I Got More Data, My Model is More Refined, but My Estimator is Getting Worse! Am I Just Dumb?

Submitted to Econometric Reviews on Bayesian Inference and Information Theoretic Methods: In Memory of Arnold Zellner

I GOT MORE DATA, MY MODEL IS MORE REFINED, BUT MY ESTIMATOR IS GETTING WORSE! AM I JUST DUMB?

By Xiao-Li Meng and Xianchao Xie

Harvard University

Possibly, but more likely you are merely a victim of conventional wisdom. More data or better models by no means guarantee better estimators (e.g., with a smaller mean squared error), when you are not following probabilistically principled methods such as MLE (for large samples) or Bayesian approaches. Estimating equations are particularly vulnerable in this regard, almost a necessary price for their robustness. These points will be demonstrated via common tasks of estimating regression parameters and correlations, under simple models such as bivariate normal and ARCH(1). Some general strategies for detecting and avoiding such pitfalls are suggested, including checking for self-efficiency (Meng, 1994, *Statistical Science*) and adopting a guiding working model.

Using the example of estimating the autocorrelation ρ under a stationary AR(1) model, we also demonstrate the interaction between model assumptions and observation structures in seeking additional information, as the sampling interval s increases. Furthermore, for a given sample size, the optimal s for minimizing the asymptotic variance of $\hat{\rho}_{MLE}$ is $s = 1$ if and only if $\rho^2 \leq 1/3$; beyond that region the optimal s increases at the rate of $\log^{-1}(\rho^{-2})$ as ρ approaches a unit root, as does the gain in efficiency relative to using $s = 1$. A practical implication of this result is that the so-called "non-informative" Jeffreys prior can be far from non-informative even for stationary time series models, because here it converges rapidly to a point mass at a unit root as s increases. Our overall emphasis is that intuition and conventional wisdom need to be examined via critical thinking and theoretical verification before they can be trusted fully.

Keywords and phrases: AR(1) model, Estimating equation, Fraction of missing information, Fisher information, Generalized method of moments (GMM),, Jeffreys prior, Non-informative prior, Partial plug-in, Observation structures, Relative information, Self-efficiency, Unit root.

1. What Does *Information* Really Mean?. Information is a buzzword in the information age. To the general public, information is a buzzword because it is interwoven into every fabric of our lives—it is now nearly impossible to find a "digital information free zone". To those of us who study or use information as a quantitative measure, it is a buzzword because we have trouble quantifying it generally enough so that its technical meaning would match its daily usage with appreciable accuracy. But nevertheless we continue trying.

In the context of statistical analysis, the technical meaning of information often is directly linked to the amount of data we have and to a measure of the quality (e.g., confidence coverage, testing power) of our inferential conclusions. An insightful reader may already be troubled by an undertone in the previous sentence: the more data we have, the more information and hence the higher quality of our statistical findings. As a matter of fact, this seemingly trivially logical intuition is false, because more data lead to better conclusions only when we know how to take advantage of their information. In other words, size does matter, but only if it is used appropriately.

Here is a simple example taken from Xie and Meng (2012). We have a sequence of 100 (time ordered) observations from a heteroscedastic regression model

$$Y_t = \beta X_t + \epsilon_t, \quad \epsilon_t \overset{\text{indep.}}{\sim} N(0, X_t^2), \quad t = 1, \ldots, n. \tag{1.1}$$

The ordinary least squares (OLS) estimator for β enjoys a celebrated robustness, that is, it is consistent even in the presence of heteroscedasticity, and its variance can be consistently estimated via the usual "sandwich estimator". Much less well known, however, is the fact that as a necessary price for this robustness, OLS is not *self-efficient* (Meng, 1994), because it can yield a more accurate estimator with less data.

For instance, in the above example, if $X_t = (101 - t)^{-1}$, then using the first 64 observations will lead to the variance of OLS $V_{1:64}^{OLS} = 0.0214$, yet if we use all our observations, $V_{1:100}^{OLS} = 0.4049$. That is, by adding about 1/3 of the data, we end up inflating the variance almost 19 times instead of reducing it. In other words, if we measure how much more information is gained by having the additional 36 observations (that happen to be the last 36 observations), we will have to conclude that there is actually a tremendous loss—or very negative gain—of information for OLS.

In contrast, if we use the maximum likelihood estimator (MLE), which for the current case amounts to a weighted least squares estimator with weight $W_t = X_t^{-1}$, its variance becomes n^{-1} if we use the first n observations; consequently, for example, doubling n will cut its variance by half, as we

normally expect. Intuitively, the failure of OLS is due to its equal weighting: putting those observations with very large variances on equal footing with those with very small ones. If the added observations are from those with larger variances, the noise they bring in can easily outweigh the gain in sample size. This is very much like diversification. It can be a sound strategy for constructing low risk portfolios, but simply allocating an equal amount share of a new stock can greatly increase the risk of the portfolio if the new stock is sufficiently volatile. The MLE properly weights observations so the amount of noise each of them can bring in is equalized, and hence the size of the weighted sample is proportional to the amount of information being accumulated.

This somewhat dramatic numerical illustration is chosen to highlight the fact that the conventional wisdom "more data imply better estimators" can be trusted only under further qualifications. The time trend in the variance function of (1.1) may seem to be artificial to a casual reader, but it is a common phenomenon in time series analysis, especially with ARCH/GARCH type of models, as we shall detail in Section 2 via a simple ARCH model. Using the same model, Section 3 provides an overview of the aforementioned concept of self-efficiency, which requires more than merely ensuring decreasing in uncertainty (e.g., variance) as we have more data (and hence we delay its discussion after presenting the simpler case in Section 2). We emphasize here that this is merely one of many approaches (and information measures) for studying such problems. As one of many examples, Abel and Singpurwalla (1994) and Ebrahami, Soofi and Soyer (2012) used the entropy to study whether one prefers to observe more failures than survivals in survival models. Also see Ebrahami, Soofi and Soyer (2010) for an excellent overview and investigation of measuring mutual information for estimation and for prediction.

In Section 4, we review how to define, calculate and interpret relative information in the context of MLE, where by relative information we mean relative gain by making additional assumptions. We use the problem of estimating correlation to illustrate why great caution is needed to leverage additional information when we are not using a principled procedure such as MLE or Bayesian methods. Section 5 then carries out the information calculations in the context of time series data with conditionally normal errors. The calculation re-confirms how the MLE approach sensibly accumulates information as we add more assumptions and/or more observations. This detailed re-examination of the power of the MLE approach also suggests a general strategy for incorporating additional information when estimating equations are employed for dependent data where only the first two condi-

tional moments are specified. Section 6 then applies the results in Section 5 to the simple AR(1) model, which in particular leads to the result on how the Fisher information depends on the sampling intervals, as summarized in the Abstract.

We doubt any technical results reported in this paper have not been noted previously, given they are based on rather standard calculations, despite the fact that we have not been able to locate a reference for the aforementioned sampling interval results. (We nevertheless note that these results are derived from the parameter estimation perspective, not from the usual signal recovering perspective underlying the well-known Nyquist-Shannon sampling theorem; see Nyquist, 1928 and Unser, 2000). We do, however, emphasize that the key theoretical insights these results convey have not received the general attention they deserve, because they have direct practical implications and can help to guide practitioners to use their resources for data collection and analysis in a more economical way. As raising the general awareness of such issues is the main purpose of our article, we conclude in Section 7 with a brief discussion on how the sampling results help to clarify a common mis-perception that the Jeffreys prior is "safe" to use because it comes with the label of being "non-informative". Whereas there are many theoretical, practical or even philosophical reasons to adopt a Jeffreys prior, a topic about which one can learn much from Professor Arnold Zellner's writing[1], its "non-informative" label is generally misleading. As we will show, as the sampling interval increases, the Jeffreys prior will converge to a point mass at a unit root, hardly non-informative by any measure.

2. Do Additional (Correct) Data Always Help?. The answer is clearly no from the simple example in Section 1. Here we use the well-known ARCH(1) regression model to further illustrate this point.

Let $\{Y_t, t = 1, \ldots, T\}$ represent the *entire* time series within a given finite-time horizon T, for which we believe a single (parametric) model is adequate. That is, we do not consider the process starting from a hypothetical infinite past, but rather from a fixed time, which will be labelled as the origin of the process $t = 0$, with a fixed but potentially unknown value Y_0. For simplicity of algebra, let us assume that the model we adopt is the simplest ARCH regression model with a single predictor (Engle, 1982):

$$Y_t|\mathcal{F}_{t-1} \sim N(X_t\beta, \tau_t^2), \qquad t = 1, \ldots, T \tag{2.1}$$

where $\mathcal{F}_{t-1}$ is the σ-field generated by $\{Y_1, \ldots, Y_{t-1}\}$, with $\mathcal{F}_0$ being defined

[1]See his CV at `http://faculty.chicagobooth.edu/arnold.zellner/more/vita.pdf`

as the trivial σ field, and the conditional variance

$$\tau_t^2 \equiv \alpha_0 + \alpha_1 \epsilon_{t-1}^2, \quad \text{where } \epsilon_{t-1} = Y_{t-1} - X_{t-1}\beta, \tag{2.2}$$

with $\alpha_0 > 0$ and $\alpha_1 \geq 0$. Here, we consider $\epsilon_0 \equiv Y_0$ (i.e., we assume $X_0 = 0$) to be an unknown fixed parameter. [It is known that, although ARCH(1) is not an AR process, the squared process $\{\epsilon_t^2, t = 1, \ldots, \}$ can be viewed as an AR(1) process in a general sense; see Bollerslev (1986).]

Suppose we observe only $Z_{obs} = \{Y_{t_1}, \ldots, Y_{t_n}\}$, where n is the size of the data and $1 \leq t_1 \leq t_n \leq T$. It is well-known that, conditioning on the (observed) X_t's, OLS estimator

$$\hat{\beta}_{t_1:t_n} = \frac{\sum_{j=1}^n Y_{t_j} X_{t_j}}{\sum_{j=1}^n X_{t_j}^2} \tag{2.3}$$

is unbiased for β regardless of the values of α_0 and α_1. Its variance is given by

$$V_{t_1:t_n} = \frac{\sum_{j=1}^n \sigma_{t_j}^2 X_{t_j}^2}{[\sum_{j=1}^n X_{t_j}^2]^2}, \tag{2.4}$$

where σ_t^2 is the *marginal variance* of Y_t given by

$$\sigma_t^2 = Y_0^2 \alpha_1^t + \alpha_0 \sum_{j=0}^{t-1} \alpha_1^j, \quad \text{for } t \geq 1, \tag{2.5}$$

with σ_0^2 defined as Y_0^2. Note in deriving (2.5) we have used the fact that, conditioning on X_t's, Y_t and Y_s are un-correlated whenever $t \neq s$ even though they are not independent.

Because of the dependence of σ_t^2 on t, we see that $V_{t_1:t_n}$ is not necessarily a monotonic decreasing function of the size n even when all the X_t's are equal, that is, even when we simply estimate the mean of Y_t and hence ignore the variations among X_t's. As a matter of the fact, it can be a monotonic *increasing* function, indicating that we would be worse off with more data, if we were seduced by OLS for its simplicity and robustness. Indeed, Engle (1982) pointed out the potential of 100% loss of efficiency by OLS compared to MLE as $\alpha_1 \uparrow 1$. (See also Pantula, 1988, for comparisons of MLE, OLS, and a generalized least squared estimator.) The condition $\alpha_1 < 1$ was needed in Engle (1982) to ensure finite variance because he was considering the ARCH process with an infinite time horizon, a condition that is unnecessary when we restrict ourselves to a finite time horizon. However, our finite-time

horizon formulation permitting $\alpha_1 = 1$ provides an insight of why OLS will lose efficiency completely as $\alpha_1 \uparrow 1$, as shown below.

