environmental factor and a binary genetic factor, and D be the binary disease indicator; value "1" of any these binary variables indicates the presence (e.g., exposed, carrier, or with disease). One key interest here is to assess if there is a gene-environment (G-E) interaction in their impact on the odds of developing the disease. Let

$$O(G,E) = \frac{\Pr(D=1|G,E)}{\Pr(D=0|G,E)}, \tag{2.1}$$

that is, the odds of disease in the sub-population defined by the pair $\{G,E\}$. Then the so-called *multiplicative interaction parameter* ψ is defined as

$$\psi = \frac{O(0,0)O(1,1)}{O(1,0)O(0,1)}, \tag{2.2}$$

which can be remembered as "odds ratio of odds", by analogy with the well-known ratio of cross-product expression of an odds ratio (OR), namely, the OR for a bivariate binary distribution $P(i,j) = \Pr(X=i, Y=j)$, expressed as

$$OR_{(X,Y)} = \frac{P(0,0)P(1,1)}{P(0,1)P(1,0)}. \tag{2.3}$$

This mathematical analogy also helps us to see why ψ is a useful parameter for assessing whether the factors G and E contribute to the odds of disease in a multiplicative fashion, that is, whether we can write $O(G,E) = g(G)e(E)$ for some functions g and e. This is because the mathematical reasoning behind the theorem "$OR_{(X,Y)} = 1$ if and only if $P(i,j)$ factors" is identical to that for "$\psi = 1$ if and only if $O(G,E)$ factors."

Consequently, by assessing whether $\beta = \log(\psi) = 0$, we can infer whether the effects of G and E are additive on the logit scale of the disease rate $\Pr(D=1|G,E)$. In general, to estimate β directly (and therefore to assess it) would require a representative sample of $\{D,G,E\}$, as hinted by its expression in (2.2). Note however, by Bayes' Theorem,

$$O(G,E) = \frac{P(G,E|D=1)\Pr(D=1)}{P(G,E|D=0)\Pr(D=0)} \propto \frac{P(G,E|D=1)}{P(G,E|D=0)}. \tag{2.4}$$

It is then easy to verify that $\psi = OR_1/OR_0$, and hence

$$\beta = \log(\psi) = \log(OR_1) - \log(OR_0) \equiv \beta_0 - \theta, \tag{2.5}$$

where $OR_i = OR_{(G,E|D=i)}$ is the odds ratio for the conditional bivariate binary distribution $P(G, E|D = i), i = 0, 1$. Consequently, if *G and E are conditionally independent given* $D = 0$, an assumption that will be labeled Assumption (0), then $\theta \equiv \log(OR_0) = 0$. This means that under Assumption (0), estimating β would be the same as estimating $\beta_0 \equiv \log(OR_1)$, the log odds ratio of *the diseased population* (a.k.a., the "cases"). This suggests the use of methods from retrospective sampling design, which typically is more effective, in terms of sampling cost and/or statistical efficiency, than prospective designs, especially when the disease prevalence is low; see Section 1 of M&C and the references therein.

2.2. **A Potentially Misleading Insight.** There is, of course, no free lunch. From a statistical inference perspective, the increased precision comes at the expense of possible serious bias when the assumption $\theta = 0$ fails. Incidently, in M&C, following an argument in Schmidt and Schaid (1999), this assumption is made as a consequence of another two assumptions: Assumption (1) G and E are independent in the general population (that is, not conditioning on the disease status) and Assumption (2) the disease is rare. Whereas these two assumptions do imply Assumption (0) hold *approximately* because when the diseased population is very small, the odds ratio between G and E for the disease-free population can be approximated by that of the general population, these two assumptions were needed by Schmidt and Schaid (1999) apparently because they did not recognize that the second factor in their equation (1) is simply OR_0^{-1}, using our notation above. Consequently, instead of invoking the theoretically more insightful Assumption (0), they had to invoke the assumption that "*the disease risk is small at all levels of both study variables*" ("both study variables" here means the gene variable and environmental variable) in order to justify that the aforementioned second factor is (approximately) 1 (and hence $\theta \approx 0$). This unnecessary assumption apparently was inherited from Piegorsch *et. al.* (1994), who correctly pointed out the usefulness of the case-only studies.

This is a good demonstration of the value of precise theoretical derivation, because identity (2.5) shows clearly that $\beta = \beta_0$ if and only if $\theta = 0$, a condition that has little to do with the disease being rare. That is, it would be quite unfortunate if the quote above is interpreted as declaring that the so-called "case-only" approach for estimating ψ is useful only for rare diseases. Indeed, the only rationale for relying on Assumption (1) (and hence Assumption (2)) I can think of is if checking the independence of G and E in the general population is easier than in the disease-free population. This could be a case when we do not trust the disease diagnosis, because the former does not require knowing each individual's disease status. But this advantage seems rather inconsequential in gene-environment interaction

studies—if we do not have the disease status or do not trust them, then we have much more to worry about than assessing the independence between G and E.

Indeed, M&C's approach did not actually use Assumption (1) or Assumption (2). Instead, they directly use (2.5) by writing $\beta = \beta(\theta) = \beta_0 - \theta$ and then reexpress (2.5) as

$$\beta(\theta) = \beta(0) - \theta. \tag{2.6}$$

This re-expression allows M&C to invoke a *partially Bayes* approach (Cox, 1975; McCullagh, 1990), which puts a prior on the nuisance parameter θ only. Since $\beta(0) = \log(OR_1)$ is a characteristic of the diseased population (i.e., $D = 1$), its inference does not involve $\theta = \log(OR_0)$, which is a characteristic of the disease-free population (i.e, $D = 0$). This separation allows M&C to first infer $\beta(0)$ via maximum likelihood estimation, and once $\beta(0)$ is replaced by its MLE, to infer $\beta = \beta(\theta)$ as a Bayesian inference problem of a *function* of the nuisance parameter θ. This is the essence of M&C's method, though their derivation contains a couple of theoretical complications that do not seem necessary (see Section 3.4).

Of course, for a pure Bayesian, such a hybrid "two-stage" method is neither necessary nor justifiable. However, as I argued in Meng (1994) in the context of a posterior predictive p-value (which is a posterior mean of a classic p-value as a function of a nuisance parameter under a prior on the nuisance parameter only, and hence a squarely partially Bayes entity), the value of such partially Bayesian methods should not be underestimated. Minimally, they allow some Bayesian perks to be enjoyed by those who do not wish to join the full B-club. For example, in the current setting, it allows the use of the prior knowledge/belief that the dependence between G and E is weak in the disease-free population. To see more clearly the pros and cons of the partially Bayes framework, the next section will examine it in detail—and compare it with the fully Bayes approach— in the context of a normal regression model with one predictor.

