Journal of Data Science 24 (1), 4–25
January 2026
DOI: 10.6339/26-JDS1217
Philosophies of Data Science

The Typicality Principle and Its Implications for Statistics and Data Science

Yiran Jiang[1,*], Zeyu Zhang[2], Ryan Martin[3], and Chuanhai Liu[2]

[1]*Dr. Bing Zhang Department of Statistics, University of Kentucky, USA*
[2]*Department of Statistics, Purdue University, USA*
[3]*Department of Statistics, NC State University, USA*

Abstract

A central focus of data science is the transformation of empirical evidence into knowledge. By "knowledge," we mean claims that are (i) supported by data through an explicit inferential procedure and (ii) accompanied by calibrated measures of uncertainty. As such, the scientific insights and attitudes of deep thinkers like Ronald A. Fisher, Karl R. Popper, and John W. Tukey are expected to inspire exciting new advances in machine learning and artificial intelligence in years to come. Along these lines, the present paper advances a novel *typicality principle* which states, roughly, that if the observed data is sufficiently "atypical" in a certain sense relative to a posited theory, then that theory is unwarranted. This emphasis on typicality brings familiar but often overlooked background notions like model-checking to the inferential foreground. One instantiation of the typicality principle is in the context of parameter estimation, where we propose a new typicality-based regularization strategy that leans heavily on goodness-of-fit testing. The effectiveness of this new regularization strategy is illustrated in three non-trivial examples where ordinary maximum likelihood estimation fails miserably. We also demonstrate how the typicality principle fits within a bigger picture of reliable and efficient uncertainty quantification.

Keywords *falsification; goodness-of-fit; inferential model; likelihood; model-checking; regularization; uncertainty quantification*

1 Introduction

Data science has captured the attention of researchers across all areas of science, business, and government. While acknowledging that deductive and abductive steps play roles in scientific workflows, like statistics, data science as a discipline is primarily concerned with the transformation of empirical evidence into knowledge, which falls under the umbrella of inductive logic:

> *In inductive reasoning we are performing part of the process by which new knowledge is created. The conclusions normally grow more and more accurate as more data are included.* (Fisher, 1935b, p. 54)

Despite statistics' 100-year head start, data science has arguably already surpassed it as the leader in applied inductive logic (Breiman, 2001; Donoho, 2017). This shift is more than just a change in terminology; it reflects an evolution in our approach to inference, prediction, and decision-making, drawing on tools and insights from machine learning and artificial intelligence.

*Corresponding author. Email: yiran.jiang@uky.edu.

Received January 4, 2026; Accepted January 15, 2026

As data science continues to advance, it is sure to inspire—and be inspired by—foundational work along lines championed by Ronald A. Fisher, Karl R. Popper, and John W. Tukey.

The complexity inherent in modern data sets means many sources of uncertainty and ambiguity, thus making the analysis and ensuing inductive argument highly non-trivial. As such, insights from epistemology are germane, and the dominant school of thought is falsificationism, as laid out in Karl Popper's *Logic of Scientific Discovery* (Popper, 1959a). His key insight was that, in a sequence of experiments that severely test relevant theories, those theories that withstand this scrutiny will have "proved their mettle" (Popper, 1959b, p. 10), a necessary, but not sufficient, condition for any theory to be called *not-false*. It is only in a limiting sense, as the number/severity of the tests increases, that a theory can earn the not-false distinction. A challenge in modern empirical sciences, not found in the "all swans are white" style examples often considered in philosophy texts, is that the empirical data cannot logically contradict any legitimate theory, so there will inevitably be uncertainty in drawing inferences. This necessitates the reliable quantification of said uncertainty and this is the focus of our present contribution.