To see this clearly, we notice that when $\{t_1, \ldots, t_n\}$ forms a consecutive sequence and $X_t = 1$ for all t, (2.4) becomes

$$V_{t_1:t_n} = \begin{cases} \frac{1}{n}\frac{\alpha_0}{1-\alpha_1} + \frac{\alpha_1^{t_1}}{n^2}\left(Y_0^2 + \frac{\alpha_0}{\alpha_1 - 1}\right)\frac{\alpha_1^n - 1}{\alpha_1 - 1}, & \text{if } \alpha_1 \neq 1; \\ \frac{1}{n}\left[Y_0^2 + \alpha_0 \frac{t_1 + t_n}{2}\right], & \text{if } \alpha_1 = 1. \end{cases} \tag{2.6}$$

From (2.6), we observe several facts. First and trivially, if we let both t_1 and t_n go to infinity, that is, when we invoke the infinite time horizon formulation, then we recover the well-known result (Engle, 1982) that the process is stationary and has variance $\sigma^2 = \alpha_0/(1 - \alpha_1)$ (and hence $V_{t_1:t_n} = \sigma^2/n$) if and only if $\alpha_1 < 1$. Second, for a finite t_1 and t_n, $V_{t_1:t_n}$ is affected by both the data size n and the beginning and ending times of observation: t_1 and t_n. The initial value Y_0 can also have very large impact as long as $\alpha_1 \geq 1$.

Third, for a fixed t_1, when $\alpha_1 > 1$, we see that as long as n is larger than a small (often very small) threshold, $V_{t_1:t_n}$ is a monotone increasing function of n, implying that as we increase the sample size, OLS becomes progressively worse, with (essentially) an exponential rate α_1^n/n^2 approaching infinity. Note this explosive nature of the model often makes it unsuitable for modeling financial data (e.g., Tsay, 2001), but nevertheless we include it here both for completeness and for illustrating the possibility that the variance of a seemingly appropriate estimator can increase with the data size at an arbitrarily fast rate if one is not using a principled method such as MLE.

The more interesting and relevant case is when $\alpha_1 = 1$. In such case, $V_{t_1:t_n}$ is a monotone decreasing function of n for fixed t_1, but it does not converge to zero as $n \to \infty$; rather, it converges to $\alpha_0/2$. Therefore, with an increase of the data size, $\hat{\beta}_{t_1:t_n}$ fails to converge to the true β because it does not appropriately utilize information in the data, resulting in a total loss of information compared to MLE as $\alpha_1 \uparrow 1$.

We emphasize here that, although we have invoked a finite-time horizon setup, taking limits such as $n \to \infty$ is still relevant both as a mathematical tool for approximation and for gaining theoretical insight. For example, one may question if there is any estimator of β that can have vanishing variance (i.e., converging in L^2) as n grows without limit, which would indicate that there is a complete loss of information by OLS. The answer is yes because the inverse t weighted estimator

$$\hat{\beta}_n = \frac{\sum_{t=1}^n Y_t/t}{\sum_{t=1}^n 1/t} \tag{2.7}$$

is unbiased and has variance

$$\text{Var}(\hat{\beta}_n) = \frac{Y_0^2 H_n^{(2)}}{[H_n^{(1)}]^2} + \frac{\alpha_0}{H_n^{(1)}}, \text{ where } H_n^{(i)} = \sum_{t=1}^{n} t^{-i}, i = 1, 2. \tag{2.8}$$

Since $H_\infty^{(2)} = \pi^2/6$ but the harmonic series $H_n^{(1)}$ goes to infinity as $\log(n)$ does, we know that $\text{Var}(\hat{\beta}_n)$ converges to zero at the rate of $\log^{-1}(n)$. Whereas this is a very slow rate, it nevertheless establishes the existence of perfect information accumulation in the sense of eliminating uncertainty eventually, in contrast to OLS, which will still have variance $\alpha_0/2$ even if we have an infinite amount of data. In the next section, we will prove that OLS fails to be self-efficient asymptotically when $\alpha_1 \geq 1$ (the case for $\alpha_1 > 1$ of course is obvious), which is a stronger indication of its improper extraction of information from the available data.

3. What is Self-Efficiency?. The general notion of self-efficiency was introduced in Meng (1994), in the context of investigating the consistency of Rubin's (1987) variance combining rule for multiple imputation inference. The following (refined) definition is from Xie and Meng (2012).

DEFINITION 1. *Let Z_{com} be a data set and Z_{obs} be a subset of Z_{com} created by a selection mechanism. A statistical estimation procedure $\hat{\theta}(\cdot)$ for θ is said to be self-efficient (with respect to the selection mechanism) if for any (constant) $\lambda \in (-\infty, \infty)$, $\hat{\theta}_{com}$ dominates $\lambda\hat{\theta}_{obs} + (1-\lambda)\hat{\theta}_{com}$ in terms of the mean squared error (MSE), where $\hat{\theta}_{com} = \hat{\theta}(Z_{com})$ and $\hat{\theta}_{obs} = \hat{\theta}(Z_{obs})$.*

In a nutshell, self-efficiency eliminates any procedure that can be improved upon by "bootstrapping", in its original (non-technical) meaning. Imagine the following scenario. A user of an estimation procedure applies it to all the data he has and obtains $\hat{\theta}_{full}$. Upon calculating its variance, the user finds that the variance of $\hat{\theta}_{full}$ is too large to be acceptable. He has no new data nor any new prior information that can help him. However, he discovers that he can improve upon $\hat{\theta}_{full}$ by first applying the same procedure to a subset of the data to obtain $\hat{\theta}_{part}$, and then forming a weighted average $\hat{\theta}_\lambda = \lambda\hat{\theta}_{part} + (1-\lambda)\hat{\theta}_{full}$ for some λ. Regardless of how he chooses the subset or λ, the fact that $\hat{\theta}_\lambda$ can beat $\hat{\theta}_{full}$ means that the original procedure he adopts is not efficient even with respect to itself (given the information he has, which can be reflected in his choice of subset or λ), because it allows itself to be improved upon applying it to a subset of the data after it has already been applied to the full data set. We remark here that in principle $\hat{\theta}_{part}$ and $\hat{\theta}_{full}$ can be combined in any way; we restricted ourselves to linear

combinations mainly because typically we apply the notion asymptotically, for which linear combinations suffice. See also Xie and Meng (2012) for a discussion on how self-efficiency is a special case of strong efficiency, which itself is closely related to Rao-Blackwellization.

To illustrate why the self-efficiency is a stronger requirement than only requiring the MSE to be non-increasing as we increase sample size, consider the simple example in Section 1. Given the nature of increasing variances over time, one may wonder whether the issue of "self-inefficiency" is avoided if the selection mechanism always takes the last (say) 64 observations (even though one clearly should avoid as much as possible any mechanism that selects the worst n observations!). Indeed this will lead to $V^{OLS}_{37:100} = 0.4076 > V^{OLS}_{1:100} = 0.4049$. (The proximity of these two variances also reminds us how OLS is dominated by those observations with larger variability). However, this fact does not imply that OLS is self-efficient with respect to the specified selection mechanism. This is because we can still find a linear combination of the complete-data OLS and the observed-data OLS that enjoys a smaller variance. For example, let $\hat{\beta}_\lambda = \lambda \cdot \hat{\beta}^{OLS}_{37:100} + (1-\lambda) \cdot \hat{\beta}^{OLS}_{1-100}$ and choose $1-\lambda = 274.5$, we have $V(\hat{\beta}_\lambda) = 0.0276$, much smaller than $V^{OLS}_{1:100}$.

What happens here is that the large negative value of $\lambda = -273.5$ allows $\hat{\beta}_\lambda$ to essentially subtract out a "bad part" of the complete-data OLS, that is, the part using the last 64 observations. Indeed, $V(\hat{\beta}_\lambda) = 0.0276$ is quite close to $V^{OLS}_{1:36} = 0.0296$. It is not surprising to see that our optimally-chosen $\hat{\beta}_\lambda$ does slightly better than $\hat{\beta}^{OLS}_{1:36}$. This is because the equal-weighting nature of OLS permits $\hat{\beta}^{OLS}_{1:36}$ to be written as a linear combination of $\hat{\beta}^{OLS}_{1:100}$ and $\hat{\beta}^{OLS}_{37:100}$, and hence the optimal λ would allow $\hat{\beta}_\lambda$ to beat $\hat{\beta}^{OLS}_{1:36}$ by extracting out a tiny amount of useful information from the last 64 observations that was not already in the first 36 observations, *as available to the OLS approach.* Of course the best procedure is simply to use the MLE based on all 100 observations, which has a much smaller variance, $V^{MLE}_{1:100} = 0.01$.

For the ARCH model in Section 2, clearly OLS is not self-efficient when $\alpha_1 > 1$ because any self-efficient procedure will necessarily reduce MSE (or variance when we deal only with unbiased estimators, such as OLS) as the data size grows. But, as the example above demonstrates, the non-increasing MSE property is not a sufficient condition for self-efficiency. This is further illustrated by the case of $\alpha_1 = 1$. For simplicity, let us assume $t_1 = 1$. Then the relative efficiency of $\hat{\beta}_{1:n} = \bar{Y}_n$ to that of the full-data estimator $\hat{\beta}_{1:N} = \bar{Y}_N$ is given by

$$RE_n = \frac{V_{1:N}}{V_{1:n}} = 1 - f\frac{\alpha_0 + 2Y_0^2}{(n+1)\alpha_0 + 2Y_0^2} = 1 - f\left[\frac{\sigma_1^2 + \sigma_0^2}{\sigma_1^2 + \sigma_n^2}\right], \tag{3.1}$$

where $f = (N-n)/N$ is the fraction of missing data, and the rightmost expression is due to (2.5) with α_1=1. Whereas this RE_n never exceeds one, it approaches one rather rapidly as σ_n^2 increases, in contrast to the situation when Y_t's are i.i.d., a case where $RE_n = n/N = 1-f$ (which can also be seen in (3.1) if σ_t^2 does not change with t). That is, if one uses OLS, then the information in much of the later part of the time series is essentially wasted.

This fact suggests that it should be fairly easy to dominate $\bar{Y}_N$ by combining it with $\bar{Y}_n$. To see this, consider $\hat{\beta}_\lambda = \lambda\bar{Y}_n + (1-\lambda)\bar{Y}_N$. Using the fact $\mathrm{Cov}(\bar{Y}_n, \bar{Y}_N) = (1-f)\mathrm{Var}(\bar{Y}_n)$, we have

$$\mathrm{Var}(\hat{\beta}_\lambda) = \left[\lambda^2 + 2\lambda(1-\lambda)(1-f)\right] V_{1,n} + (1-\lambda)^2 V_{1,N}. \tag{3.2}$$

Hence $\mathrm{Var}(\hat{\beta}_\lambda) < \mathrm{Var}(\bar{Y}_N)$ is equivalent to (assuming $0 < \lambda < 2$)

$$\frac{\lambda + 2(1-\lambda)(1-f)}{2-\lambda} < \frac{V_{1:N}}{V_{1:n}} = RE_n. \tag{3.3}$$

This leads to

$$0 < \lambda < 2\frac{RE_n - 1 + f}{RE_n - 1 + 2f} = \frac{2n}{(2n+1) + 2Y_0^2/\alpha_0}. \tag{3.4}$$

That is, for any λ that is not too close to 1 in the sense of satisfying (3.4), $\hat{\beta}_\lambda$ has a smaller MSE than the full-data OLS $\bar{Y}_N$, which implies that OLS is not self-efficient for ARCH (when $\alpha_1 = 1$). Note the choice of λ here does not depend on the estimand β, an obvious requirement for the notion of self-efficiency to be meaningful. The expression (3.4) also illustrates the impact of Y_0^2. On one hand, when $Y_0^2 = 0$ or is relatively small compared to α_0, there would be no or little impact of the value of α_0 on the possible range of λ . On the other hand, when Y_0^2 is large (relative to α_0) it will take some quite large n to diminish its initial impact on σ_n^2, and hence the linear combination $\hat{\beta}_\lambda = \lambda\bar{Y}_n + (1-\lambda)\bar{Y}_N$ needs to give more weight to $\bar{Y}_N$ in order to dominate the full-data estimator $\bar{Y}_N$.