3. Understanding Partially Bayes Methods

3.1. **A Partially Bayes Approach for Bivariate Normal.** M&C presented their general approach via a heuristic argument, which essentially amounts to assuming normality whenever needed, with variances treated as known. To avoid the distractions of the heuristics, which cannot be made precise in general because a Taylor expansion was invoked for approximating *a prior distribution*, let us assume directly that we have an i.i.d. sample $\{y_1, \ldots, y_n\}$ from the following bivariate normal model:

$$Y = \begin{pmatrix} X \\ Z \end{pmatrix} \sim N\left(\begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right), \tag{3.1}$$

where ρ is a known constant. Our interest here is to estimate β, with α being treated as a nuisance parameter.

As is well-known, without any prior knowledge, the MLE of β is $\hat{\beta}^{\rm MLE} = \bar{Z}_n = \sum_i Z_i/n$, and the MLE for α is $\hat{\alpha}^{\rm MLE} = \bar{X}_n = \sum_i X_i/n$. On the other hand, if we happen to know α, then the MLE of β is the regression estimator

$$\hat{\beta}(\alpha) = \bar{Z}_n + \rho(\alpha - \bar{X}_n). \tag{3.2}$$

Note that this definition of the $\hat{\beta}(\alpha)$ function allows us to reexpress (3.2) as

$$\hat{\beta}(\alpha) = \hat{\beta}(0) + \rho\alpha. \tag{3.3}$$

Clearly, given the data, the only unknown quantity in $\hat{\beta}(\alpha)$ is α (recall ρ is known here). Suppose we are willing to put down the prior $N(0, \tau^2)$ for α, where τ^2 represents our prior belief about how close α is to zero. Under this prior, the partially Bayes approach combines it with the (partial) likelihood from $\bar{X}_n|\alpha \sim N(\alpha, n^{-1})$ to arrive at the usual "shrinkage" posterior (e.g., Efron and Morris, 1973)

$$\alpha|\bar{X}_n \sim N\left(w_\tau \bar{X}_n,\ (n+\tau^{-2})^{-1}\right), \tag{3.4}$$

where $w_\tau = n/(n+\tau^{-2})$. Given this posterior of α, we can infer any of its functions, such as $\hat{\beta}(\alpha)$ of (3.2). In particular, M&C suggested to replace the α in (3.3) by the posterior mean in (3.4), which results in, after noting from (3.2) that $\hat{\beta}^{\rm MLE} - \hat{\beta}(0) = \rho\bar{X}_n$, their estimator

$$\hat{\beta}^{\rm part}_\tau \equiv \hat{\beta}(0) + w_\tau(\hat{\beta}^{\rm MLE} - \hat{\beta}(0)) = w_\tau\hat{\beta}^{\rm MLE} + (1-w_\tau)\hat{\beta}(0). \tag{3.5}$$

Therefore, for a given hyperparameter τ^2, the *partially* Bayes estimator $\hat{\beta}^{\rm part}_\tau$ for β is a compromise between the MLE under the restrictive model with $\alpha = 0$, $\hat{\beta}(0)$, and the MLE of β under the full model, $\hat{\beta}^{\rm MLE} = \hat{\beta}(\hat{\alpha}^{\rm MLE})$, as weighted by the usual shrinkage factor w_τ.

3.2. **Comparing Full Bayes with Simultaneous Partially Bayes.** Before we discuss the issue of choosing τ^2, it is informative to compare the above partially Bayes solution to a full Bayes one, which of course would require a joint prior for $\{\alpha, \beta\}$. To simplify the algebra, let us assume that *a priori* β and α are independent, and $\beta \sim N(0, \varsigma^2)$, with ς^2 given. Under this setup, the joint posterior of $\{\alpha, \beta\}$ obviously follows the usual regression calculation:

$$\begin{pmatrix} \alpha \\ \beta \end{pmatrix} \Bigg| \begin{pmatrix} \bar{X}_n \\ \bar{Z}_n \end{pmatrix} \sim N\left((\Omega^{-1} + \Sigma_n^{-1})^{-1}\Sigma_n^{-1} \begin{pmatrix} \bar{X}_n \\ \bar{Z}_n \end{pmatrix},\ (\Omega^{-1} + \Sigma_n^{-1})^{-1} \right), \tag{3.6}$$

where $\Sigma_n = \frac{1}{n}\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ and $\Omega = \begin{pmatrix} \tau^2 & 0 \\ 0 & \varsigma^2 \end{pmatrix}$.

To understand the difference between the full Bayes and the partially Bayes methods, however, it is more informative to invoke the following indirect derivation. Following the partially Bayes argument, we first treat α as a known constant. Then it is easy to see that

the $\hat{\beta}(\alpha)$ of (3.2) is a sufficient statistic for β and

$$\hat{\beta}(\alpha)|\beta \sim N\left(\beta,\ n_\rho^{-1}\right), \tag{3.7}$$

where $n_\rho = n/(1-\rho^2)$ (larger than n due to gained information via regressing on α). Together with the prior $\beta \sim N(0, \varsigma^2)$, the identical calculation for (3.4) yields

$$\beta|\hat{\beta}(\alpha) \sim N\left(w_{\varsigma,\rho}\hat{\beta}(\alpha),\ \left(\varsigma^{-2}+n_\rho\right)^{-1}\right), \tag{3.8}$$

where $w_{\varsigma,\rho} = n_\rho/(n_\rho+\varsigma^{-2})$. The sufficiency of $\hat{\beta}(\alpha)$ for β for given α implies that $\mathrm{E}[\beta|\bar{X}_n, \bar{Z}_n, \alpha] = \mathrm{E}[\beta|\hat{\beta}(\alpha)]$, and hence, by iterated expectations and (3.3),

$$\mathrm{E}[\beta|\bar{X}_n, \bar{Z}_n] = w_{\varsigma,\rho}\mathrm{E}\left[\hat{\beta}(\alpha)|\bar{X}_n, \bar{Z}_n\right] = w_{\varsigma,\rho}\left(\hat{\beta}(0) + \rho\mathrm{E}[\alpha|\bar{X}_n, \bar{Z}_n]\right). \tag{3.9}$$

At first glance, (3.9) achieves nothing because it simply transfers the calculation of $\mathrm{E}[\beta|\bar{X}_n, \bar{Z}_n]$ to the equally difficult (or easy) problem of calculating $\mathrm{E}[\alpha|\bar{X}_n, \bar{Z}_n]$. But this observation should also remind us that we can simply switch β with α (and accordingly ς with τ and $\bar{Z}_n$ with $\bar{X}_n$) to arrive at its dual identity

$$\mathrm{E}[\alpha|\bar{X}_n, \bar{Z}_n] = w_{\tau,\rho}\left(\hat{\alpha}(0) + \rho\mathrm{E}[\beta|\bar{X}_n, \bar{Z}_n]\right), \tag{3.10}$$

where $w_{\tau,\rho} = n_\rho/(n_\rho + \tau^{-2})$, and $\hat{\alpha}(\beta) = \bar{X}_n + \rho(\beta - \bar{Z}_n)$.