Behind the falsificationist perspective is the understanding that experiments tend to produce data that are *typical*, i.e., look like what is expected in the real world. Consequently, if the observed data are atypical—or look sufficiently different from what is expected—relative to a posited theory, then it is fair to consider that theory falsified. This is our proposed *typicality principle* in its basic form. But what does it mean for data to "look like" what is expected? (Strictly speaking, there's a sense in which all data are "atypical." For instance, in models that involve absolutely continuous distributions, all possible data realizations have zero probability under such models. Fortunately, strict atypicality is not necessary for reliable inference—meaningful notions of atypicality can be developed with a little creativity.) It is common to interpret the model's likelihood function as a measure of its quality of fit to the observed data and, in turn, it is common to judge whether data "look like" what is expected by the magnitude of the likelihood. The *law of likelihood* (e.g., Edwards, 1992; Hacking, 1976) formalizes this. In some cases, however, the likelihood function value could be large as a result of some degeneracy, not because the data "look like" what is expected under the corresponding theory. This highlights a shortcoming in the likelihood-centric approach to inductive inference and a need for new perspectives. The familiar regularization strategies depend only on the posited theory—e.g., penalize theories incompatible with assumed "sparsity"—and not on the data itself, hence cannot repair the aforementioned flaw on their own. Alternatively, our notion of typicality focuses on *fit* in a nonparametric goodness-of-fit sense as opposed to a parametric model-based, large-likelihood sense. The typicality principle advanced here is inspired by Tukey's insights on model building/checking (Tukey, 1977, 1962). While philosophical principles tend to be "top down," i.e., instructions passed down from a higher authority, Tukey's brand of philosophy is "bottom up" (e.g., Dempster, 2002; Tukey, 1986) and, hence, Popperian in spirit. Indeed, in model-building, for example, no candidate model is God-given, so a model's merits must be earned by providing a satisfactory explanation of the observations. Our proposed typicality principle builds on these key ideas.

Our main methodological innovation here is a new brand of *typicality-focused* regularization strategies. More specifically, we recommend adoption of the familiar penalized likelihood framework but with a twist: rather than penalizing theories that are incompatible with prior knowledge, we penalize those theories for which the data could be judged atypical, thus ensuring our derived procedures reward theories that fit the data well, aiding interpretation and enhancing efficiency. In this sense, typicality-focused regularization is not a structural constraint or a

prior specification, but a principled mechanism for integrating model checking and likelihood-based fitting within a single, calibrated inferential framework. We evaluate the performance of our proposed typicality-focused regularization by applying it to several challenging problems that have historically served as points of contention in the foundations of statistics. Our results demonstrate its efficiency in point estimation and uncertainty quantification more broadly, highlighting its potential to address some of the deepest unresolved issues in statistical science. Beyond its practical applications, we delve into the theoretical underpinnings of the typicality principle, uncovering connections—or lack thereof—to other familiar statistical principles. These connections underscore the broader importance of typicality, not only as a methodological tool but also as a conceptual bridge linking various aspects of statistical reasoning. By positioning the typicality principle within this rich theoretical and applied context, this paper lays the groundwork for future exploration of its implications in data science and beyond. This perspective aligns with the goals of trustworthy machine learning, where one seeks not only accurate predictions but also reliable uncertainty statements and diagnostics that detect when a fitted model is not behaving as assumed. Recent work has emphasized that reliable uncertainty quantification is a central component of trustworthy artificial intelligence, particularly for detecting model limitations and supporting robust decision-making (Deuschel et al., 2024; Liu et al., 2024). In particular, typicality-based assessment can be viewed as a statistical mechanism for (i) stress-testing model adequacy and (ii) moderating likelihood-driven fitting when diagnostic evidence suggests lack-of-fit, thereby improving the trustworthiness of downstream decisions.

The remainder of the paper is organized as follows. Section 2 sets the context of our discussion and introduces some key concepts and notation. Section 3 introduces a first basic version of the typicality principle, and our discussion there focuses on statistical intuitions and philosophical considerations. In the context of parameter estimation, a particular instantiation of the typicality principle results in our new typicality-based regularization strategy which is also detailed there. The performance of our proposed typicality-based regularization strategy is investigated in Section 4 in the context of three nontrivial, paradox-laden examples: a mixture model due to Le Cam, the Neyman–Scott problem, and Stein's mean vector length. The challenges faced in these three examples are also common in modern data science applications where there is a risk of over-fitting, so our contributions here extend beyond the simple parametric models considered here. Section 5 digs deeper, advancing a formal typicality principle and showing how this fits into a general framework that offers provably reliable uncertainty quantification, beyond point estimation, hypothesis testing, etc. Connections to other statistical principles are offered and a numerical illustration demonstrates the validity and efficiency of the proposed framework in a challenging marginal inference problem. We conclude in Section 6 with a few remarks.