Since OLS is a root of a second-order regular estimating equation (SOREE, which means that it satisfies a set of standard differentiability and integrability assumptions, as listed in Xie and Meng, 2012), there is an indirect but typically easier way to ascertain whether the root is self-efficient. Specifically, Xie and Meng (2012) established that if our complete-data estimator $\hat{\theta}_{com}$ is derived from a SOREE $h_{com}(Z_{com};\theta) = 0$, and the observed-data estimator $\hat{\theta}_{obs}$ from a SOREE $h_{obs}(Z_{obs};\theta) = 0$, then the corresponding estimating procedure is self-efficient asymptotically if and only if asymptotically

$$\left[E\left(-\frac{\partial h_{obs}}{\partial\theta}\right)\right]^{-1} E\left(h_{obs}h_{com}^\top\right) = \left[E\left(-\frac{\partial h_{com}}{\partial\theta}\right)\right]^{-1} E\left(h_{com}h_{com}^\top\right). \tag{3.5}$$

Note here the negative sign on both sides is unnecessary for the equality but is needed to make a connection with the expected Fisher information, as discussed below. Whereas the meaning of (3.5) may not be immediately clear, a special case, which is applicable to our ARCH example, helps to reveal the essence of (3.5). It is obvious that if (asymptotically)

$$\left[E\left(-\frac{\partial h_{obs}}{\partial \theta}\right)\right]^{-1} E\left[h_{obs}(h_{com}-h_{obs})^{\top}\right]=0, \tag{3.6}$$

then (3.5) becomes (asymptotically)

$$\left[E\left(-\frac{\partial h_{obs}}{\partial \theta}\right)\right]^{-1} E\left(h_{obs}h_{obs}^{\top}\right)=\left[E\left(-\frac{\partial h_{com}}{\partial \theta}\right)\right]^{-1} E\left(h_{com}h_{com}^{\top}\right). \tag{3.7}$$

When $h_{com}(Z_{com};\theta)$ and $h_{obs}(Z_{obs};\theta)$ are, respectively, the score functions from complete-data and observed-data likelihoods, we see (3.7) holds trivially because of the second Bartlett identity, namely, the expected Fisher information is the same as the variance of the score function (see Section 5), and therefore both sides are the identity matrix. In this sense, (3.5) or (3.7) can be viewed as an extension of the second Bartlett identity, which plays a key role in likelihood inference and alike; see Meng (2009, 2011) for a recent investigation of the role of Bartlett identities in H-likelihood.

For our ARCH example, we let $Z_{com}=\{Y_t, t=1,\ldots,N\}$, and $Z_{obs}=\{Y_t, t=t_1,\ldots,t_n\}$, and accordingly (note here $\theta=\beta$)

$$h_{com}(Z_{com};\theta)=\sum_{t=1}^{N}(Y_t-\beta X_t)X_t, \quad h_{obs}(Z_{obs};\theta)=\sum_{t=t_1}^{t_n}(Y_t-\beta X_t)X_t. \tag{3.8}$$

Using the fact that Y_t and Y_s are uncorrelated whenever $t\neq s$ (conditioning on X_t's), we see (3.6) holds trivially in general but (3.7) does not. This is because that even in the case where $X_t\equiv 1$, the left-hand side of (3.7), which is $nV_{t_1:t_n}$, can be asymptotically the same as its right-hand side $NV_{1:N}$, if and only if $\alpha_1<1$, in which case both sides converge to $\sigma^2=\alpha_0/(1-\alpha_1)$. Therefore, OLS is (asymptotically) self-efficient for ARCH with $X_t\equiv 1$ if and only if $\alpha_1<1$, assuming we do not put any constraints on what t_1 and t_n can be.

If we do allow ourselves to put restrictions on where we collect our observations $Z_{obs}=\{Y_t, t_1\le t\le t_n\}$, we can see from (2.6) that there is one special class of sub-samples for which OLS can be regarded as self-efficient even when $\alpha_1=1$. This is because when $\alpha_1=1$, $V_{t_1:t_n}$ depends on t_1 and t_n via its length n and the "center" $(t_1+t_n)/2$. Therefore, if the selection mechanism for Z_{obs} (see Definition 1) is such that it selects only middle segments

of the entire series, namely, with $t_1 + t_n = 1 + N$, exactly or asymptotically, then (3.7) holds in the same fashion. This, of course, says little about OLS being a good procedure, but rather that it cannot be improved upon by using itself in such cases. In other words, self-efficiency is a form of minimal requirement for a sensible estimating procedure. Without it, we can expect paradoxical behaviors and a great loss of information. But with it, the procedure can still be much inferior to other procedures (e.g., MLE), including being inconsistent (since Definition 1 says nothing about being consistent).

4. Do Additional (Correct) Assumptions Always Help?. The answer to this question is the same as that to the question in the title of Section 2, that is, "NO." Also as before, the answer becomes "YES" when we use probabilistically principled methods such as MLE and Bayesian approaches. The following brief review reminds us how Fisher information necessarily increases when we make (relevant) assumptions to reduce our model class.

For simplicity, let us assume our model parameter $\theta = (\theta_1, \theta_2)^\top$ is two dimensional, with θ_1 being of primary interest and θ_2 serving as the nuisance parameter. For notation, we adopt the convention

$$I(\theta) = \begin{pmatrix} i_{11} & i_{12} \\ i_{21} & i_{22} \end{pmatrix} \quad \text{and} \quad I^{-1}(\theta) = \begin{pmatrix} i^{(11)} & i^{(12)} \\ i^{(21)} & i^{(22)} \end{pmatrix}; \tag{4.1}$$

here $I(\theta)$ can be either the expected Fisher information or the observed Fisher information. The difference between these two versions can be quite important in practice (e.g., Efron and Hinkly, 1978), but the following discussion is relevant for either version. We will also assume all the usual regularity conditions for justifying the asymptotical arguments in the following derivations.

As is well known, without any restrictions on the nuisance parameter θ_2, the asymptotic variance for the MLE for θ_1 is given by $i^{(11)}$. Therefore, if we define Fisher information for an individual parameter by the inverse of the variance of its MLE, we see

$$I(\theta_1) = [i^{(11)}]^{-1} = i_{11} - \frac{i_{12}^2}{i_{22}} \quad \text{and} \quad I(\theta_1|\theta_2) = i_{11}, \tag{4.2}$$

where $I(\theta_1|\theta_2)$ means the Fisher information for estimating θ_1 conditioning on the value of θ_2 being given. Let

$$\mathcal{G}(\theta_1|\theta_2) = I(\theta_1|\theta_2) - I(\theta_1) \quad \text{and} \quad \mathcal{R}(\theta_1|\theta_2) = \frac{\mathcal{G}(\theta_1|\theta_2)}{I(\theta_1)} \tag{4.3}$$

be respectively the absolute gain and relative gain in Fisher information for estimating θ_1 due to knowing θ_2. Then (4.2) justifies the phrase "*gain*"

because $\mathcal{G}(\theta_1|\theta_2) \geq 0$ and it is zero if and only if $i_{12} = 0$, that is, when the two parameters θ_1 and θ_2 are orthogonal in the sense that the Fisher information matrix $I(\theta)$ is diagonalized. The meaning of $i_{12} = 0$ can be seen even clearer from a Bayesian perspective, because

$$\mathcal{R}(\theta_1|\theta_2) = \frac{r^2(\theta_1, \theta_2)}{1 - r^2(\theta_1, \theta_2)}, \tag{4.4}$$

where

$$r(\theta_1, \theta_2) = \frac{i^{(12)}}{\sqrt{i^{(11)} i^{(22)}}} = \frac{-i_{12}}{\sqrt{i_{11} i_{22}}}$$

is the (limit of the) asymptotic posterior correlation between θ_1 and θ_2 (this is most easily seen when we use *observed* Fisher information). Hence the relative gain is completely determined by how correlated θ_1 and θ_2 are. If $i_{12} = 0$, then the two parameters are uncorrelated, and hence information on θ_2 provides no help for estimating θ_1. At the other extreme, the gain is infinity when θ_1 is determined by θ_2, and hence knowing θ_2 implies that θ_1 is also known with certainty.

All the calculations above are well known and seem to only confirm the obvious. What is less well known or obvious is that as soon as we move from MLE to an estimating equation setting, then the "obvious" no longer holds in general. Below is a non-trivial example demonstrating how assumptions on the nuisance parameters can actually do serious harm when we employ them in a seemingly very natural but actually very flawed way, even when these assumptions are known to be true (and therefore the issue addressed here is not the usual bias-variance trade-off due to incorrect assumptions).

Suppose we want to estimate the correlation ρ for bivariate normal data $\{(x_i, y_i), i = 1, \ldots, n\}$. Without making any restriction on other model parameters $\phi = \{\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2\}$, we know the sample correlation

$$\hat{\rho}_n = \frac{S_{XY} - n\bar{X}\bar{Y}}{S_X S_Y} \equiv h_n(\hat{\phi}_n, S_{XY}) \tag{4.5}$$

is the MLE and hence it is asymptotically efficient with asymptotic variance $(1 - \rho^2)^2/n$ (see Ferguson, 1996); here $\hat{\phi}_n = \{\bar{X}, \bar{Y}, S_X, S_Y\}$ is the MLE of ϕ (and hence the sample variances S_X^2 and S_Y^2, as well as the sample cross product S_{XY}, are defined with denominator n instead of the usual $n - 1$). Now suppose the data are such that both X and Y are marginally standard normal $N(0, 1)$, hence we know that $\phi = \phi_0 = \{0, 0, 1, 1\}$. The Fisher information for this restrictive model then is $n(1 + \rho^2)/(1 - \rho^2)^2$. That is, in our notation we have

$$I(\rho) = \frac{n}{(1 - \rho^2)^2} \quad \text{and} \quad I(\rho|\phi_0) = \frac{n(1 + \rho^2)}{(1 - \rho^2)^2}. \tag{4.6}$$

Therefore $\mathcal{R}(\rho|\phi_0) = \rho^2 \geq 0$. However, this gain is achieved by the MLE of ρ, which is a root of a cubic equation determined by $\{S_X^2, S_Y^2, S_{XY}\}$ (see Jeffreys 1983). It is *not* achieved by performing a seemingly very natural and intuitive "plug-in" step, that is, to replace the sample estimate $\hat{\phi}_n$ in the h_n function in (4.5) by its true value ϕ_0, which would lead to the estimator

$$\hat{r}_n \equiv h_n(\phi_0, S_{XY}) = S_{XY} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i. \tag{4.7}$$

Whereas this estimator can also be justified from an estimating equation derived from $E(XY) = \rho$ under the restrictive model, it is clearly not efficient because it only uses a part of the minimal sufficient statistics $\{S_X^2, S_Y^2, S_{XY}\}$.