It is now a simple matter to solve (3.9)-(3.10) to arrive at

$$\hat{\beta}^{\mathrm{full}}_{\tau,\varsigma} \equiv \mathrm{E}[\beta|\bar{X}_n, \bar{Z}_n] = \frac{w_{\varsigma,\rho}[\hat{\beta}(0) + w_{\tau,\rho}\rho\hat{\alpha}(0)]}{1-\rho^2 w_{\varsigma,\rho}w_{\tau,\rho}}, \tag{3.11}$$

where the superscript "full" highlights the fact that it is identical to the fully Bayes answer from (3.6), as can be verified directly.

To express (3.11) in a more insightful way, we can use the fact that $\rho\hat{\alpha}(0) = \rho\bar{X}_n - \rho^2\bar{Z}_n = \hat{\beta}^{\mathrm{MLE}} - \hat{\beta}(0) - \rho^2\hat{\beta}^{\mathrm{MLE}}$ to arrive at

$$\hat{\beta}^{\mathrm{full}}_{\tau,\varsigma} = w_{\varsigma\tau,\rho}\hat{\beta}^{\mathrm{part}}_{\tau}, \tag{3.12}$$

where $\hat{\beta}^{\mathrm{part}}_{\tau}$ is from (3.5),

$$w_{\varsigma\tau,\rho} = \frac{w_{\varsigma,\rho} - \rho^2 w_{\varsigma,\rho}w_{\tau,\rho}}{1-\rho^2 w_{\varsigma,\rho}w_{\tau,\rho}} = \frac{n_{\rho,\tau}}{n_{\rho,\tau}+\varsigma^{-2}}, \tag{3.13}$$

and

$$n_{\rho,\tau} = \frac{n}{1-\rho^2(1-w_\tau)}, \tag{3.14}$$

with $w_\tau = n/(n+\tau^{-2})$, as in (3.4). This means that, as far as point estimator goes, the full Bayes estimator $\hat{\beta}^{\mathrm{full}}_{\varsigma,\tau}$ can be viewed as a further shrinkage of the partially Bayes estimator $\hat{\beta}^{\mathrm{part}}_{\tau}$ towards zero. In particular, we notice that regardless of the value of ρ, $\lim_{\tau\to\infty} n_{\rho,\tau} = n$ and hence $\lim_{\tau\to\infty} w_{\varsigma\tau,\rho} = n/(n+\varsigma^{-2}) \equiv w_\varsigma$. This means that when $\tau = \infty$, the fully Bayes estimator for β would reduce to the usual shrinkage estimator $w_\varsigma\bar{Z}_n$ based on the $\bar{Z}_n$ margin

alone. Intuitively, when $\tau = \infty$, there is no information to borrow from the prior knowledge of α for estimating β even if $\rho \neq 0$, and hence all the information is in the $\{Z, \beta\}$ margin.

3.3. **Sequential Partially Bayes Methods and Partial Shrinkage.** Intuitively, the fully Bayes method takes into account the prior information $\beta \sim N(0, \varsigma^2)$, which was not used by $\hat{\beta}^{\text{part}}_{\tau}$. The above derivation shows how one can achieve the full Bayes efficiency by performing two partially Bayes steps *simultaneously*, namely, by solving (3.9)-(3.10) as a pair, which is a special case of applying the "self-consistency" principle (Meng, Lee and Li, 2009). In contrast, if we have carried out the partially Bayes method *sequentially*, that is, in two stages, then the full efficiency is not guaranteed even if priors for both β and α are used.

To see this more clearly, suppose we follow M&C's general argument and first treat the nuisance parameter α as known. Then conditioning on α, but taking into account the prior information on β via $N(0, \varsigma^2)$, our Bayes estimator for β is as given in (3.8),

$$\hat{\beta}_{\varsigma}(\alpha) \equiv \mathrm{E}[\beta|\hat{\beta}(\alpha)] = w_{\varsigma,\rho}\hat{\beta}(\alpha) = w_{\varsigma,\rho}\left(\hat{\beta}(0) + \rho\alpha\right). \tag{3.15}$$

Now, unlike in the simultaneous method described above, if we follow the general argument as in M&C to treat $\hat{\beta}_{\varsigma}(\alpha)$ as the objective of our inference, we would replace α in the right most side of (3.15) by its (partial) posterior mean $E(\alpha|\bar{X}_n) = w_{\tau}\bar{X}_n$. This substitution then will lead to the sequential partially Bayes estimator

$$\hat{\beta}^{\text{seqe}}_{\varsigma\tau,\rho} = w_{\varsigma,\rho}\left(\hat{\beta}(0) + w_{\tau}(\rho\bar{X}_n)\right) = w_{\varsigma,\rho}\beta^{\text{part}}_{\tau}. \tag{3.16}$$

Comparing (3.16) to (3.12), we see that although both of them are further shrinkages of the same $\beta^{\text{part}}_{\tau}$ of (3.5) and both shrinkage factors depend on ς, $\hat{\beta}^{\text{seqe}}_{\varsigma\tau,\rho}$ shrinks less towards zero than the full Bayes estimator $\hat{\beta}^{\text{full}}_{\varsigma\tau,\rho}$. This is because

$$w_{\varsigma\tau,\rho} \equiv \frac{n_{\rho,\tau}}{n_{\rho,\tau} + \varsigma^{-2}} < \frac{n_{\rho}}{n_{\rho} + \varsigma^{-2}} \equiv w_{\varsigma,\rho}, \tag{3.17}$$

provided that

$$n_{\rho,\tau} \equiv \frac{n}{1 - \rho^2(1 - w_{\tau})} < \frac{n}{1 - \rho^2} \equiv n_{\rho}, \tag{3.18}$$

which is the case as long as $\rho \neq 0$ because $w_{\tau} = n/(n + \tau^{-2}) > 0$.

Intuitively, $\hat{\beta}^{\text{seqe}}_{\varsigma\tau,\rho}$ only achieves *partial shrinkage* compared to $\hat{\beta}^{\text{full}}_{\varsigma\tau,\rho}$ because it fails to take into account the prior information $\beta \sim N(0, \varsigma^2)$ *when estimating* α. Even when β and α are *a priori* independent, as long as X and Z are correlated conditional on the model parameter, X and Z are correlated with respect to the *predictive distribution*, that is, with the model parameter integrated out according to the prior. In our current setting, the correlation between (X, Z) with respect to their predictive distribution is $\rho_{\tau,\varsigma} = \rho/\sqrt{(1 + \tau^2)(1 + \varsigma^2)}$. As long as $\rho_{\tau,\varsigma} \neq 0$, the information on the marginal distribution of Z via α will have an

impact on the marginal distribution of X (and vice versa). However, as $\varsigma \to \infty$, $\rho_{\tau,\varsigma} \to 0$ and hence this impact disappears as the prior information for β becomes diffuse. This can also be seen from (3.17), which becomes equality and hence $\hat{\beta}^{\text{seqe}}_{\varsigma\tau,\rho} = \hat{\beta}^{\text{fall}}_{\varsigma\tau,\rho}$ whenever $\varsigma = \infty$, regardless of the value of ρ or τ.

