GOF Tests via Machine Learning

Some reflections and a number of questions

PHYSTAT Informal Review

March 12, 2025

Alessandra R. Brazzale

alessandra.brazzale@unipd.it

Department of Statistical Sciences | University of Padova

Ministero dell'Università e della Ricerca | PRIN | Università degli Studi di Padova

Ongoing work with. . .

Prof. Anthony C. Davison (EPFL)

Max Jeffrey LoConte (MSc Statistics @EPFL)

Premise

- I have far more questions than answers…
- Feedback most welcome!

2023 PhyStat Workshop on Systematics @Banff

April 23 – 28, 2023

Image Credit: https://www.birs.ca/events/2023/5-day-workshops/23w5096

Struck by…

- goodness-of-fit
- likelihood ratio test
- *exponential tilting*
- two-sample testing
- optimal test

Parametrisation of the alternative hypothesis:

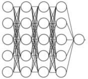
$$n(x\,|\,\mathrm{H}_{\hat{\mathbf{w}},\hat{\nu}}) = n(x\,|\,\mathrm{R}_0)\,\frac{n(x\,|\,\mathrm{R}_{\hat{\nu}})}{n(x\,|\,\mathrm{R}_0)}\,e^{f(x;\,\hat{\mathbf{w}})}$$

central value SM (ν=0)

Model of systematic uncertainties effects

$$\hat{r}(x;\nu) = \frac{n(x\,|\,\mathrm{R}_\nu)}{n(x\,|\,\mathrm{R}_0)} = \exp\left[\hat{\delta}_1(x)\,\nu + \hat{\delta}_2(x)\,\nu^2 + \ldots\right]$$

NN model



Won’t talk about. . .

goodness-of-fit testing vs two-sample testing

Try and talk about. . .

- (1933)

IX. *On the Problem of the most Efficient Tests of Statistical Hypotheses.*

By J. NEYMAN, *Nencki Institute, Soc. Sci. Lit. Varsoviensis, and Lecturer at the Central College of Agriculture, Warsaw, and* E. S. PEARSON, *Department of Applied Statistics, University College, London.*

- (1938)

THE LARGE-SAMPLE DISTRIBUTION OF THE LIKELIHOOD RATIO FOR TESTING COMPOSITE HYPOTHESES[1]

BY S. S. WILKS

When is the LR optimal?

given sample x and significance level α

- scalar parameter of interest θ

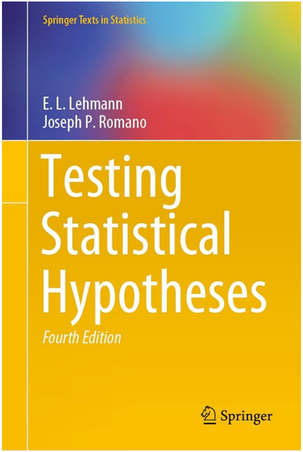
$H_0 : p_0(x)$ vs $H_1 : p_1(x)$	**most powerful** test based on LR $\frac{p_1(x)}{p_0(x)}$
$H_0 : \theta \leq \theta_0$ vs $H_1 : \theta > \theta_0$	**uniformly most powerful** test[1] if LR is *monotone* in $T(x)$
$H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$	**uniformly most powerful unbiased** test if $T(x)$ has *symmetric* distribution

- $\theta = (\psi, \lambda)$, with ψ scalar parameter of interest

$H_0 : \psi = \psi_0$ vs $H_1 : \psi \neq \psi_0$	**uniformly most powerful unbiased similar** test, e.g. *natural exponential family*

[1] Karlin–Rubin theorem

A most enjoyable book



Likelihood-based theory

given likelihood $L(\theta; x)$ and MLE $\hat{\theta}$

- UMP tests need be worked out on a case-by-case basis
- Wilks (1938) characterizes the asymptotic (large n) distribution of log LR

$$2\left\{\log L(\hat{\theta}; X) - \log L(\theta_0; X)\right\} \sim \chi^2_\nu$$

- large- and small-sample neo-Fisherian theory based on likelihood-type functions (score statistic, Wald test, profile likelihood, . . .)

Exponential family

given $\theta \in \Theta \in \mathbb{R}^d$, $T(y) \in \mathscr{T} \in \mathbb{R}^d$, $h(\cdot) \geq 0$ and $\kappa(\theta)$ finite

- natural exponential family

$$p(y;\theta) = h(y)\exp\{T(y)^\top \theta - \kappa(\theta)\}$$

- obtained by embedding single fixed density $p_0(y)$

$$p(y;\theta) \propto p_0(y)e^{t(y)^\top \theta}$$

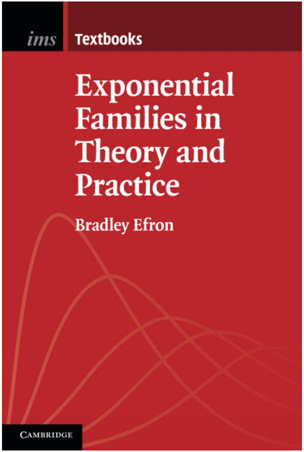
in a larger class by *exponential tilting*

Many nice properties

$$p(y;\theta) = h(y)\exp\{t^\top(y)\theta - \kappa(\theta)\}$$

- same support
- $t(y)$ is minimal sufficient for θ
- $\kappa(\theta)$ is cumulant generating function
- If $\theta = (\psi, \lambda)$ and $t(y) = (t_1(y), t_2(y)) \Longrightarrow$ conditional distribution of t_1 given $T_2 = t_2$ only depends on ψ.