Indeed it is a terrible estimator because it is not even guaranteed that $|\hat{r}_n| \leq 1$, the necessary range for any correlation. One can easily see that $\mathrm{Var}(\hat{r}_n) = (1+\rho^2)/n$ for any n, hence the corresponding information is $I^{(h_n)}(\rho|\phi_0) = n/(1+\rho^2)$. Consequently, the "gain" in information from knowing the $\hat{\phi}_n$ part of $h_n(\hat{\phi}_n, S_{XY})$ is actually negative for this seemingly obvious "plug-in" method because

$$\mathcal{R}^{(h_n)}(\rho|\phi_0) \equiv \frac{I^{(h_n)}(\rho|\phi_0) - I(\rho)}{I(\rho)} = -(3-\rho^2)\frac{\rho^2}{1+\rho^2}.$$

This loss of efficiency therefore approaches 100% as ρ^2 approaches 1. Intuitively, this is because the sample correlation will estimate ρ perfectly when $\rho^2 = 1$, but $\hat{r}_n$ would still incur a variance of $2/n$. The loss compared with the MLE of ρ given $\phi = \phi_0$ is even more substantial because by (4.6):

$$\frac{I^{(h_n)}(\rho|\phi_0) - I(\rho|\phi_0)}{I(\rho|\phi_0)} = -\frac{4\rho^2}{(1+\rho^2)^2}.$$

As a simple illustration, when $\rho^2 = 1/2$, the variance of $\hat{r}_n$ is 6 times the (asymptotic) variance of the sample correlation, and 9 times that of the optimal MLE.

We detail this example because it provides an excellent reminder of a number of theoretical insights that have direct general practical implications. First, whenever possible, more well-developed and probabilistically principled methods such as MLE and Bayesian estimation should be preferred. The MLE for ρ, which is a root of a cubic equation, is not something intuitive enough to derive without following the MLE recipe. (Incidentally we remark that the maximum entropy distribution in the class of distributions $\{f(X,Y), (X,Y) \in R^2 : E_f[XY] = \rho\}$ does not exist, which can be proved by using results given in Ebrahimi, Soofi, and Soyer (2008), as

pointed out by a reviewer.) Indeed, the substantial gain of information for estimating ρ from knowing ϕ has posed a great puzzle to even some professional statisticians (as we have tried it on some) because our common intuition may suggest that the information about the marginal distributions should provide no information (at least not asymptotically) for estimating the correlation, as the correlation is invariant to affine transformations of each margin. Not wanting to spoil a great opportunity for a good mental exercise, we will not reveal the answer to the puzzle but only mention that a clue can be found in Jeffreys (1983).

Second, this example illustrates that the common ad hoc method of "partial plug-in" can do much more harm than good, and it is not enough to use our "intuition" as the safeguard. After all, it was a conventional wisdom that got us into trouble in the first place, as many of us had reasoned "how could one do worse by replacing an estimated quantity with its truth?" The above example shows how, and not just worse but disastrously so! It is particularly worth mentioning that the estimator we started with in this example is the most efficient MLE under the unrestrictive model, and the way the replacement was made also seems to be very natural. We simply express the MLE as a function of minimal sufficient statistics $\{\hat{\phi}, S_{XY}\} = \{\bar{X}, \bar{Y}, S_X, S_Y, S_{XY}\}$, and substitute those components with their known estimands under the restrictive model. The negative result we obtained therefore is not due to any defect of the original estimator or a rather contrived "plug-in" step for the sake of constructing a pathological counterexample. Rather, it is squarely due to the fact that "partial plug-in" is not a valid general strategy (without further qualification) because there are no sound statistical principles behind it other than its seemingly "very natural and intuitive" appearance.

Third, as far as statistical estimation goes, it is always a good idea to be mindful of the potential *substantial* loss of efficiency when we invoke a criterion that does not address the full efficiency. Here $\hat{r}_n$ of (4.7) is both unbiased and self-efficient (the latter because it is an iid sum), but neither of these properties prevents it from being a terrible estimator. Hence this example also provides a great illustration of the difference between self-efficiency and full efficiency, and it re-confirms the well understood fact that unbiasedness is typically not sought for its own sake but rather as a consequence of other considerations (see, for example, the desirable SOUP property studied by Meng and Zaslavsky, 2002). Incidently, even when we consider small-sample MSEs, the unbiased $\hat{r}_n$ is still typically dominated by the sample correlation $\hat{\rho}_n$ of (4.5), because the latter is nearly unbiased in the sense that its relative bias is less than n^{-1} for any $n \geq 2$ for bivariate normal data, as proved in Meng (2005).

In a nutshell, moment estimators and more generally estimating equations are used frequently in practice especially in the economics literature because of their simplicity and robustness (to model assumptions). Whereas both simplicity and robustness are of great practical importance, this example reminds us of the need to exercise caution when we try to improve upon these methods by incorporating additional assumptions. When it is feasible, it would be much safer (than relying on "partial plug-in") to entertain a reasonable working model and use the score equation under the working model as a guideline to construct an improved estimating equation that incorporates the additional assumption. We will illustrate such a strategy in Section 5 for general time series data.

5. How Does the MLE Extract Information?. Let us consider a general class of conditionally normal models for dependent data, $\{Y_{t_1}, \ldots, Y_{t_n}\}$, where the subscripts $\{t_1, \ldots, t_n\}$ do not have to index time nor do they necessarily form a consecutive or even equal-spaced sequence, that is, the spacing $s_j = t_j - t_{j-1}$ can be arbitrary (but positive). Let $\mathcal{F}_{j-1}$ be the σ-field generated by $\{Y_{t_1}, \ldots, Y_{t_{j-1}}\}$, with $\mathcal{F}_0$ being the trivial σ-field. By conditionally normal we mean $p(Y_{t_j}|\mathcal{F}_{j-1}; \theta)$ is given by $N(\mu_j(\theta), \tau_j^2(\theta)), j = 1, \ldots, n$ where for notational simplicity we suppress—but not forget—the dependence of $\mu_j(\theta)$ and of $\tau_j^2(\theta)$ on $\mathcal{F}_{j-1}$. Indeed such unrestricted dependence makes this class of models rather general, in spite of the seemingly restrictive conditional normality assumption, because the resulting joint distribution of the entire data sequence $Z_{obs} = \{Y_{t_1}, \ldots, Y_{t_n}\}$ can be far from normal (e.g., the aforementioned ARCH model).

Under such a setting, the score function $S(\theta|Z_{obs})$, which again for notational simplicity we suppress Z_{obs} and use only its size n to remind ourselves of the data source, is given by

$$S_n(\theta) = -\sum_{j=1}^{n} \left[\frac{\tau_j'(\theta)}{\tau_j(\theta)} + d_j(\theta) d_j'(\theta) \right], \tag{5.1}$$

where $d_j(\theta) = [Y_{t_j} - \mu_j(\theta)]/\tau_j(\theta)$ and $d_j'(\theta)$ denotes its derivative (similar notation for $\tau_j'(\theta)$); we obviously assume the usual differentiability and other regularity conditions as needed (we will not repeat any such mathematical qualification in subsequent discussions unless a qualification is important for conveying a central idea). Because

$$d_j'(\theta) = -\frac{1}{\tau_j(\theta)} \left[\mu_j'(\theta) + d_j(\theta) \tau_j'(\theta) \right], \tag{5.2}$$

we see that the MLE of θ is a root of

$$\sum_{j=1}^{n} d_j(\theta) \left[\frac{\mu_j'(\theta)}{\tau_j(\theta)}\right] = \sum_{j=1}^{n} \left(1 - d_j^2(\theta)\right) \left[\frac{\tau_j'(\theta)}{\tau_j(\theta)}\right]. \tag{5.3}$$

This general expression reveals how the MLE extracts additional information from each Y_j beyond the information already captured by $\mathcal{F}_{j-1}$.

To see this clearly, let us start with $n = 1$. In such a case, (5.3) becomes

$$d_1(\theta)\mu_1^{'}(\theta) = (1 - d_1^2(\theta))\tau_1^{'}(\theta). \tag{5.4}$$

We first note that in order for the left-hand side of (5.4) to have mean zero, we only need to correctly specify the (conditional) mean $\mu_1(\theta)$. In contrast, in order for the right-hand side to have mean zero, we must also correctly specify the (conditional) variance $\tau_1^2(\theta)$ after $\mu_1(\theta)$ is correctly specified. Intuitively speaking, the left-hand side of (5.4) extracts the information in Y_1 by fitting its mean, and the right-hand side extracts *additional* information by fitting its variance.

This separation of fitting tasks is seen even more clearly when $\tau_1(\theta)$ does not depend on θ and hence fitting the variance will not provide any additional information about θ once the mean has been fitted. The estimating equation (5.4) correctly recognizes this fact via $\tau_1^{'}(\theta) = 0$, which leads to using only its left-hand side to estimate θ. Similarly, when $\mu_1^{'}(\theta) = 0$, all the information for θ will come from the variance side only, that is, the right-hand side of (5.4). Of course, if both $\mu_1^{'}(\theta)$ and $\tau_1^{'}(\theta)$ are zero, then θ is not identifiable from Y_1 alone, and (5.4) captures this non-identifiability correctly by yielding the "0=0" trivial identity. When both $\mu_1^{'}(\theta)$ and $\tau_1^{'}(\theta)$ are non-zero, the best estimator is obtained when the fitness or rather non-fitness incurred by the two sides is balanced.

Now suppose we have a second data point Y_2. The MLE approach again performs this balancing act by adding to each side the same type of non-fitness measure, but with two additional considerations. First, the mean $\mu_2(\theta)$ and the standard deviation $\tau_2(\theta)$ are not that of the marginal distribution of Y_2, but rather of the conditional distribution of Y_2 given Y_1. This makes perfect sense, because if there is any gain from having Y_2, it must come from the information that is not already captured by Y_1 (assuming we have already used up all the information in Y_1).

Second, to appropriately weight the information in Y_1 and Y_2, the derivatives $\mu_j'(\theta)$ and $\tau_j'(\theta)$ for $j = 1, 2$ need to be weighted by the scaling factor $\tau_j(\theta)$. This also makes intuitive sense because, without such a scaling factor, one would be able to arbitrarily exaggerate or diminish the information in a particular Y_j by changing its scale, which would not affect $d_j(\theta)$, but it

would affect the corresponding $\mu_j^{'}(\theta)$ and $\tau_j'(\theta)$. This fact also explains why the usual optimal linear combination of several independent estimators for the same θ is the one with inverse variance as its weight, not the inverse of standard deviation. That is, the inverse weighting by variance actually is the product of two standard deviation scalings, one for individual estimators (i.e., as in $d_j(\theta)$) and one for $\mu'(\theta)$, which takes value 1 when $\mu(\theta) = \theta$.

This scaling for appropriately measuring the accumulation of information is vivid in the Fisher information formulation. Specifically, let us define $\xi_j(\theta) = \mu_j'(\theta)/\tau_j(\theta)$, $\eta_j(\theta) = \tau_j'(\theta)/\tau_j(\theta)$ and $I_n(\theta) = -S_n'(\theta)$. Equations (5.1)-(5.2) then lead to

$$I_n(\theta) = \sum_{j=1}^{n} \left[\xi_j(\theta)\xi_j^\top(\theta) + 2\eta_j(\theta)\eta_j^\top(\theta)\right] + R_{1,n}(\theta), \tag{5.5}$$

where, suppressing the function argument θ for simplicity,

$$R_{l,m} = \sum_{j=l}^{m} \left[d_j\left(\xi_j\eta_j^\top + 2\eta_j^\top\xi_j - \xi_j'\right) - (1-d_j^2)\left(2\eta_j\eta_j^\top - \eta_j'\right)\right], \tag{5.6}$$

with ξ_j' and η_j' being matrices. By the definition of $d_j(\theta)$, we have for $j \geq 1$ and any $\theta \in \Theta$,

$$\mathrm{E}[d_j(\theta)|\mathcal{F}_{j-1}] = 0 \text{ and } \mathrm{E}[1 - d_j^2(\theta)|\mathcal{F}_{j-1}] = 0. \tag{5.7}$$

Consequently, $\mathrm{E}[R_{1,n}(\theta)] = 0$ and hence the *expected Fisher information* in $\{Y_{t_j}, j = 1, \ldots, n\}$ is (note here we use $\mathcal{I}_n$ to denote the expectation of I_n):

$$\mathcal{I}_n(\theta) = \sum_{j=1}^{n} \mathrm{E}[\xi_j(\theta)\xi_j^\top(\theta)] + 2\sum_{j=1}^{n} \mathrm{E}[\eta_j(\theta)\eta_j^\top(\theta)] \equiv \mathcal{I}_n^{(\mu)}(\theta) + \mathcal{I}_n^{(\tau)}(\theta). \tag{5.8}$$

This form of $\mathcal{I}_n(\theta)$ of course is well-known for the AR(1) model under normality. By recasting it in this more general form and by viewing n as a generic index, we see more clearly from (5.8) that the total information as measured by the expected Fisher information is accumulated via two kinds of additivity.