3.4. **Completing M&C's Argument.** For a given value of τ^2, M&C's general approach is essentially an approximate version of what is presented in Section 3.1, resulting in the same partially Bayes estimator general expression as in (3.5). I say essentially because M&C apparently introduced a technical complication that is not necessary. The derivation in Section 3.1 relies on treating $\hat{\beta}(\alpha)$ of (3.2) as our *estimand*. Note $\hat{\beta}(\alpha)$ actually depend on data, but from the Bayesian perspective, treating it as a known function of the unknown α only presents no conceptual or technical complication. However, M&C introduced $\beta(\theta)$ (using their generic notation θ, which is the same as α for the bivariate normal example), the limit of $\hat{\beta}(\theta)$, as the data-free estimand, and then derive a partially Bayes estimator for $\beta(\theta)$ via the delta method $\beta(\theta) - \beta(0) \approx \beta'(0)\theta$ and the (partially Bayes) posterior on θ.

In the example of the gene-environment interaction, this definition of $\beta(\theta)$ worked well, because (2.6) holds for both the population version and sample version. However, for the bivariate normal example, although the sample version $\hat{\beta}(\alpha)$ of (3.2) is a linear function of α, the limit version, according to M&C's definition, would be a constant function because $\beta(\alpha) = \beta$ for all α and hence $\beta'(0) = 0$. Consequently, in general, the aforementioned delta method can be meaningless. Fortunately, this complication is really unnecessary, as we can work directly with $\hat{\beta}(\alpha)$ as the estimand for the partially Bayes method.

Another unnecessary complication is in M&C's treatment of estimating the prior variance as a hyperparameter. Given the prior $\theta \sim N(0, \tau^2)$, M&C first approximated the prior for $\phi \equiv \beta(\theta)$ by $N(\phi_0, \tau^2_\phi)$, where $\phi_0 = \beta(0)$ and $\tau^2_\phi = [\beta'(0)]^2\tau^2$. (Perhaps this is where M&C felt the need to introduce the population version $\beta(\theta)$ because it might seem odd to put a prior on a data-dependent quantity $\hat{\beta}(\theta)$; but there is actually nothing incoherent in the partially Bayes framework for the latter operation). To estimate τ_ϕ, M&C invoked an empirical Bayes argument, which estimates the hyperparameter τ^2 by $\max\{\hat{\theta}^2 - \hat{v}^2, 0\}$ when the approximation $\hat{v}^{-1}(\hat{\theta} - \theta)|\theta \sim N(0, 1)$ holds for some statistic v. A critical ingredient of M&C's proposal is to use $\hat{\theta}^2$ as a *conservative* estimate of τ^2, which then leads to a conservative estimator of the corresponding hyperparameter τ^2_ϕ as $\hat{\tau}^2_\phi = [\hat{\beta}'(0)]^2\hat{\theta}^2$. Substituting this estimator for the hyperparameter in a general version of (3.5) leads to M&C's general proposal. But $\hat{\beta}'(0)\hat{\theta}$ is nothing but the first-term Taylor expansion of $\hat{\beta}(\hat{\theta}) - \hat{\beta}(0) = \hat{\beta} - \hat{\beta}_0$ (though note the hidden assumption that $\hat{\beta}(\hat{\theta}) = \hat{\beta}$). This suggests that we can bypass the calculation of $\hat{\beta}'(0)$ and directly use $(\hat{\beta} - \hat{\beta}_0)^2$ as a conservative estimator of τ^2_ϕ. Indeed, with this modification, M&C's proposal has the simpler expression as given in the abstract.

What is necessary is that once the hyperparameter is estimated from the data, the operating characteristics of the resulting estimator must be evaluated specifically according to the estimation method used. That is, we can no longer rely on the established general properties of the (fully) Bayesian estimators to justify their corresponding empirical counterparts. We do tend to believe that such empirical estimators are reasonably accurate in a variety of situations in practice, as demonstrated in M&C via simulations. But the same belief sometimes can get us into deep trouble when we put too much faith on simulations, which are necessarily limited. Indeed, intuitively speaking, the idea that we can achieve a good universal compromise between $\hat{\beta}$ and $\hat{\beta}_0$ only using themselves plus an estimate of $\mathrm{V}(\hat{\beta})$ (see the formula in the Abstract or (4.1) below) is just too good to be true. It is true that when $\hat{\beta}$ is an unbiased estimator of β, $\hat{B} \equiv \hat{\beta}_0 - \hat{\beta}$ provides an unbiased estimator of the bias in $\hat{\beta}_0$. But it would be illogical for us to worry about $\hat{\beta}$ having too large a variance—and hence the need to seek a reduction by bringing in a more efficient estimator $\hat{\beta}_0$—but not to worry about the large variability in $\hat{B}$, which depends on $\hat{\beta}$ critically. How can we be sure that the large error in the estimated weight $\hat{w}_{\tau_\phi} = w_{\hat{\tau}_\phi}$, which in turn depends critically on $\hat{B}$, would not offset the gain in mean-squared error due to the (correct) weighting via w_{τ_ϕ}?

Indeed, we are not sure at all, as demonstrated in the following Ph.D. qualifying exam problem. (Again, both Section 4 and Section 5 are reproduced in verbatim, other than correcting a few typographical errors.)

4. Learning through Exam: The Actual Qualifying Exam Problem

During a recent departmental seminar, our speaker made an assertion along the following lines: "*I have two estimators, $\hat{\beta}$ and $\hat{\beta}_0$ for the same parameter β. The former is more robust because it is derived under a more general model, and the second is more efficient because it is obtained assuming a more restrictive model. The following is a compromise between the two:*

$$\hat{\beta}_c = \frac{(\hat{\beta} - \hat{\beta}_0)^2}{\hat{\mathrm{V}}(\hat{\beta}) + (\hat{\beta} - \hat{\beta}_0)^2}\hat{\beta} + \frac{\hat{\mathrm{V}}(\hat{\beta})}{\hat{\mathrm{V}}(\hat{\beta}) + (\hat{\beta} - \hat{\beta}_0)^2}\hat{\beta}_0, \tag{4.1}$$

where $\hat{\mathrm{V}}(\hat{\beta})$ is a consistent estimate of the variance of $\hat{\beta}$. This should work better because when the more restrictive model is true, $\hat{\beta}_c$ tends to give more weight to the more efficient $\hat{\beta}_0$, and at the same time, $\hat{\beta}_c$ remains consistent because asymptotically it is the same as $\hat{\beta}$."

As some of you might recall, I was both intrigued by and skeptical about this assertion. This problem asks you to help me to understand and investigate the speaker's assertion. To do so, let's first formalize the meaning of a general model and a more restrictive one.