A second most enjoyable book



And here we are...

$$n(x|\mathrm{H}_{\mathbf{w}}) = e^{f_{\mathbf{w}}(x)} n(x|\mathrm{R})$$

- exponential tilting on intensity (instead of density)
- $f_{\mathrm{w}}(x)$ reconstructed by M(achine) L(earning)[2] to yield

$$-2\log \frac{\mathcal{L}(\mathrm{R}|\mathcal{D})}{\mathcal{L}(\mathrm{H}_{\widehat{\mathbf{w}}}|\mathcal{D})} = -2\left[\mathrm{N}(\mathrm{H}_{\widehat{\mathbf{w}}}) - \mathrm{N}(\mathrm{R}) - \sum_{x\in\mathcal{D}} f_{\widehat{\mathbf{w}}}(x)\right]$$

- which, given $\mathrm{N}(\mathrm{H}_{\hat{\mathrm{w}}})$ and $\mathrm{N}(\mathrm{R})$ is monotone in $\sum_{x\in D} f_{\hat{\mathrm{w}}}(x)$

[2] not maximum likelihood!

What are we working on

New Physics Learning Machine

- When does it give rise to χ^2 limiting distributions?
- What about the df's?
- (Can we connect it to *generalized* likelihood ratio testing?)

On Profile Likelihood

S. A. MURPHY and A. W. VAN DER VAART

We show that semiparametric profile likelihoods, where the nuisance parameter has been profiled out, behave like ordinary likelihoods in that they have a quadratic expansion. In this expansion the score function and the Fisher information are replaced by the efficient score function and efficient Fisher information. The expansion may be used, among others, to prove the asymptotic normality of the maximum likelihood estimator, to derive the asymptotic chi-squared distribution of the log-likelihood ratio statistic, and to prove the consistency of the observed information as an estimator of the inverse of the asymptotic variance.

KEY WORDS: Least favorable submodel; Likelihood ratio statistic; Maximum likelihood; Nuisance parameter; Semiparametric model; Standard error.

What are we working on

C2ST

- What if we train $p_0(x)$ with R and $p_1(x)$ with H_w?
- (Need strategies for post-selection inference?)

Splitting strategies for post-selection inference

BY D. GARCÍA RASINES AND G. A. YOUNG

Department of Mathematics, Imperial College London, London SW7 2AZ, U.K.

daniel.garcia-rasines16@imperial.ac.uk alastair.young@imperial.ac.uk

SUMMARY

We consider the problem of providing valid inference for a selected parameter in a sparse regression setting. It is well known that classical regression tools can be unreliable in this context because of the bias generated in the selection step. Many approaches have been proposed in recent years to ensure inferential validity. In this article we consider a simple alternative to data splitting based on randomizing the response vector, which allows for higher selection and inferential power than the former, and is applicable with an arbitrary selection rule. We perform a theoretical and empirical comparison of the two methods and derive a central limit theorem for the randomization approach. Our investigations show that the gain in power can be substantial.

Thank you!

Alice
Ye Mouse
Ye Lobster
Ye White King
Ye Mad Hatter
Ye Youth
Old Father William
Ye Queen's Gardeners
Ye Slithy Tove
Ye Borogove
Tweedledum

References

- Efron, B. (2022). *Exponential Families in Theory and Practice*. Cambridge University Press, Cambridge.
- García Rasines, E. and Young, G.A. (2023). Splitting strategies for post-selection inference. *Biometrika*, **110** 597–614.
- Grosso, G., Letizia, M., Pierini, M. and Wulzer, A. (2024). Goodness of fit by Neyman-Pearson testing. *SciPost Physics*, **16**, 123.
- Lehmann, E. and Romano, J. (2022). *Testing statistical hypotheses*. Springer-Verlag, New York, Fourth edition.
- Murphy, S.A. and van der Vaart, A.W. (2000). On profile likelihood. *Journal of the American Statistical Association*, **95**, 449–465.
- Neyman, J. and Pearson, E. (1932). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London*. **Series A**
- Wilks, S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, **9** 60–62.

Acknowledgments

The publication was produced with funding from the Italian Ministry of University and Research under the Call for Proposals related to the scrolling of the final rankings of the PRIN 2022 call. — Project title "Latent variable models and dimensionality reduction methods for complex data" – Project No. 20224CRB9E, CUP C53C24000730006 – PI Prof. Paolo Giordani - Grant Assignment Decree No. 1401 adopted on 18.9.2024 by MUR.

Ministero dell'Università e della Ricerca

PRIN

Università degli Studi di Padova