The first kind additivity is due to *data augmentation*[2], that is, we gain information from having more data. Here, the incremental information contained in the conditional model $Y_{t_j}|\mathcal{F}_{j-1} \sim N(\mu_j(\theta), \tau_j^2(\theta))$ is added to the

[2]The term "data augmentation" (Tanner and Wong, 1987; 2010) is also well-known in the EM and MCMC literature, where it refers to creating artificial (missing) data for the purpose of constructing useful statistical algorithms. The connection with the discussion here is that the algorithmic efficiencies of these algorithms are (almost) exactly determined by the amount of augmented Fisher information; see van Dyk and Meng (2001, 2010) for an overview and some detailed investigations.

information already measured for $\mathcal{F}_{j-1}$, namely $\mathcal{I}_{j-1}(\theta)$. This kind of additivity of course is a direct consequence of the conditional model formulation, which permits sequential input of non-redundant information. In the time series literature, this is known as a consequence of factoring the likelihood using *orthogonal errors* (see Wilson, 2001). It is trivial to see that the total information is invariant to the order we choose to factor the likelihood. (Even when our data follow a natural time order, there is no particular mathematical reason that would prohibit us from choosing another ordering for modeling. Indeed, in the general context of Markov chains, it is common to consider time-reversed chains.)

The second kind is additivity due to *model reduction*, that is, we gain information by reducing the model class via adding more restrictions. From (5.8) we see the expected Fisher information is the sum of two parts: the information from fitting the mean part $\mu_j(\theta)$, as represented by the jth term in $\mathcal{I}_n^{(\mu)}(\theta)$, and the additional information from fitting the variance part $\tau_j^2(\theta)$, as represented by the jth term in $\mathcal{I}_n^{(\tau)}(\theta)$. This second kind of additivity is less well-known, but it reflects more fundamentally how MLE extracts the maximum amount of information. As a matter of fact, if we let $S_n^{(\mu)}(\theta) = \sum_{j=1}^n d_j(\theta)\xi_j(\theta)$ and $S_n^{(\tau)}(\theta) = \sum_{j=1}^n \left(d_j^2(\theta) - 1\right)\eta_j(\theta)$, then

$$S_n(\theta) = S_n^{(\mu)}(\theta) + S_n^{(\tau)}(\theta), \tag{5.9}$$

where both $S_n^{(\mu)}(\theta)$ and $S_n^{(\tau)}(\theta)$ behave like a genuine score function. This is because $S_n^{(\mu)}(\theta)$ is the score function when we assume $\tau_j(\theta)$ is free of θ, and $S_n^{(\tau)}(\theta)$ is the score function when we assume $\mu_j(\theta)$ is free of θ. Furthermore, $S_n^{(\mu)}(\theta)$ and $S_n^{(\tau)}(\theta)$ share no redundant information because $\mathrm{Cov}(S_n^{(\mu)}(\theta), S_n^{(\tau)}(\theta)) = 0$, which holds whenever $E(d_j^3|\mathcal{F}_{j-1};\theta) = 0$ for all $j \geq 1$, a condition that is certainly satisfied under the conditional normality.

These facts imply that the information in both $S_n^{(\mu)}(\theta)$ and $S_n^{(\tau)}(\theta)$ can be measured by their corresponding expected Fisher information even when they are treated as estimating equations, because the root of $S_n^{(\mu)}(\theta) = 0$ has the asymptotic variance $\left[\mathcal{I}_n^{(\mu)}(\theta_0)\right]^{-1}$, and the root of $S_n^{(\tau)}(\theta) = 0$ has the asymptotic variance $\left[\mathcal{I}_n^{(\tau)}(\theta_0)\right]^{-1}$, where θ_0 is the true value of θ. (Note, to establish such results rigorously, we will need regularity conditions and martingale theory, neither of which is central to the key messages in the current paper.) The reason for not needing the usual "sandwich" estimator for the asymptotic variance is that $S_n^{(\mu)}(\theta) = 0$ and $S_n^{(\tau)}(\theta) = 0$ possess a necessary condition for being an optimal estimating equation in the sense

that both satisfy the second Bartlett identity (see Godambe, 1960 and 1976; Meng, 2011):

$$\text{Var}(S_n^{(\zeta)}(\theta)) = \text{E}\left[-\frac{\partial S_n^{(\zeta)}(\theta)}{\partial\theta}\right] \equiv \mathcal{I}_n^{(\zeta)}(\theta), \quad \text{for all } \theta, \tag{5.10}$$

where ζ can be either μ or τ.

In other words, under some regularity conditions (e.g., see Xie and Meng, 2012), either side of (5.3) can be used as an (optimal) estimating equation in itself for consistently estimating θ. However, the amount of information utilized by the left-hand side estimating equation $S_n^{(\mu)}(\theta) = 0$ is only $I_n^{(\mu)}(\theta_0)$, and the amount of information used by the right-hand side estimating equation $S_n^{(\tau)}(\theta) = 0$ is only $I_n^{(\tau)}(\theta_0)$. The MLE utilizes both parts of information by summing them, as in (5.9), and the additivity

$$\mathcal{I}_n(\theta) = \text{Var}(S_n(\theta)) = \text{Var}(S_n^{(\mu)}(\theta)) + \text{Var}(S_n^{(\tau)}(\theta)) = \mathcal{I}_n^{(\mu)}(\theta) + \mathcal{I}_n^{(\tau)}(\theta) \tag{5.11}$$

holds due to the aforementioned orthogonality, $\text{Cov}(S_n^{(\mu)}(\theta), S_n^{(\tau)}(\theta)) = 0$.

An important implication of the above results is that, although we invoked the conditional normality in deriving the initial score function equation (5.3) or equivalently (5.9), the resulting estimating equations behave coherently in the sense of preserving the aforementioned two kinds of additivities under a much weaker assumption, namely, the conditional skewness is zero, $E(d_j^3|\mathcal{F}_{j-1};\theta) = 0$ (note technically this is even weaker than requiring $p(Y_j|\mathcal{F}_{j-1};\theta)$ be symmetric, though for many practical purposes, it is difficult to achieve zero skewness without symmetry).

The preservation of the information additivity over data augmentation ensures us that as we collect more data the efficiency of our estimate cannot decrease. Similarly, the preservation of information additivity under model reduction guarantees the same is true when the added assumptions are valid. Estimating equation (5.9) therefore is an appealing general estimating equation when only the first two conditional moments are specified. For example, when additional assumptions are made about θ, they will be reflected in the derivatives $\mu_j'(\theta)$ and $\tau_j'(\theta), j = 1, \ldots, n$. Because (5.9) acts coherently just as a real score function under the (working) conditionally normal model, the additional information from the added assumptions will be at least partially realized and properly reflected in the resulting expected Fisher information, instead of unintentionally causing damage as in the correlation example of Section 4.

6. How Much More Can We Learn From the AR(1) Model?. The AR(1) model with normal error appears to be the simplest non-trivial

example (the so-called "S-NoTE") in the context of time series modeling, and the relevant literature is enormous in both econometrics and statistics. Early classic work in statistics include White (1958) and Anderson (1959), studying the limiting distributions of least square estimators for the AR(1) models. A good number of later developments were listed (and some were reviewed) in Chan (2006), including generalizations to multivariate cases (e.g., Tsay and Tiao, 1990). An excellent reference book containing much of the advanced theory on AR(1) and much more is Tanaka (1996). Most recent examples include Chan and Ing's (2011) results on a uniform bound for the inverse Fisher information matrix, as well as an intriguing result on predictive error by Yu, Lin and Cheng (2012). On the econometrics side, much work has been devoted to unit-root AR(1) models and related problems, for example the work by Phillips (1987) and Phillips and Perron (1988) developing a number of tests for the unit root. The applied interest in AR(1) and related models is highlighted by an entire issue of *Journal of Applied Econometrics* (1991, October/December issue) devoted to discussions on using AR(1) models for GNP growth and related analyses.

The AR(1) model is also one of the simplest (continuous) Markov chains, defined by

$$Y_t = \rho Y_{t-1} + \epsilon_t, \quad \epsilon_t \overset{\text{iid}}{\sim} N(0, \sigma^2), \quad t = 1, \ldots, T, \tag{6.1}$$

where Y_0 is often set to zero but, in a finite-time horizon framework, it would be more appropriate to treat it as a parameter. Typically we are interested in estimating ρ, with σ^2 (and Y_0) treated as a nuisance parameter. Under this setup, it is well-known that for any $t > s \geq 0$,

$$Y_t | Y_s \sim N(\rho^{t-s} Y_s,\ k_{t-s}(\rho)\sigma^2), \text{ with } k_l(\rho) = \sum_{j=0}^{l-1} \rho^{2j}. \tag{6.2}$$

To apply (5.8) with $Z_{obs} = \{Y_{t_j}, j = 1, \ldots, n\}$, we know from (6.2) that $\mu_j(\theta) = \rho^{s_j} Y_{t_{j-1}}$ and $\tau_j^2(\theta) = k_{s_j}(\rho)\sigma^2$, where $s_j = t_j - t_{j-1}, j = 1, \ldots, n$. Consequently, for $\theta = \{\rho, \sigma^2, Y_0\}, \xi_j(\theta) = (s_j \rho^{s_j - 1} Y_{t_{j-1}}, 0, \rho^{t_1} 1_{\{j=1\}})^\top / \tau_j(\theta)$. It then follows

$$\mathcal{I}_n^{(\mu)}(\theta) = \frac{1}{\sigma^2} \begin{pmatrix} Y_0^2 A_{1,n}(\rho) + \sigma^2 B_{2,n}(\rho) & 0 & t_1 \gamma_1(\rho) Y_0 \\ 0 & 0 & 0 \\ t_1 \gamma_1(\rho) Y_0 & 0 & \rho \gamma_1(\rho) \end{pmatrix}, \tag{6.3}$$

where $A_{\ell,n}(\rho) = \sum_{j=\ell}^n \alpha_j(\rho), B_{\ell,n}(\rho) = \sum_{j=\ell}^n \beta_j(\rho)$, with

$$\alpha_j(\rho) = \frac{s_j^2 \rho^{2(t_j - 1)}}{k_{s_j}(\rho)}, \quad \beta_j(\rho) = \frac{s_j^2 \rho^{2(s_j - 1)} k_{t_{j-1}}(\rho)}{k_{s_j}(\rho)}, \ \gamma_1(\rho) = \frac{\rho^{2t_1 - 1}}{k_{t_1}(\rho)}. \tag{6.4}$$

Here we have used the fact that $\mathrm{E}(Y_{t_{j-1}}^2) = \rho^{2(t_{j-1})}Y_0^2 + k_{t_{j-1}}(\rho)\sigma^2$ with the convention that $k_0(\rho) \equiv 0$. Similarly, because $\eta_j(\theta) = 0.5(\delta_j(\rho), 1/\sigma^2, 0)^\top$, where $\delta_j(\rho) = k'_{s_j}(\rho)/k_{s_j}(\rho), j = 1, \ldots, n$, we have

$$\mathcal{I}_n^{(\tau)}(\theta) = \frac{1}{2\sigma^4} \begin{pmatrix} \sigma^4 \sum_{j=1}^n \delta_j^2(\rho) & \sigma^2 \sum_{j=1}^n \delta_j(\rho) & 0 \\ \sigma^2 \sum_{j=1}^n \delta_j(\rho) & n & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{6.5}$$

From these expressions, we can examine many factors that affect the amount of information in the data or in our model assumptions. Such investigations are useful for both data collection and data analysis. For example, in the context of collecting future data we may wonder if we should take a daily observation for 30 consecutive days or we should take one observation every other day for 60 day period, assuming we can afford waiting for 60 days but no more than 30 observations (and that the same model is applicable to the longer period). Or for analyzing historical data with missing observations, we may wonder how much information is lost due to the particular pattern of the missing observations, and how much of the lost information can be compensated by introducing assumptions on the nuisance parameters (perhaps from other studies in the literature)?