Suppose we have i.i.d. data $\vec{Y} = \{y_1, \ldots, y_n\}$ from a model $f(y|\theta)$, where $\theta = \{\alpha, \beta\}$, both of which are scalar quantities, with β the parameter of interest, α the nuisance parameter,

and the *meaning* of β does not depend on the value of α. Suppose the restrictive model takes the form $f_0(y|\beta) = f(y|\alpha = 0, \beta)$, i.e., under the restrictive model we know the true value of α is zero. Let $\hat{\theta} = \{\hat{\alpha}, \hat{\beta}\}$ be a consistent estimator of θ under the general model $f(y|\theta)$, and let $\hat{\beta}_0$ be a consistent estimator of β_0, which is guaranteed to be β only when the restrictive model $f_0(y|\beta)$ holds. We further assume all the necessary regularity conditions to guarantee their *joint* asymptotic normality, that is,

$$\sqrt{n}\left[\begin{pmatrix}\hat{\theta}\\ \hat{\beta}_0\end{pmatrix} - \begin{pmatrix}\theta\\ \beta_0\end{pmatrix}\right] \to N\left(\begin{pmatrix}\mathbf{0}\\ 0\end{pmatrix}, \Sigma = \begin{pmatrix}\Sigma_\theta & C^T\\ C & \sigma^2_{\beta_0}\end{pmatrix}\right). \tag{4.2}$$

For simplicity of derivation, we will assume $\Sigma \geq 0$ (i.e., a semi-positive definite matrix) is *known*, and the convergence in (4.2) is in the L^2 sense (i.e., $X_n \to X$ means $\lim_{n\to\infty} E||X_n - X||^2 = 0$).

(A) The speaker clearly was considering a variance-bias trade-off, assuming that $\hat{\beta}_0$ is more efficient than $\hat{\beta}$ when the more restrictive model is true. Under the setup above, prove this is true asymptotically when $\hat{\theta}$ and $\hat{\beta}_0$ are maximum likelihood estimators (MLE, as in the superscript below) under the general model and restrictive model respectively and when we use the Mean-Squared Error (MSE) criterion (we can then assume Σ_θ and σ^2_β are given by the inverse of the corresponding Fisher information). That is, prove that if the restrictive model holds, the (asymptotic) relative efficiency (RE) of $\hat{\beta}_0$ to that of $\hat{\beta}$ is no less than 1:

$$RE \equiv \lim_{n\to\infty} \frac{E[\hat{\beta}^{\text{MLE}} - \beta]^2}{E[\hat{\beta}_0^{\text{MLE}} - \beta]^2} \geq 1, \tag{4.3}$$

and give a necessary and sufficient condition for equality to hold. Provide an intuitive statistical explanation of this result, including the condition for equality to hold.

(B) Give a counterexample to show that (4.3) no longer holds if we drop the MLE requirement. What is the key implication of this result on the speaker's desire to improve $\hat{\beta}$ via $\hat{\beta}_0$?

(C) Since we assume Σ is known, we can replace $\hat{\text{V}}(\hat{\beta})$ in (4.1) by σ^2_β/n, where σ^2_β is an appropriate entry of Σ_θ. We can therefore re-express (4.1) as

$$\hat{\beta}_c = (1 - W_n)\hat{\beta} + W_n\hat{\beta}_0, \quad \text{where} \quad W_n = \frac{\sigma^2_\beta}{\sigma^2_\beta + n(\hat{\beta} - \hat{\beta}_0)^2}. \tag{4.4}$$

Prove that, under our basic setup (4.2), $\lim_{n\to\infty} E(W_n) = 0$ if and only if $\beta \neq \beta_0$.

(D) Using Part (C) to prove that whenever $\beta \neq \beta_0$,

$$\lim_{n\to\infty} \frac{E[\hat{\beta}_c - \beta]^2}{E[\hat{\beta} - \beta]^2} = 1. \tag{4.5}$$

Which aspect of the speaker's assertion this result helps to establish?

(E) To show that the condition $\beta \neq \beta_0$ cannot be dropped in Part (D), let us consider that our data $\{y_1, \ldots, y_n\}$ are i.i.d. samples from the following bivariate normal model:

$$Y = \begin{pmatrix} X \\ Z \end{pmatrix} \sim N\left(\begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right), \tag{4.6}$$

where ρ is *known*. Show that under this model, *when we use MLEs for* $\hat{\beta}$ and $\hat{\beta}_0$, $\sqrt{n}(\hat{\beta}_c - \beta)$ has exactly the same distribution as

$$\xi = Z_0 - \rho(X_0 + \sqrt{n}\alpha)\tilde{W}_n = (Z_0 - \rho X_0) + \rho[(1 - \tilde{W}_n)X_0 - \tilde{W}_n\sqrt{n}\alpha], \tag{4.7}$$

where $(X_0, Z_0)^\top$ has the same distribution as in (4.6) but with both α and β set to zero, and

$$\tilde{W}_n \equiv \tilde{W}_n(\rho, \alpha) = \frac{1}{1 + \rho^2(X_0 + \sqrt{n}\alpha)^2}.$$

Use the right-most expression in (4.7) to then show that

$$nE[\hat{\beta}_c - \beta]^2 = 1 - \rho^2 + \rho^2 G_n(\rho, \alpha), \tag{4.8}$$

where

$$G_n(\rho, \alpha) = E[(1 - \tilde{W}_n(\rho, \alpha))X_0 - \tilde{W}_n(\rho, \alpha)\sqrt{n}\alpha]^2. \tag{4.9}$$

(F) Continuing the setting of Part (E), use (4.8) to prove that when $\alpha = 0$, for all n,

$$E[\hat{\beta}_0^{\text{MLE}} - \beta]^2 < E[\hat{\beta}_c - \beta]^2 < E[\hat{\beta}^{\text{MLE}} - \beta]^2, \tag{4.10}$$

as long as $\rho \neq 0$. Why does this result imply that $\beta \neq \beta_0$ cannot be dropped in Part (D)? What happens when $\rho = 0$?

(G) Still under the setting of Parts (E) and (F), verify that $G_n(0, \alpha) = n\alpha^2$, and then use this fact to prove that as long as $n\alpha^2 > 1$, there exists a $\rho^*_{n,\alpha} > 0$ such that for all $0 < |\rho| < \rho^*_{n,\alpha}$,

$$nE[\hat{\beta}_c - \beta]^2 > 1 = nE[\hat{\beta}^{\text{MLE}} - \beta]^2. \tag{4.11}$$

Does this contradict Part (D)? Why or why not?

(H) What do all the results above tell you about the speaker's proposed estimator $\hat{\beta}_c$? Does it have the desired property as the speaker hoped for? Would you or when would you recommend it? Give reasons for any conclusion you draw.