2 Problem Setup

To set the scene, we introduce our notation, model assumptions, and objectives. Let X, taking values in the sample space $\mathbb{X}$, denote the observable data. We posit a statistical model $\{\mathsf{P}_\theta : \theta \in \mathbb{T}\}$ for X, which consists of a collection of probability distributions supported on $\mathbb{X}$, indexed by a parameter θ in the parameter space $\mathbb{T}$. This model could be the kind of low-dimensional models presented in textbooks, e.g., $\mathsf{P}_\theta = \mathsf{Gamma}(\alpha, \beta)$ with $\theta = (\alpha, \beta)$, or the model could be high-dimensional as in the case of a neural network, so θ can be scalar-, vector-, or even function-

valued. The distributions P_θ included in the model have associated probability density or mass functions, denoted by p_θ, and for each $x \in \mathbb{X}$, the function $L_x(\theta) = p_\theta(x)$ is the likelihood.

We will assume that there exists a "true value" of the uncertain model parameter θ, denoted by Θ. Cases in which the true distribution does not belong to the posited model can also be considered, but we will not consider this here; see Jiang and Liu (2025). Of course, data $X = x$ is observed, and our objective is to make inference on the uncertain Θ, relative to the posited statistical model. The most basic inferential objective is point estimation, which amounts to selecting a most likely or plausible parameter value based on observation x; this will be discussed in more detail in Section 3. Beyond point estimation, a goal is to quantify uncertainty about Θ. More specifically, uncertainty associated with relevant hypotheses "$\Theta \in H$" is quantified with a data-dependent, numerical degree of plausibility and, if a formal test was desired, hypotheses determined to be sufficiently implausible would be rejected on those grounds. Henceforth, for simplicity of presentation, both H and "$\Theta \in H$" will refer to hypotheses about the uncertain Θ. Given the plausibility assignments to hypotheses $H \subseteq \mathbb{T}$, these can be inverted to find, e.g., relatively small hypotheses H whose complements are relatively implausible. This brings confidence set construction under our general inference umbrella.

We will further assume that prior information about Θ is vacuous, so the aforementioned degrees of belief/confidence cannot be obtained via application of Bayes's theorem with a genuine or informative prior distribution for Θ. "Vacuous prior information" technically means that we lack the information required to eliminate any candidate prior distributions for Θ (Martin, 2022a), hence, all priors and all corresponding Bayesian solutions are plausible. Since it is impossible to learn about Θ, given x, from a Bayesian analysis under such an extreme credal state (e.g., Kyburg, 1987), the confirmationist solution is strictly out of reach. Consequently, achieving our top-level goal of quantifying uncertainty about Θ requires considerations beyond the familiar Bayesian and fiducial-like frameworks; details are presented in Section 5 below.

3 Typicality Principle

We adopt Popper's falsificationist view, that is, a hypothesis H concerning the uncertain Θ cannot be confirmed based on data x, it can only be refuted or not depending on whether or not truthfulness of H and data x are sufficiently contradictory. Importantly, falsificationism is really our only option: employing, e.g., the opposing Carnapian–Jeffreysian–Jaynesian confirmationist view (e.g., Carnap, 1962; Jeffreys, 1998; Jaynes, 2003) in a scientifically and mathematically rigorous way requires a genuine prior probability distribution and application of Bayes's theorem, which is out of reach as discussed in Section 2. As discussed in Section 1, our falsificationist perspective suggests the specification of a strategy for assessing if—and, if so, in what sense and to what extent—the data x is "typical" relative to a stated hypothesis H about Θ. Given such an assessment, inference is at least conceptually straightforward.

We start here with an informal and intuitive statement of the key principle, one that focuses on simple, singleton hypotheses. More formality is given in Section 5.

Typicality Principle (intuitive version)**.** *If data x is sufficiently atypical relative to the posited model with parameter θ, then the hypothesis $H = \{\theta\}$ is unwarranted.*