Before we answer any of such questions, we emphasize that, although the expressions $\mathcal{I}_n^{(\mu)}(\theta)$ and $\mathcal{I}_n^{(\tau)}(\theta)$ are valid for any value of ρ, using the *expected* Fisher information $\mathcal{I}_n(\theta) = \mathcal{I}_n^{(\mu)}(\theta) + \mathcal{I}_n^{(\tau)}(\theta)$ to determine the asymptotic variance for the MLE is valid only when $|\rho| < 1$. (But for such asymptotics to work well, one should also avoid near-unit root cases, such as those described in Chan, 1988.) As it is well-known (see Tanaka, 1996), for $|\rho| \geq 1$, the usual normal asymptotics fail and hence a more involved calculation is needed. In the current paper, we will focus on the stationary case with $|\rho| < 1$, which already provides a rich setting to investigate the questions raised above.

6.1. *Is there as interaction between model assumptions and data patterns?.* As in Section 4, let i_{st} and $i^{(st)}$ be the $\{s,t\}$th element of $\mathcal{I}_n(\theta)$ and of $\mathcal{I}_n^{-1}(\theta)$, respectively. Using the fact that $i_{23} = 0$ and $\mathcal{I}_n(\theta)$ is symmetric, we have, as a generalization of (4.2),

$$\mathcal{I}_n(\rho) \equiv [i^{(11)}]^{-1} = i_{11} - \frac{i_{12}^2}{i_{22}} - \frac{i_{13}^2}{i_{33}} < i_{11} = \mathcal{I}_n(\rho|\sigma^2, Y_0), \tag{6.6}$$

where $\mathcal{I}_n(\rho|\sigma^2, Y_0)$ denotes the expected Fisher information for ρ when we assume both σ^2 and Y_0 are known. Furthermore, here i_{12}^2/i_{22} measures the (absolute) gain in information for estimating ρ by conditioning on σ^2 (i.e.,

treating it as known), and i_{13}^2/i_{33} measures the gain due to conditioning on Y_0. That is, using the conditional information notation, we have

$$\mathcal{I}_n(\rho|Y_0) = i_{11} - \frac{i_{12}^2}{i_{22}}; \quad \mathcal{I}_n(\rho|\sigma^2) = i_{11} - \frac{i_{13}^2}{i_{33}}. \tag{6.7}$$

As in Section 4, we let $\mathcal{G}_n(\rho|K) = \mathcal{I}_n(\rho|K) - \mathcal{I}_n(\rho)$ be the gain in information due to the added knowledge K. Then from (6.6)-(6.7) we see that the two gains are additive:

$$\mathcal{G}_n(\rho|\sigma^2, Y_0) = \mathcal{G}_n(\rho|\sigma^2) + \mathcal{G}_n(\rho|Y_0). \tag{6.8}$$

This is because $i_{23} = 0$, or in Bayesian terms, because σ^2 and Y_0 are conditionally independent (asymptotically) given ρ, in the same spirit as with the relationship between the relative gain in information and Bayesian posterior correlation described by (4.4). Intuitively this makes good sense because if ρ is known, then knowing the starting value Y_0, which is the same as knowing the mean, tells us little about the residual variance σ^2, and vice versa. Consequently, the information obtained from knowing Y_0 and σ^2 has no redundancy, leading to the additivity in (6.8).

The practical relevance of these results is that they can provide useful insights and general guidelines on what gain or loss of information is important and for what (missing) data patterns. For example, suppose our observed data are a *consecutive segment* $Z_{obs} = \{Y_t, \ldots, Y_{t+n-1}\}$, for which $s_j = 1, j = 2, \ldots, n$ and $s_1 = t$. Then, $k_{s_j}(\rho) = 1$ for $j \geq 2$ and $k_1(\rho) = (1-\rho^{2t})/(1-\rho^2)$. It follows then (recall $\delta_j(\rho) = k'_{s_j}(\rho)/k_{s_j}(\rho)$)

$$\mathcal{R}_n(\rho|\sigma^2) \equiv \frac{\mathcal{G}_n(\rho|\sigma^2)}{\mathcal{I}_n(\rho)} = \frac{\frac{1}{2n}\delta_1^2(\rho)}{\frac{Y_0^2}{\sigma^2}A_{2,n}(\rho) + B_{2,n}(\rho) + \frac{1}{2}(1-\frac{1}{n})\delta_1^2(\rho)} < \frac{1}{n-1},$$

where the inequality holds because both $A_{2,n}(\rho)$ (note it is not $A_{1,n}(\rho)$) and $B_{2,n}(\rho)$ are positive. Therefore, knowing the value of σ^2 essentially does not help the estimation of ρ; indeed for fixed t, the relative gain $\mathcal{R}_n(\rho|\sigma^2)$ goes to zero at the n^{-2} rate.

In contrast, consider cases where the observations are not consecutive and there is a sufficient amount of gaps among them. That is, let $\mathcal{J}_n = \{j \geq 2 : s_j > 1\}$, and we assume $r_n = |\mathcal{J}_n|/n$ approaches $r > 0$ as n increases, where $|\mathcal{J}_n|$ is the cardinality of $\mathcal{J}_n$. By the definition of $\delta_j(\rho)$, it is non-zero if and only if $j \in \mathcal{J}_n$. Let $\bar{\delta}_{\mathcal{J}_n}(\rho)$ be the sample average of $\{\delta_j(\rho), j \in \mathcal{J}_n\}$, and $V_{\mathcal{J}_n}(\rho)$ be its sample variance (with denominator $|\mathcal{J}_n|$). Then simple algebra yields

$$\mathcal{R}_n(\rho|\sigma^2) = \frac{0.5 r_n^2 [\bar{\delta}_{\mathcal{J}_n}(\rho)]^2}{\frac{Y_0^2}{\sigma^2}\frac{A_{2,n}(\rho)}{n} + \frac{B_{2,n}(\rho)}{n} + 0.5\left\{r_n(1-r_n)[\bar{\delta}_{\mathcal{J}_n}(\rho)]^2 + r_n V_{\mathcal{J}_n}(\rho)\right\}}.$$

Because for any $\ell \geq 1$

$$A_{\ell,n}(\rho) = \sum_{j=\ell}^{n} \frac{s_j^2 \rho^{2(t_j-1)}}{k_{s_j}(\rho)} < \sum_{t=1}^{\infty} t^2 \rho^{2(t-1)},$$

which is a converging series when $\rho^2 < 1$, we know $A_{2,n}(\rho)/n \to 0$ for any $|\rho| < 1$. Using the fact that $k_{t_{j-1}}(\rho) = (1-\rho^{2t_{j-1}})/(1-\rho^2)$, we can write

$$\frac{B_{2,n}(\rho)}{n} = \frac{1}{1-\rho^2}\left[(1-r_n) + r_n \bar{C}_{\mathcal{J}_n}(\rho) - \frac{A_{2,n}(\rho)}{n}\right], \tag{6.9}$$

where $\bar{C}_{\mathcal{J}_n}(\rho)$ is the sample average of $\{C_j(\rho) = \frac{s_j^2 \rho^{2(s_j-1)}}{k_{s_j}(\rho)}, j \in \mathcal{J}_n\}$. Note that $C_j(\rho) \leq [e\rho \log(\rho)]^{-2}$ for all values of s_j, and hence $\bar{C}_{\mathcal{J}_n}(\rho)$ must be bounded above regardless of the choices of the s_j's.

Combining the results above, we see that as n increases, the relative gain by assuming knowledge of σ^2 approaches the limit

$$\mathcal{R}_\infty(\rho|\sigma^2) = \frac{0.5 r^2 [\bar{\delta}_{\mathcal{J}}(\rho)]^2}{\frac{1}{1-\rho^2}[1-r+r\bar{C}_{\mathcal{J}}(\rho)] + 0.5\left\{r(1-r)[\bar{\delta}_{\mathcal{J}}(\rho)]^2 + r V_{\mathcal{J}}(\rho)\right\}},$$

where any quantity with the $\mathcal{J}$ subscript is the limit of the same quantity with the $\mathcal{J}_n$ subscript (assuming the limit exists of course). Here we see that the relative gain will no longer be negligible once we have an appreciable amount of gaps. Qualitatively this is naturally expected from (6.2) because, as soon as $s > 1$, both ρ and σ^2 enter the conditional variance function $\tau_j^2(\theta)$. Hence, knowledge of σ^2 helps to better fit the value of ρ via fitting $\tau_j^2(\theta)$ because its dependence on the unknown σ^2 is eliminated (and here we are using MLE as the estimation method, not an ad hoc "partial plug-in").

To see how the relative gain from the same information about σ^2 depends critically on the observed data patterns, let us consider the case where we sample every $s(\geq 2)$ observations, that is $s_j = s$ for all $j \geq 2$. Then $r = 1$ and $V_{\mathcal{J}}(\rho) = 0$, hence

$$\mathcal{R}_\infty^{(s)}(\rho|\sigma^2) = \frac{(1-\rho^2)[\bar{\delta}_{\mathcal{J}}(\rho)]^2}{2\bar{C}_{\mathcal{J}}(\rho)} = \frac{2[1-s\rho^{2(s-1)} + (s-1)\rho^{2s})]^2}{s^2(1-\rho^2)^2(1-\rho^{2s})\rho^{2(s-2)}}. \tag{6.10}$$

Here we index the limit of $R_n(\rho|\sigma^2)$ explicitly by the superscript s to emphasize the interaction between the sample design and model assumption. That is, the knowledge about σ^2 leads to very different relative gain in information depending on the sampling interval even with the same size fixed. For example, when $s = 2$ and $s = 3$, we have

$$\mathcal{R}_\infty^{(2)}(\rho|\sigma^2) = \frac{1-\rho^2}{2(1+\rho^2)}; \quad \text{and} \quad \mathcal{R}_\infty^{(3)}(\rho|\sigma^2) = \frac{2(1-\rho^2)(1+2\rho^2)^2}{9\rho^2(\rho^4+\rho^2+1)}. \tag{6.11}$$

These expressions remind us that interaction between data patterns and model assumptions is an intricate issue even for such a seemingly simple situation. We see that, for $s = 2$, the relative gain is bounded with the gain ranging from 0% when ρ^2 approaches one to 50% when ρ^2 approaches zero. This corresponds to a reduction of asymptotic variance from 0% to 33.3% (because the the percentage of variance reduction is given by $\mathcal{R}_\infty/(1+\mathcal{R}_\infty)$). However, although for $s = 3$ the reduction is also a monotone decreasing function of ρ, the range now increases to $(0, \infty)$, resulting in variance reduction of potentially 100% as ρ approaches zero.