5. Interweaving Research and Pedagogy: The Actual Annotated Solution

(A) *This part tests a student's understanding of the most basic theory of likelihood inference, especially the calculation of Fisher information, and the fact that the MLE approach is efficient/coherent in the sense that when more assumptions are made its efficiency is guaranteed to be non-decreasing.*

The result (4.3) is easily established using the fact that if we write the expected Fisher information under the general model (with $n = 1$) as

$$I(\theta) = \begin{pmatrix} i_{\alpha\alpha} & i_{\alpha\beta} \\ i_{\alpha\beta} & i_{\beta\beta} \end{pmatrix}, \quad \text{and notationally} \quad I^{-1}(\theta) = \begin{pmatrix} i^{\alpha\alpha} & i^{\alpha\beta} \\ i^{\alpha\beta} & i^{\beta\beta} \end{pmatrix}, \tag{5.1}$$

then $i^{\beta\beta} = [i_{\beta\beta} - i_{\alpha\beta}^2 i_{\alpha\alpha}^{-1}]^{-1}$. The Fisher information under the restrictive model of course is given by $i_{\beta\beta}$ with $\alpha = 0$. Consequently, under our basic setup, when $\alpha = 0$,

$$RE = \frac{i^{\beta\beta}}{i_{\beta\beta}^{-1}} = \left[1 - \frac{i_{\alpha\beta}^2}{i_{\alpha\alpha} i_{\beta\beta}}\right]^{-1} \geq 1, \tag{5.2}$$

where equality holds if and only if $i_{\alpha\beta} = 0$ *when* $\alpha = 0$, that is, when β and α are *orthogonal* (asymptotically) under the restrictive model. Intuitively, the gain of efficiency of $\hat{\beta}_0^{\text{MLE}}$ over $\hat{\beta}^{\text{MLE}}$ is due to $\hat{\beta}^{\text{MLE}}$'s *covariance adjustment* via $\hat{\alpha}^{\text{MLE}} - \alpha$ when $\alpha = 0$. However, this adjustment can take place if and only if $\hat{\beta}^{\text{MLE}}$ is correlated with $\hat{\alpha}^{\text{MLE}}$ when $\alpha = 0$, which is the same as $i_{\alpha\beta} \neq 0$.

(B) *This part in a sense is completely trivial, but it carries an important message. That is, the common notation/intuition that "the more information (e.g., via model assumptions) or the more data, the more efficiency" can be true only when the procedure we use processes information/data in an efficient way (e.g., as with MLE).*

There are many trivial and "absurd" counterexamples. For example, in Part (A), if we use the same MLE under the general model, but only use 1/2 our samples when applying the MLE under the restrictive model, then the RE ratio in (5.2) obviously will be *deflated* by a factor 2, and hence it can easily be made to be less than one.

[A much less trivial or absurd example is when we want to estimate the correlation parameter ρ with bivariate normal data $\{(x_i, y_i), i = 1, \ldots, n\}$. Without making any restriction on other model parameters, we know the sample correlation is asymptotically efficient with asymptotic variance $(1 - \rho^2)^2/n$ (see Ferguson, 1996, Chapter 8). Now suppose our restrictive model is that both X and Y have mean zero and variance 1. The Fisher information for this restrictive model is $(1 + \rho^2)/(1 - \rho^2)^2$, therefore $RE = 1 + \rho^2 \geq 1$, which confirms Part (A). However, since $E(XY) = \rho$ under the restrictive model, someone might be tempted to use the obvious moment estimator $\hat{r}_n = \sum_i x_i y_i/n$ for ρ. But one can easily calculate that the variance (and hence MSE) of $\hat{r}_n$ is $(1 + \rho^2)/n$ for any n. Consequently, the RE of $\hat{r}_n$ compared to the sample correlation is (asymptotically) $(1 - \rho^2)^2/(1 + \rho^2)$, which is always less than one and actually approaches zero when ρ^2 approaches 1. So the additional assumption can hurt tremendously if one is not using an efficient estimator! (Students may recall that my qualifying exam problem from a previous year was about this problem.) Moments estimators are used frequently in practice because of their simplicity and robustness

(to model assumptions), but this example shows that one must exercise great caution when using moment estimators, especially when making claims about their relative efficiency when adding assumptions or data.]

(C) *Intuitively this result is obvious, because when $\beta \neq \beta_0$, the denominator in W_n can be made arbitrarily large as n increases, and hence its expectation should go to zero. But this part tests a student's ability to make such "hand-waving" argument rigorous without invoking excessive technical details, which is an essential skill for theoretical research.*

Let $\Delta_n = \sqrt{n}(\hat{\beta}-\hat{\beta}_0-\delta)$, where $\delta = \beta-\beta_0$. Then by (4.2), Δ_n converges in L^2 to $N(0,\tau^2)$, where $\tau^2 = a^\top \Sigma a$, with $a = (0,1,-1)^\top$. Therefore, there exists a n_0 such that for all $n \geq n_0$, $\mathrm{V}(\Delta_n) \leq 2\tau^2$. Consequently, for any $\epsilon > 0$, if we let $M_\epsilon = \sqrt{2\tau^2/\epsilon}$, and $A_n = \{|\Delta_n| \geq M_\epsilon\}$, then by Chebyshev's inequality, we have

$$\Pr(A_n) = \Pr(|\Delta_n| \geq M_\epsilon) \leq \frac{\mathrm{V}(\Delta_n)}{M_\epsilon^2} \leq \epsilon. \tag{5.3}$$

Now if $\delta \neq 0$, then as long as $n \geq M_\epsilon^2/\delta^2$, we have, noting $0 < W_n = \frac{\sigma_\beta^2}{\sigma_\beta^2+(\Delta_n+\sqrt{n}\delta)^2} \leq 1$,

$$0 \leq E(W_n) = E(W_n \mathbf{1}_{A_n}) + E(W_n \mathbf{1}_{A_n^c}) \leq \Pr(A_n) + \frac{\sigma_\beta^2}{\sigma_\beta^2 + (\sqrt{n}|\delta| - M_\epsilon)^2}, \tag{5.4}$$

where in deriving the last inequality we have used the fact that $(u+v)^2 \geq (|u|-|v|)^2$. That $E(W_n) \to 0$ then follows from (5.3) and (5.4) by first letting $n \to \infty$ in (5.4), and then letting $\epsilon \to 0$ in (5.3).

To prove the converse, we note that when $\delta = 0$, $W_n = \frac{\sigma_\beta^2}{\sigma_\beta^2+\Delta_n^2}$. Therefore, by (Jensen's) inequality $E(X^{-1}) \geq [E(X)]^{-1}$, we have

$$E(W_n) \geq \frac{\sigma_\beta^2}{\sigma_\beta^2 + E(\Delta_n^2)} \to \frac{\sigma_\beta^2}{\sigma_\beta^2 + \tau^2} > 0.$$

(D) *This part is rather straightforward, as long as the student is familiar with the Cauchy-Schwarz inequality (which is a must!)*

From (4.4), we have $\sqrt{n}(\hat{\beta}_c-\beta) = \sqrt{n}(\hat{\beta}-\beta) - W_n D_n$, where $D_n = \sqrt{n}(\hat{\beta}-\hat{\beta}_0)$. It follows then

$$nE(\hat{\beta}_c-\beta)^2 = nE(\hat{\beta}-\beta)^2 + E(W_n^2 D_n^2) - 2E[\sqrt{n}(\hat{\beta}-\beta)(W_n D_n)]. \tag{5.5}$$

