Annals of the Institute of Statistical Mathematics (2022) 74:643–647
https://doi.org/10.1007/s10463-022-00829-3

INVITED ARTICLE: THIRD AKAIKE MEMORIAL LECTURE

Discussion of "Akaike Memorial Lecture 2020: Some of the challenges of statistical applications"

Masataka Taguri[1,2]

Received: 28 December 2021 / Revised: 28 December 2021 / Accepted: 17 January 2022 /
Published online: 28 May 2022

Keywords Statistical science · Shrinkage · Selection bias · Likelihood · Publication bias

I would like to first cerebrate Professor Copas for his thought-provoking contributions to statistical sciences. We all know of his deep and pioneering work on shrinkage estimators and publication bias and so on. Professor Copas' paper is on four topics (shrinkage of predictions, non-random samples, model sensitivity, and publication bias). Although the topics are broad, they all have one thing in common: they all extracted statistical problems from specific applications and developed methodologies and/or theories to solve problems that cannot be solved directly by the application of standard methods. I believe that this kind of research is an ideal way to advance both statistical theory and applications in the field of empirical science. Here, I will briefly comment on each of the four topics.

1 Shrinkage of predictions

Copas (1983) revealed that $E[y|\hat{y}]$ gives a shrinkage predictor:

$$E[y|\hat{y}] = K\hat{y} = Kx^T\hat{\beta}, \;\; K = 1 - \frac{1}{F},$$

where F is Fisher's F statistic for the regression model. More importantly, Copas (1983) and Houwelingen and Le Cessie (1990) also showed that this shrinkage predictor $Kx^T\hat{\beta}$ minimizes the mean squared prediction error,

✉ Masataka Taguri
taguri@tokyo-med.ac.jp

[1] Department of Data Science, Yokohama City University (Now at Tokyo Medical University), 6-6-2 Nishishinjyuku, Shinjyuku-ku, Tokyo 160-0023, Japan

[2] Research Center for Medical and Health Data Science, The Institute of Statistical Mathematics, Tokyo, Japan

$$\mathrm{MSE}(c) = E[(y - \hat{y}(c))^2],$$

where $\hat{Y}(c) = cx^T\hat{\beta}$ and c is a constant. The shrinkage coefficient K can be used to quantify overfitting and one can use the coefficient to calibrate the model for future data. This kind of idea (shrinkage and penalization) had a large impact on the practical modeling strategies (Harrell, 2015). For example, Greenland (2000) discussed that traditional variable selection procedures such as stepwise selection on confounders leave important confounders uncontrolled and shrinkage methods are superior to variable selection for both confounding control and prediction purposes.

Although the shrinkage predictor has very good properties, it is biased toward the null and a trade-off exists (see Fig. 1). The shrinkage predictor has smaller MSE but biased. On the other hand, a maximum likelihood estimator (MLE) has larger MSE but asymptotically unbiased. Thus, we should cautiously select an appropriate method depending on the objectives (e.g., prediction, estimation, and testing). Here is an example on the sparse modeling. On the regression problem with high-dimensional covariates, we often use the Lasso (Tibshirani, 1996). The Lasso estimator of the regression coefficients β is defined as the minimizer of the following objective function:

$$||Y - X\beta||_2^2 + \lambda||\beta||_1$$

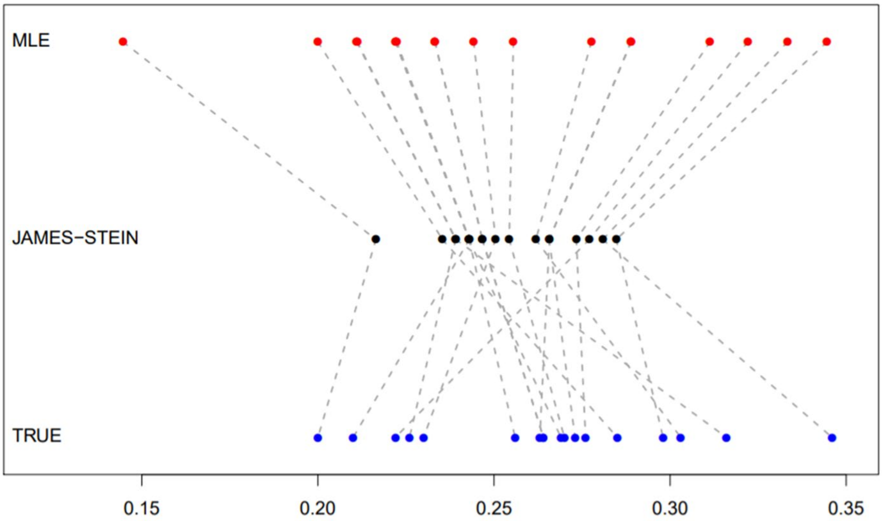

Fig. 1 A comparison between maximum likelihood estimator (MLE), the James–Stein estimator (a shrinkage estimator), and the true batting averages of eighteen baseball players in 1970 (Efron amd Hastie, 2016). *Computer Age Statistical Inference*. Cambridge University Press.). MLE is batting average in first 90 at bats and the true is average in remainder of 1970 season. It is clearly seen that the James–Stein estimator shrinks toward the grand mean (0.254) on each MLE value. It is also evident that the James–Stein estimator over-shrinks the data compared to the truth

where Y denotes a vector of the outcome variables and X is a design matrix. Usually the regularization parameter λ is selected to minimize the prediction error by cross-validation. However, if our objective is to select important covariates or risk factors, then λ could be selected to keep the false discovery rate at a level (Huang, 2017). If our objective is to select confounding factors and estimate a treatment effect, then we should also consider the association of potential confounders and a treatment to choose an appropriate λ (Koch et al., 2018).