6.2. *What is the optimal spacing?.* The above calculation also reveals that the optimal sampling interval for estimating ρ increases with ρ when the sample size n is fixed. Qualitatively, this is expected because as ρ increases, consecutive observations become increasingly similar to each other and hence, for a given sample size n, the effective sample size decreases. Increasing the sampling interval then helps to combat this problem, for the very same reason that "thinning" in Markov chain Monte Carlo is useful when one can process only a fixed number of draws (see Gelman and Shirley, 2011). The Fisher information calculation helps us to see quantitatively how the optimal spacing increases with the value of ρ. For example, it is not obvious how large $|\rho|$ must be before the optimal spacing jumps from $s = 1$ to $s = 2$.

To simplify the derivation, we consider large-n approximation, in which case the same calculation above yields

$$\lim_{n\to\infty} \frac{\mathcal{I}_n^{(s)}(\rho)}{n} = \frac{\bar{C}_{\mathcal{J}}(\rho)}{1-\rho^2} = \frac{s^2\rho^{2(s-1)}}{1-\rho^{2s}} \equiv H(s,\rho), \tag{6.12}$$

which recovers the well-known case for $s = 1$, namely, the asymptotic variance of $\hat{\rho}$ is $(1 - \rho^2)/n$ with consecutive observations. To maximize $H(s, \rho)$ for a given ρ, we let $x = -s \log \rho^2$, which yields

$$H(s(x),\rho) = [\rho \log \rho^2]^{-2} \frac{x^2}{e^x - 1}. \tag{6.13}$$

But $x^2/(e^x - 1)$ has the global maximizer at the x that satisfies $(2 - x)e^x = 2$ or equivalently $(x - 2)e^{x-2} = -2e^{-2}$; hence $x_{max} = 2 + W(-2e^{-2}) \approx 1.59362$, where W is the so-called Lambert W function, the inverse function of $f(W) = We^W$. The maximum of $x^2/(e^x - 1)$ is $x_{max}(2 - x_{max}) \approx 0.64761$.

Consequently, for any given ρ, $H(s, \rho)$ is maximized at

$$s_{max}(\rho) = \frac{x_{max}}{-\log \rho^2}.$$

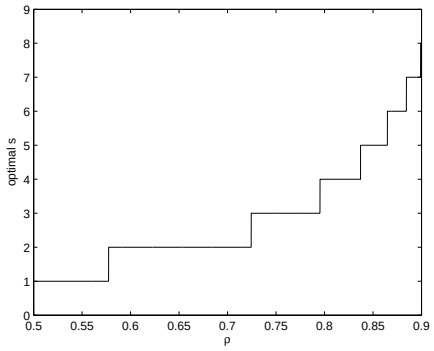

FIG 1. *Optimal spacing.*

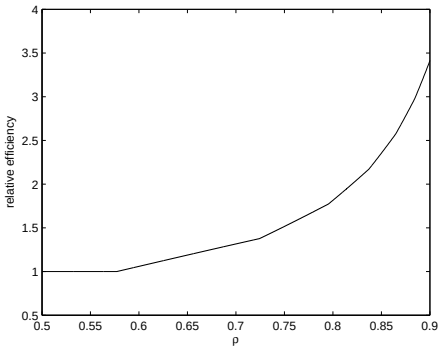

FIG 2. *Relative efficiency.*

It is also easy to verify that $H(s,\rho)$ is monotone increasing for $s < s_{max}(\rho)$ and monotone decreasing after $s > s_{max}(\rho)$.

However, since for our current setting s can only take positive integer values, the solution for the integer maximization is a bit more involved. Evidently, when $s_{max}(\rho) \leq 1$, the optimal integer spacing must be $s_{opt}(\rho) = 1$. But $s_{opt}(\rho)$ remains equal to one as long as $H(1,\rho) > H(2,\rho)$. In general, because for any given positive integer s,

$$R_s(\rho) \equiv \frac{H(s+1,\rho)}{H(s,\rho)} = (1+s^{-1})^2 \left[1 - \left(\sum_{i=0}^{s} \rho^{2i}\right)^{-1}\right] \tag{6.14}$$

is a strictly monotone increasing function of ρ^2, we see that if we find $\rho_s (\geq 0)$ such that $R_s(\rho_s) = 1$, then $\{\rho_1, \rho_2, \ldots..\}$ forms the sequence of cut-off points such that the optimal spacing is $s_{opt}(\rho) = s$ whenever $\rho_{s-1}^2 \leq \rho^2 \leq \rho_s^2$, where $\rho_0 = 0$. This is simply because inside such a range $H(s,\rho)$ dominates both $H(s+1,\rho)$ and $H(s-1,\rho)$, and the unimodality of $H(s,\rho)$ as a function of s then establishes our assertion. The first couple of values of ρ_s are easy to obtain from setting $R_s(\rho) = 1$ for $s = 1$ and 2, which give $\rho_1^2 = 1/3$ and $\rho_2^2 = (\sqrt{105} - 5)/10$, yielding $\rho_1 = \pm 0.577$ and $\rho_2 = \pm 0.724$. The rest can be obtained easily numerically. Figure 1 plots the optimal spacing $s_{opt}(\rho)$ as a function of ρ for $\rho \in [0.5, 0.9]$, and Figure 2 plots the corresponding gain in efficiency by using the optimal spacing relative to using $s = 1$.

We see that the maximal gain in efficiency can be quite large, and in fact it goes to infinity at the rate of $0.64761 \log^{-1}(\rho^{-2})$ as $\rho^2 \to 1$, as can be verified from (6.12)-(6.13). Note in this verification we have used the fact

that because $s_{max}(\rho) \in [s-1, s]$ for any $\rho^2 \in [\rho^2_{s-1}, \rho^2_s]$, we have

$$\lim_{\rho^2 \to 1} \frac{s_{opt}(\rho)}{s_{max}(\rho)} = 1 \quad \text{and} \quad \lim_{s \to \infty} |\rho^2_s - \rho^2_{s-1}| = 0.$$

The diminishing of the gap $\rho_s - \rho_{s-1}$, which can be verified as with the rate of s^{-2}, is quite visible from Figure 1.

Of course in real applications we do not know the value of ρ before we estimate it. This is the usual trouble with mathematical optimality results for experimental designs (and beyond), because the optimal design necessarily depends on the estimand we are after. Nevertheless, if one is interested in finding an approximately economical sampling plan for a given budget (e.g., the sample size n), then the above results do provide useful guidelines. It is not unreasonable to assume that in many practical situations we would have a rough idea before collecting data about the magnitude of the autocorrelations, for example, small versus large. For instance, if our prior knowledge tells us that it is almost certain that $|\rho|$ is below say 0.5, then $s = 1$ is the choice. On the other hand, if one is suspecting a near unit-root phenomenon, say $|\rho| \geq 0.8$, then $s = 3$ is a better choice. Although one should not expect the optimal gain, the minimal gain is often substantial, as shown in Figure 3, where the vertical axis is on the $\log_{10}$ scale. Hence even an ordinate of 0.25, as is the case at about $\rho = 0.8$, corresponds to more than 75% gain in efficiency. More precisely, if we use $s = 3$, then the gain will be anywhere from 80% to 300% as ρ moves from 0.8 to 1. In general, it is easy to derive from (6.12) that

$$\lim_{\rho^2 \to 1} \frac{H(s, \rho)}{H(1, \rho)} = s, \tag{6.15}$$

providing a handy rule of thumb for possible improvement near the unit root.

We remark here that in some applications the setup of letting s grow but fixing n may be considered irrelevant because the total time horizon N is moderate or even small, e.g., there were only 64 quarterly observations available. Consequently, the constraint $ns = N$ means that in order to increase s we must decrease n. In such cases, the relevant calculation is how much information is lost by *sub-sampling*. One important fact is that although sub-sampling a consecutive sequence always leads to loss of information under the MLE approach, a sub-sample with a larger size may not necessarily dominate one with a smaller size because the latter may not be nested within the former. For example, taking observations every third quarter is not a sub-sampling of taking observations every other quarter.

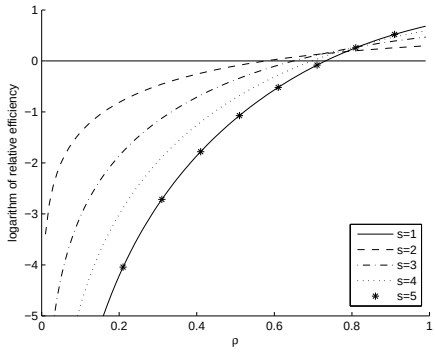

FIG 3. *The logarithm of relative efficiency.*

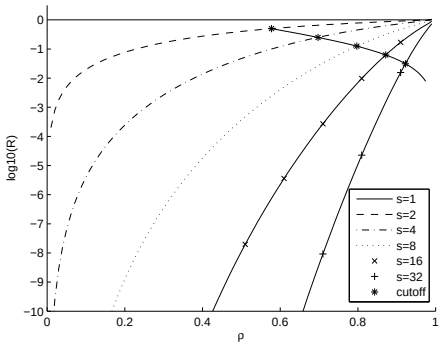

FIG 4. *The logarithm of* $R_K(L,\rho)$.

Even when the sub-samples are nested (and hence there is no surprise as far as "negative information" goes), the Fisher information calculation can still reveal phenomena that might not be completely expected. For example, for estimating the autocorrelation ρ in an AR(1) model, how much efficiency is lost if we reduce 64 quarterly observations to 32 by sub-sampling every other quarter? The answer will depend on the value of ρ.

To see this clearly, consider a simple case where $Y_0 = 0$, $N = 2^K$, and $s = 2^L$, and hence $n = 2^{K-L}$ and the sub-samplings are nested as we increase L. By (5.8), (6.2), and (6.6), we obtain

$$\mathcal{I}_n^{(s)}(\rho) = \frac{s^2\rho^{2(s-1)}}{1-\rho^{2s}}\left[n - \frac{1-\rho^{2N}}{1-\rho^{2s}}\right]. \tag{6.16}$$

Consequently, under the constraint $ns = N$, the relative information in using a spacing s compared with using $s = 1$ is given by

$$R_K(L,\rho) \equiv \frac{\mathcal{I}_{N/s}^{(s)}(\rho)}{\mathcal{I}_N^{(1)}(\rho)} = \frac{\mathcal{I}_{2^{K-L}}^{(2^L)}(\rho)}{\mathcal{I}_{2^K}^{(1)}(\rho)}.$$

Figure 4 displaying $\log_{10}[R_K(L,\rho)]$ for $K = 10, L = 1, ..., 5$, and $0 < \rho \leq 1$. We see that $R_K(L,\rho)$ gets closer to 1 when ρ approaches 1. (Figure 4 shows only the $\rho \geq 0$ case, because $R_k(L,\rho)$ is symmetric about $\rho = 0$.) Indeed, it is not hard to show that in general

$$\lim_{\rho^2 \to 1} R_K(L,\rho) = \frac{N-s}{N-1} = \frac{n-1}{n-s^{-1}},$$

which is far from the usual n/N ratio when the data are i.i.d. The "cut-off" line in Figure 4 corresponds to the ρ at which $R_k(L,\rho) = n/N = s^{-1}$.

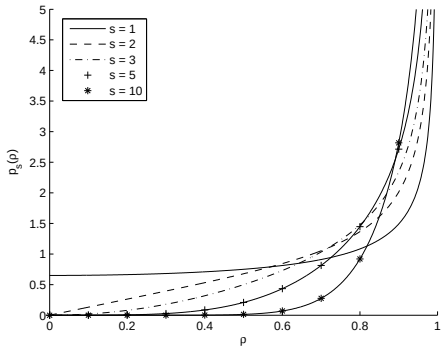

FIG 5. *Jeffreys prior density for ρ, as a functional of spacing s.*

We therefore learn that as long as n is not too small ($n > 30$), the loss of information by using a sub-sample is minor when ρ is close to a unit root. Figure 4 also reveals that the loss is very substantial if we sub-sample too much and ρ is not near a unit root.