Under our assumptions, the first term on the right hand side of (5.5) converges to $\sigma_\beta^2 > 0$, so (4.5) follows if we can establish that the second term on the right hand side of (5.5) converges to zero. This is because, by the Cauchy-Schwarz inequality, the third term on the right hand side of (5.5) is bounded above in magnitude by $2\sqrt{nE(\hat{\beta}-\beta)^2 E(W_n^2 D_n^2)}$, and

hence it must then converge to zero as well if the second term does so. But by the definition of W_n in (4.4),

$$E(W_n^2 D_n^2) = E\left[W_n \frac{\sigma_\beta^2 D_n^2}{\sigma_\beta^2 + D_n^2}\right] \le \sigma_\beta^2 E(W_n), \tag{5.6}$$

which converges to zero by Part (C) when $\delta = \beta - \beta_0 \neq 0$. The implication of this result is that the speaker's assertion that $\hat{\beta}_c$ is asymptotically the sane as $\hat{\beta}$ is correct, as long as $\beta \neq \beta_0$. [Note there is a subtle difference between $\beta = \beta_0$ and $\alpha = 0$. The latter implies the former, but the reverse may not be true because one can always choose $\hat{\beta}_0$ to be $\hat{\beta}$ even if the restrictive model is not true.]

(E) *This part tests a student's understanding of multi-variate normal models and the basic regression concepts, with which one can complete this part without any tedious algebra.*

The most important first step is to recoganize/realize that under the general model, $\hat{\beta}^{\text{MLE}} = \bar{Z}_n$, and under the restrictive model, $\hat{\beta}_0^{\text{MLE}} = \bar{Z}_n - \rho \bar{X}_n$, where $\bar{X}_n$ and $\bar{Z}_n$ are the sample averages; hence $D_n = \rho\sqrt{n}\bar{X}_n$. The first expression in (4.7) then follows from (4.4) when we re-write it as $\hat{\beta}_c = \bar{Z}_n - W_n(\rho \bar{X}_n)$ and let $X_0 = \sqrt{n}(\bar{X}_n - \alpha)$ and $Z_0 = \sqrt{n}(\bar{Z}_n - \beta)$, and the fact that (X_0, Z_0) has the same bivariate normal distribution as in (4.6) but with zero means. The second expression is there to hint the independence of the two terms, because the first term $(Z_0 - \rho X_0)$ is the residual after regressing out X_0, and the second term is a function of X_0 only. With this observation, (4.8) follows immediately because the residual variance is $1 - \rho^2$.

(F) *Again, this part does not require any algebra if a student understands the most basic calculations with bivariate normal and regression.* When $\alpha = 0$, $\tilde{W}_n(\rho, 0) = \frac{1}{1+\rho^2 X_0^2}$, and

$$G_n(\rho, 0) = E[X_0(1 - \tilde{W}_n(\rho, 0))]^2 = E\left[X_0^2 \left(\frac{\rho^2 X_0^2}{1 + \rho^2 X_0^2}\right)^2\right] \equiv C_\rho, \tag{5.7}$$

where the constant $C_\rho > 0$ is free of n and it is clearly less than $E(X_0^2) = 1$. Therefore the identity (4.8) immediately leads to $nE[\hat{\beta}_c - \beta]^2 = 1 - (1 - C_\rho)\rho^2$, which is strictly larger than $nE[\hat{\beta}_0^{\text{MLE}} - \beta]^2 = 1 - \rho^2$ and smaller than $nE[\hat{\beta}^{\text{MLE}} - \beta]^2 = 1$, as long as $\rho \neq 0$. Clearly in this case (4.5) of Part (D) will not hold because the ratio there will be $1 - (1 - C_\rho)\rho^2 < 1$, hence the condition $\beta \neq \beta_0$ cannot be dropped in Part (D) – note when $\rho \neq 0$, $\beta \neq \beta_0$ is equivalent to $\alpha \neq 0$.

When $\rho = 0$, $\hat{\beta}^{\text{MLE}} = \hat{\beta}_0^{\text{MLE}}$, and hence regardless of the value of α, Part (D) holds trivially even though the condition $\beta \neq \beta_0$ is violated. This also provides another (trivial) example that $\beta = \beta_0$ does not imply $\alpha = 0$, as we discussed at the end of the solution to Part (D) above.

(G) *This part demonstrates the need of some basic mathematical skills in order to derive important statistical results (that cannot be just "hand-waved"!).*

When $\rho = 0$, $\tilde{W}_n(0,\alpha) = 1$, and hence $G_n(0,\alpha) = n\alpha^2$. From its expression (4.9), the (random) function under expectation is continuous in ρ and bounded above by $X_0^2 + n\alpha^2$, which has the expectation $1+n\alpha^2$. Hence, by the Dominated Convergence Theorem, $G_n(\rho,\alpha)$ is a continuous function of ρ for any given α and n. Consequently, whenever $G_n(0,\alpha) = n\alpha^2 > 1$, there must exist a $\rho^*_{n,\alpha} > 0$, such that for any $|\rho| \le \rho^*_{n,\alpha}$, $G_n(\rho,\alpha) > 1$ as well. It follows then, when $0 < |\rho| \le \rho^*_{n,\alpha}$, from (4.8),

$$nE[\hat{\beta}_c - \beta]^2 = 1 - \rho^2 + \rho^2 G_n(\rho,\alpha) > 1 - \rho^2 + \rho^2 = 1 = nE[\hat{\beta}^{\mathrm{MLE}} - \beta]^2. \tag{5.8}$$

Inequality (5.8), however, does not contradict Part (D) because the choice of $\rho^*_{n,\alpha}$ depends on n, so Part (D) implies that as n increases, $\rho^*_{n,\alpha} \to 0$.

(H) Parts (A) and (B) demonstrate that in order for the proposed estimator (4.1) to achieve the desired compromise, a minimal requirement is that there should be some "efficiency" requirement on the estimation procedures, especially the one under the more restrictive model. Otherwise it would not be wise in general to bring in $\hat{\beta}_0$ to *contaminate* an already more efficient and more robust estimator $\hat{\beta}$.

Parts (C) and (D) proved that under quite mild conditions, the proposed $\hat{\beta}_c$ is equivalent asymptotically to the estimator under the general model, as long as the estimator under the more restrictive model is *asymptotically biased*, that is, as long as $\beta_0 \neq \beta$. So in that sense the speaker's proposal is not harmful but not helpful either asymptotically, and therefore any possible improvement must be a finite-sample one (which apparently is what the speaker intended and indeed the only possible way if one uses MLE to start with).