2 Non-random samples

The non-random sample selection (selection bias) is an important problem to analyze an observational study data. Professor Copas considered the two extreme models:

Model A: sample selection is random and $E[y|x]$ decreases over x.
Model B: sample selection is non-random and $E[y|x]$ does not depend on x.

As he noted in the paper, the treatment allocations were determined by the doctors in the real clinical practice so that Model A is unrealistic. Using a simultaneous equation model, he showed that under *Model B*, the non-null association between x and y can exist in the selected sample. I think the truth is usually in the middle like the following Model C:

Model C: sample selection is non-random and $E[y|x]$ depends on x to some extent.

However, in the case of Model C, identification of the regression coefficients β must be difficult without another very strong assumption, because we do not have any information on the non-selected sample. Thus in this case, we may study sensitivity of inference to a given selection effect and repeatedly apply this method under the multiple possible scenarios (Copas and Li, 1997). I believe this kind of a sensitivity analysis is very important under the presence of non-identifiable parameters in a realistic model like Model C. We have proposed similar sensitivity analysis methods under the context of the treatment non-compliance and mediation analysis (Taguri and Chiba, 2012, 2015). However, Greenland and Lash (2008) pointed out that these approaches may convey unduly pessimistic or conservative picture of the uncertainty surrounding results because sensitivity analyses treat all scenarios (i.e., range of sensitivity parameters) equally, regardless of plausibility. Thus, if we consider relatively broad range of sensitivity parameters, Bayesian probabilistic sensitivity analysis using explicit prior distributions for the sensitivity parameters will be useful.

3 Robust likelihoods

There is a discrepancy between the accuracy of the fitting in the entire parametric model and the accuracy of the estimation of the *parameter of interest* or the *target parameter*. On the likelihood inference, Copas and Eguchi (2010) considered the set of alternative models G as comparative alternatives to the working log-normal model. They proposed to use the envelope of family of likelihoods $L_{ENV}(\varphi)$ to conduct a statistical inference considering the model uncertainty.

Another possible approach to focus on the target parameter is to use a semiparametric (nonparametric) model and estimate the target parameter directly. For example, if our objective is to estimate the mean of X, we can use the sample mean as an estimator without assuming a parametric model for X. In the field of causal inference, van der Laan and Rose (2011) proposed a framework or "road map" of inference. As a first step, we specify the target parameter as a function of the observed data (for example, confounders-adjusted mean of X). Then, we construct a semiparametric efficient estimator (the targeted MLE) for the parameter. The nuisance functions in the estimator will be estimated by the super learner, an ensemble learning algorithm to avoid parametric model assumptions. I wonder that this kind of semiparametric approach is sometimes preferable in terms of robustness to model misspecifications than the approach that looks at robustness to the local misspecification.

4 Publication bias

Publication bias has long been recognized as a major difficulty in systematic reviews in medical research and Professor Copas and his colleagues have made tremendous contributions to this field. In the paper, he assumed that the selection probability is an increasing function of t^2:

$$P(S|t) = 1 - \exp\left\{-(\alpha + \beta t^2)\right\} \text{ with } \alpha \geq 0,\ \beta \geq 0.$$

Given that the effect size is positively associated with the probability of publication, this assumption seems reasonable. I wonder there should be an identification problem for (α, β) because we observe only the data of publication studies. Another question is that although in the funnel plot the study estimate was negatively associated with 1/sd, there might be a possibility that there is a little selection bias but the true effect is indeed larger in small studies due to the careful patient selection, etc. We may assume this effect in the model for y, but this may lead to identification of the model parameters more difficult.

References

Copas, J. B. (1983). Regression, prediction and shrinkage. *Journal of the Royal Statistical Society, Series B, 45*, 311–335.

Copas, J. B., Eguchi, S. (2010). Likelihood for statistically equivalent models. *Journal of the Royal Statistical Society, Series B, 72*, 193–217.

Copas, J. B., Li, H. G. (1997). Inference for non-random samples (with discussion). *Journal of the Royal Statistical Society, Series B, 59*, 55–95.

Efron, B., Hastie, T. (2016). *Computer age statistical inference*. Cambridge, UK: Cambridge University Press.

Greenland, S. (2000). When should epidemiologic regressions use random coefficients? *Biometrics, 56*, 915–921.

Greenland, S., Lash, T. L. (2008). Bias Analysis. In K. J. Rothman, S. Greenland, T. L. Lash (Eds.), *Modern epidemiology*, 3rd ed. (pp. 345–380). Philadelphia: Lippincott-Williams-Wilkins.

Harrell, F. E., Jr. (2015). *Regression modeling strategies*: *With applications to linear models, logistic and ordinal regression, and survival analysis*, 2nd ed. New York: Springer.

Houwelingen, J. C., Le Cessie, S. (1990). Predictive value of statistical models. *Statistics in Medicine, 9*, 1303–1325.

Huang, H. (2017). Controlling the false discoveries in LASSO. *Biometrics, 73*, 1102–1110.

Koch, B., Vock, D. M., Wolfson, J. (2018). Covariate selection with group lasso and doubly robust estimation of causal effects. *Biometrics, 74*, 8–17.

Taguri, M., Chiba, Y. (2012). Instruments and bounds for causal effects under the monotonic selection assumption. *The International Journal of Biostatistics, 8*(1), 24.

Taguri, M., Chiba, Y. (2015). A principal stratification approach for evaluating natural direct and indirect effects in the presence of treatment-induced intermediate confounding. *Statistics in Medicine, 34*, 131–144.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B, 58*, 267–288.

van der Laan, M. J., Rose, S. (2011). *Targeted learning*: *Causal inference for observational and experimental data*. New York: Springer.