7. Is the Jeffreys Prior Really Non-informative?. The Fisher information calculation above also reveals an intriguing phenomenon regarding the Jeffreys prior for ρ, adding to a set of well-known complications of "non-informative" priors for the AR(1) model (e.g., Berger and Yang, 1994; Uhlig, 1994a). Specifically, a consequence of (6.12) is that the Jeffreys prior $I_n^{1/2}(\rho)$ (for large n) amounts to assigning ρ^{2s} a $Beta(1/2, 1/2)$ distribution, given $\rho \geq 0$. It follows then that for any $0 \leq a < 1$, the prior CDF $F_s(a)$ converges to zero because

$$F_s(a) \equiv Pr(\rho^{2s} \leq a^{2s}) = F_{0.5,0.5}(a^{2s}) \to 0, \quad \text{as } s \to \infty$$

where $F_{0.5,0.5}$ is the CDF of $Beta(1/2, 1/2)$. Hence the Jeffreys prior converges to the point mass at $\rho = 1$, hardly noninformative by any measure. Note here the convergence is rather rapid at the exponential rate a^s (but not uniformly), because $\lim_{s\to\infty} F_s(a)/[2a^s/\pi] = 1$. Even for $s = 2$, the prior median of ρ is $0.5^{1/4} = 0.84$, and hence the Jeffreys prior strongly prefers values closer to the unit root than those that are away from it. Figure 5 shows how skewed the prior densities are even for modest spacing s.

The Jeffreys prior has often been used in the literature because it is associated with "objective Bayesian analysis" (e.g., Phillips, 1991). Putting aside the philosophical quibble whether a meaningful "objective analysis" is ever possible (e.g., Berger and Yang, 1994), we can see from the above

example that the Jeffreys prior can certainly put strong preference over one region of the parameter space than another, in sharp contrast to a layman's understanding of the concept of "objective prior" or non-informative prior. In the context of autoregressive models, this phenomenon is well-known and has been much debated, but primarily in the case of explosive region or at least (near) unit-root cases, that is, when $|\rho| \geq 1$ (see Uhlig, 1994a, b; and Kass and Wasserman, 1996, and the references therein). Our example shows that the same phenomenon exists even within the stationarity region $|\rho| < 1$, once we allow ourselves to go beyond consecutive sampling.

We emphasize, however, that although we use (6.12) to directly assign a univariate prior for ρ, because (6.12) measures the *marginal information* for ρ, the prior specified above for ρ is (asymptotically) equivalent to the marginal prior for ρ derived from the joint Jeffreys prior on (ρ, σ^2). Had we used the *conditional* Jeffreys prior $p(\rho|\sigma^2)$ pretending σ^2 is known, then the Fisher information will not strongly prefer $\rho = 1$ even as $s \to \infty$ because the information from fitting the variance part of AR(1) for estimating $|\rho| < 1$ will not be consumed by the need for estimating σ^2, since it is already (assumed) known. In this sense, one could attribute our finding as another example of the unreliability of multivariate Jeffreys prior (Kass and Wasserman, 1996). But as also noted in Kass and Wasserman (1996, Section 3.5), even in the case of univariate ρ (i.e., σ^2 is known), the Jeffreys prior puts too much weight in regions that correspond to nonstationarity. Similarly, Berger and Yang (1994) reported the difficulties with finding "objective" prior for ρ "even in the comparatively simple case of known σ^2."

Our feeling is that a key issue with the use of the Jeffreys prior lies in how the Fisher information is determined by the patterns of the data, not merely the size of the data. In general the use of data-dependent priors has been strongly discouraged other than for certain specific theoretical purposes (e.g., Wasserman, 2000, Mukerjee, 2008). In contrast, the dependence of the Jeffreys prior on the data pattern, which clearly is an important aspect of the *observed data*, has not received the same treatment (but see Kass and Wasserman's (1996, Section 3.5) emphasis on how the Jeffreys prior depends on the sample space). There may well be theoretically justifiable and practically useful principles or at least guidelines to tell us what aspects of the observed data can be used for constructing priors, e.g., such as when constructing "weakly informative" priors (Gelman *et. al.* 2008), where "weak" is relative to the information in the data (under a specified likelihood). But to blindly trust the conventional wisdom that the Jeffreys prior is an "objective" prior requires one to overlook the common meaning of the phrase "objective."

We conclude with this example to re-iterate the overall message of our paper. That is, serious statistical inference is an enterprise involving science, engineering, and even a bit of art. As such, it is virtually always wise to integrate good intuition and conventional wisdom with critical theoretical thinking and rigorous mathematical derivations whenever feasible. This integration is critical to ensure that our conclusions are not only scientifically defensible but also the best possible ones given our limited resources, and minimally it prevents us from unknowingly producing statistical results that are seriously inferior to what we expect them to be.

Acknowledgements. We thank Editor Ehsan Soofi for the invitation (and for his extraordinary patience) to contribute to this special volume in honor of Professor Arnold Zellner, who was a colleague and friend of one of us (Meng) during his Chicago years (1991-2001). We also thank many colleagues, especially Joseph Blitzstein, Ngai Hang Chan and Ehsan Soofi for very helpful exchanges and conversations, Alex Blocker, Steven Finch and Nathan Stein for proofreading and constructive comments, and the National Science Foundation for partial financial support.

References

Abel, P. S. and Singpurwalla N. D. (1994). To survive or to fail: That is the question. *The American Statistician* **48**, 18-21.

Anderson, T.W. (1959). On asymptotic distributions of estimates of parameters of stochastic difference equations. *The Annals of Mathematical Statistics* **30**, 676-687.

Berger, J. O. and Yang, R.-Y. (1994). Noninformative priors and Bayesian testing for AR(1) model. *Econometric Theory* **10**, 461-482.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* **31**, 307-327.

Chan, N. G. (1988). The parameter inference for nearly nonstationary time series. *Journal of the American Statistical Association* **83,** 857-862.

Chan, N. G. (2006). Inference for time series and stochastic process. *Statistica Sinica* **16**, 683-696.

Chan, N. H. and Ing, C. K. (2011). Uniform moment bounds of Fisher's information with applications to time series, *The Annals of Statistics* **39**, 1526-1550.

Ebrahimi, N., Soofi, E. S. and Soyer, R. (2008). Multivariate maximum entropy identification, transformation, and dependence. *Journal of Multivariate Analysis* **99**, 1271-1231.

Ebrahimi, N., Soofi, E. S. and Soyer, R. (2010). On the sample information

about parameter and prediction. *Statistical Science* **25**, 348-367.

Ebrahimi, N., Soofi, E. S. and Soyer, R. (2012). When is failure preferable to survival. *Submitted*.

Efron, B. and Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika* **65**, 457-487.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* **50,** 987-1007.

Ferguson, T. G. (1996). *A Course in Large Sample Theory.* Chapman & Hall, London.

Gelman, A. and Shirley, K. (2011). Inference from simulations and monitoring convergence. In *Handbook of Markov Chain Monte Carlo* (Eds: Brooks, Gelman, Jones, and Meng). CRC Press, Boca Raton.

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics* **2**, 1360-1383.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics* **31**, 1208-1211.

Godambe, V. P. (1976). Conditional likelihood and unconditional optimum estimating equations. *Biometrika* **63**, 277-84.

Jeffreys, H. (1983). *Theory of Probability.* Oxford University Press.

Kass, R. E. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association* **91**, 1343-1370

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input (with discussions). *Statistical Science* **9**, 538-573.

Meng, X.-L. (2005). From unit root to Stein's estimator to Fisher's k statistics: If you have a moment, I can tell you more. *Statistical Science* **20**, 141-162.

Meng, X.-L. (2009). Decoding the H-likelihood. *Statistical Science* **24**, 280-293.

Meng, X.-L. (2011). What's the H in H-likelihood: A Holy Grail or an Achilles' Heel? (with discussions). In *Bayesian Statistics 9* (Eds: Bernardo, Bayarri, Berger, Dawid, Heckerman, Smith and West), 473-500. Oxford University Press.

Meng, X.-L. and Zaslavsky, A. M. (2002). Single observation unbiased priors. *Annals of Statistics* **30**, 1345-1375.

Mukerjee, R. (2008). Data-dependent probability matching priors for empirical and related likelihoods. In *Pushing the Limits of Contemporary*

Statistics: Contributions in Honor of Jayanta K. Ghosh (Eds: Clarke and Ghosal), 60-70. Institute of Mathematical Statistics, Beachwood, Ohio.

Nyquist, H. (1928). Certain topics in telegraph transmission theory. *Trans. AIEE* **47**, pp. 617-644. (Reprint as classic paper in: *Proc. IEEE*, **90**, No. 2, Feb 2002.)

Pantula, S. G. (1988). Estimation of autoregressive models with ARCH errors. *Sankhya* **50**, 119-138.

Phillips, P. C. B. (1987). Testing for a unit root in time series regression. *Econometrica* **55**, 277-301.

Phillips, P. C. B (1991). To criticise the critics: An objective Bayesian analysis of stochastic trends. *Journal of Applied Econometrics* **6**, 333-364.

Phillips, P. C. B and P. Perron (1988). Testing for a unit root in time series regression. *Biometrika* **75**, 335-346.

Rubin, D. B. (1987). *Multiple imputation for non-response in Surveys.* Wiley, New York.

Tanaka, K. (1996). *Time Series Analysis – Nonstationary and Noninvertible Distribution Theory.* Wiley, New York.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* **82**, 528–540.

Tanner, M. A. and Wong, W. H. (2010). From EM to data augmentation: the emergence of MCMC Bayesian computation in the 1980s. *Statistical Science* **25**, 506–516.

Tsay, R. S. (2001). Nonlinear time series models: testing and applications. In *A Course in Time Series Analysis* (Eds: Pena, Tiao and Tsay), 267-285. Wiley, New York.

Tsay, R. S. and Tiao, G. C. (1990). Asymptotic properties of multivariate nonstationary processes with applications to autoregressions. *The Annals of Statistics* **18**, 220-250.

Uhlig, H. (1994a). What macroeconomists should know about unit roots: A Bayesian perspective. *Econometric Theory* **10**, 645-671.

Uhlig, H. (1994b). On Jeffreys prior when using the exact likelihood function. *Econometric Theory* **10**, 633-644.

Unser, M. (2000). Sampling–50 years after Shannon. *Proceedings of the IEEE* **88**, pp. 569-587.

van Dyk, D. A. and Meng, X.-L. (2001). The art of data augmentation (with discussion). *Journal of Computational and Graphical Statistics* **10**, 1–111.

van Dyk, D. A. and Meng, X.-L. (2010). Cross-fertilizing strategies for better

EM mountain climbing and DA field exploration: A graphical guide book. *Statistical Science* **25**, 429–449.

Wasserman, L. (2000). Asymptotic inference for mixture models using data-dependent priors. *Journal of the Royal Statistical Society B* **62**, 159-180.

White, J. S. (1958). The limiting distribution of the serial correlation coefficient in the explosive case. *The Annals of Mathematical Statistics* **29**, 1188-1197.

Wilson, G. T. (2001). Model fitting and checking, and the Kalman filter. In *A Course in Time Series Analysis* (Eds: Pena, Tiao and Tsay), 86-110. Wiley, New York.

Xie, X. and Meng, X.-L. (2012). Multi-party inferences: What happens when there are three uncongenial models involved? *Technical Report*, Department of Statistics, Harvard University.

Yu, S-H., Lin, C.-C., and Cheng, H.-W. (2012). A note on mean squared prediction error under the unit root model with deterministic trend. *Journal of Time Series Analysis* **33**, 276-286.