Parts (E)-(G) give an example to show that when the restrictive model is true, the speaker's proposal can achieve the desired compromise, that is, $\hat{\beta}_c$ beats $\hat{\beta}^{\mathrm{MLE}}$ in terms of MSE for all n, but it is not as good as $\hat{\beta}_0^{\mathrm{MLE}}$. The latter is not surprising at all because in this case $\hat{\beta}_0^{\mathrm{MLE}}$ is the most efficient estimator (asymptotically, but also in finite sample given its asymptotic variance is also the exact variance). However, when the restrictive model is not true, then there is no longer any guarantee that $\hat{\beta}_c$ will dominate $\hat{\beta}$ (indeed this is not possible in general whenever $\hat{\beta}$ is admissible). The result in Part (G) also hinted that in order for $\hat{\beta}_c$ to beat $\hat{\beta}$, the "regression effect" of $\hat{\beta}$ on $\hat{\alpha}$ must be strong enough (e.g., expressed in this case via $|\rho| > \rho^*_{n,\alpha}$) in order to have enough borrowed efficiency from $\hat{\beta}_0$ to make it happen.

In summary, the speaker's proposal can provide the desired compromise when the restricted model is close to being true and the original two estimators are efficient in their own right, but it cannot achieve this unconditionally. In general, it is not clear at all as when one

should use such a procedure, especially when the original two estimators are not efficient to start with.

6. A Piece of Inadmissible Cake?

M&C's $\hat{\beta}_c$ evidently was proposed as an improvement on the original $\hat{\beta}$, with MSE as the intended criterion. Adopting the classic framework of decision theory (Berger, 1985), the hope is that $\hat{\beta}_c$ is *R-better* than $\hat{\beta}$ in terms of the squared loss:

$$R(\hat{\beta};(\alpha,\beta)) = \int_y (\hat{\beta}(y)-\beta)^2 f(y|\alpha,\beta)\mu(dy), \tag{6.1}$$

where $f(y|\beta,\alpha)$ is the sampling density and μ is its corresponding baseline measure. But for $R(\hat{\beta}_c;(\alpha,\beta)) \leq R(\hat{\beta};(\alpha,\beta))$ to hold for all β and α (and with strict inequality for at least one (α,β)) means $\hat{\beta}$ is not admissible under the squared loss. The simple normal problem investigated in Section 4 and Section 5 demonstrates clearly that this would be wishful thinking in general. The question then is how do we quantify the apparently good properties of $\hat{\beta}_c$, as suggested by the empirical evidences in M&C?

If we have a joint prior on $\{\alpha,\beta\}$, of course we can compare the Bayesian risks of $\hat{\beta}_c$ and $\hat{\beta}$. But the partially Bayes approach precisely wants to avoid any prior specification about β. This leads to the notion of *partially Bayes risk*

$$r_\pi(\hat{\beta};\beta) = \int R(\hat{\beta};(\alpha,\beta))\pi(d\alpha). \tag{6.2}$$

If we adopt such a measure, then one fundamental question is under which prior $\pi(\alpha)$ the original estimator $\hat{\beta}$ is dominated by $\hat{\beta}_c$, that is, $r_\pi(\hat{\beta}_c;\beta) \leq r_\pi(\hat{\beta};\beta)$ for all β?

Intuitively, it is possible for $\hat{\beta}_c$ to dominate $\hat{\beta}$ in terms of r_π when π puts enough mass on or near $\alpha = 0$, as suggested by Part (F) of Section 4. The trouble is that in practice we will not know how close the restrictive model is to the truth *when we wish for an automated bias-variance trade-off,* because if we knew, then we surely should have included the information in our model to improve our estimator (e.g., via an informative prior), just as if we know $\alpha = 0$ for sure, then we should just use $\hat{\beta}_0$ (assuming it is an efficient estimator under the sub-model). We therefore seem to run into a circular situation. The information we need to evaluate $\hat{\beta}_c$ meaningfully makes $\hat{\beta}_c$ unnecessary, but without it, there does not seem to exist a meaningful way to establish the superiority of $\hat{\beta}_c$.

This was the main reason that I suspected that $\hat{\beta}_c$ was more a craving than a creation. I of course hope my suspicion is groundless and that M&C's proposal can lead to a real advancement at the frontier of methods for accomplishing appropriate bias-variance tradeoff. But this is a case where only hard theory, not simulations nor intuitions, can settle the matter. After all, the whole industry of shrinkage estimation came out of the counter-intuitive—at

least initially—Stein's paradox established by rigorous theory (Stein, 1956; James and Stein, 1961; Efron and Morris, 1977). There might be an empirical partially Bayes theory in parallel to the elegant one established by Efron and Morris (1973) for shrinkage via empirical Bayes, but the key ingredient in M&C, that is, estimating the prior variance via the *conservative* $(\hat{\beta}_0 - \hat{\beta})^2$ is likely to be fatal to this line of exploration because the performance of $\hat{\beta}_c$ depends critically on the reliability of this estimation.

Evidently, there is a lot to be learned from the classic theory of admissibility before we can settle this matter, because this is squarely a problem of comparing estimators under the squared loss. Professor Berger has done much to build this field, so it is only fitting for me to present the problem of comparing $\hat{\beta}_c$ and $\hat{\beta}$ in general as a piece of cake to him on the occasion of his 60th birthday.

Happy Birthday, Jim, even if the cake turns out to be inadmissible!

References

1. Berger, J.O. (1985) *Statistical Decision Theory and Bayesain Analysis*. Springer-Verlag, New York.
2. Cox, D. R. (1975) A note on partially Bayes inference and the linear model. *Biometrika*, **92**, 399-418.
3. Efron, B and Morris, C. (1973) Stein's estimation rule and its competitors–an empirical Bayes approach. *The Journal of American Statistical Association* **68** 117-130.
4. Efron, B. and Morris, C. (1977). Stein's paradox in statistics. *Scientific American* **236(5**), 119127.
5. Ferguson, T.S. (1996) *A Course in Large Sample Theory*. Chapman& Hall, London.
6. James, W. and Stein, C. (1961) Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical and Statistical Probabilities* **Vol. 1**, 361-379. University of California Press, Berkeley.
7. McCullagh, P (1990) A note on partially Bayes inference for generalized linear models. *Technical Report 284*, Department of Statistics, The University of Chicago.
8. Meng, X.-L.(1994). Posterior predictive p-values. *The Annals of Statistics* **22**, 1142-1160.
9. Meng, X.-L., Lee, T.C.M., Li, Z. (2009) What can we do when EM is not applicable? Self consistency principle for semi-parametric and non-parametric estimation with incomplete and irregularly spaced data. *Statistical Science*, under revision.
10. Mukherjee, B and Chatterjee, N. (2008) Exploiting gene-environment independence for analysis of casec-ontrol studies: an empirical Bayes-type shrinkage estimator to trade-off between bias and efficiency. *Biometrics*, **64(3)**, 685 - 694.
11. Piegorsch, W., Weinberg, C.R., and Taylor, J.A. (1994) Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Statistics in Medicine* **13(2)**, 153 - 162.
12. Schmidt, S. and Schaid, D.J. (1999) Potential misinterpretation of the case-only study to assess gene-environment interaction. *American Journal of Epidemiology.* **150(8)**, 878-885.
13. Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate distribution. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* **Vol 1**, 197 - 206. University of California Press, Berkeley.