The reader is surely tempted to make a comparison between the typicality principle as stated above and more familiar notions involving the likelihood. For instance, the law of likelihood (e.g., Edwards, 1992, p. 30) states, roughly, that data x supports the hypothesis $H = \{\theta\}$

less than hypothesis $H' = \{\theta'\}$ if $L_x(\theta) < L_x(\theta')$ and, consequently, if the gap was sufficiently wide, then hypothesis H would be unwarranted. This in turn leads naturally to the principle of maximum likelihood (e.g., Stigler, 2007; Aldrich, 1997; Fisher, 1922, 1925), which suggests using the maximally supported singleton hypothesis, i.e., the value $\hat{\theta}(x) = \arg\max_\theta L_x(\theta)$ with maximal likelihood, as a point estimator for Θ given x. It's well-known, however, that the likelihood function can be large at (x, θ) values even though x is highly atypical of a sample from P_θ. This is often the result of some inadequacy of the model, such as degeneracy or overfitting, and commonly manifests as inconsistency: as the information in the sample increases, the maximum likelihood estimator fails to converge in probability to Θ. Despite the ubiquity of maximum likelihood estimation in textbooks and in applications, and their asymptotic efficiency in regular cases, their general inadequacy has been widely documented. Indeed, as quoted in van der Vaart (2002, Sec. 11), Lucien Le Cam emphasized as early as 1960 that

> The author is firmly convinced that a recourse to maximum likelihood is justifiable only when one is dealing with families of distributions that are extremely regular. The cases in which [maximum likelihood] estimates are readily obtainable and have been proved to have good properties are extremely restricted.

and, later, in Le Cam (1986),

> The terms "likelihood" and "maximum likelihood" seem to have been introduced by RA Fisher who seems also to be responsible for a great deal of the propaganda on the merits of the maximum likelihood method... In view of Fisher's vast influence, it is perhaps not surprising that the presumed superiority of the method is still for many an article of faith promoted with religious fervor. This state of affairs remains, in spite of a long accumulation of evidence to the effect that maximum likelihood estimates are often useless or grossly misleading.

Examples highlighting the inadequacies of maximum likelihood estimators are presented in Section 4 below. Le Cam's point here is just that maximum likelihood works fine in some cases but not in others. That an estimator works well in some cases and not in others on its own creates no foundational concerns. If, however, the *principle* of maximum likelihood itself is untrustworthy, then that is a serious foundational concern: data science cannot be a science without trustworthy principles. The typicality principle aims to fill this trustworthiness gap.

One interpretation of the maximum likelihood estimator $\hat{\theta}(x)$ is that it represents the member of the posited statistical model that is closest (in the sense of Kullback–Leiber divergence (Lehmann, 1983; Pardo, 2018)) to the empirical distribution of the data. Since the latter empirical distribution necessarily represents the data, it might come as a surprise to hear that x might not look like a typical sample from the "best approximation" $\mathsf{P}_{\hat{\theta}(x)}$ in the posited model. The point is that the maximum likelihood estimator depends only on some summary features of the full data, rarely on the full data itself; on the other hand, model building/checking processes use all aspects of the data. As an intentionally oversimplified example, suppose that x consists of what is assumed to be an iid sample of size n from a normal distribution with unit variance. Then the maximum likelihood estimator is the average of those values in x and, hence, it only depends on this summary. So, of course, if two data sets x and y are of the same size and have the same sample means, then the likelihood functions and hence the maximum likelihood estimators are identical. But if the sample mean of x is 0 and all of the entries in y are 0, then x might look like a typical sample from $\mathsf{N}(0, 1)$ but y definitely does not. The catch is that the law of

likelihood, etc. all include the caveat *relative to the given model* and, since it would be absurd to model a data set y whose entries are all identically 0 by a normal distribution, this potential issue is rarely discussed. In fact, the above point boils down to the sufficiency principle and, in the right context, is considered by most (including us) to be a desirable feature of likelihood-based inference. In real applications, however, where data and model are more complex than in the toy example above, it may not be clear that $\hat{\theta}(x)$ only depends on an unacceptably incomplete summary of x and, hence, that x might not resemble a typical sample from $\mathsf{P}_{\hat{\theta}(x)}$; see Section 4. Moreover, the trope "all models are wrong but some are useful" suggests that we should anticipate some degree of model misspecification and manually build in some direct model-fit considerations rather than fully trust the data reductions inherent in likelihood-based methods.

Our proposed implementation of the typicality principle is via regularization. This will be anchored around the likelihood function, since this is efficient in regular cases, but our proposed regularization differs from that common in the data science literature in some important ways. Specifically, consider the objective function

$$\rho_\lambda^{\text{typ}}(x, \theta) = \ell_x(\theta) - \lambda\, r_x^{\text{typ}}(\theta), \quad \theta \in \mathbb{T}, \tag{3.1}$$

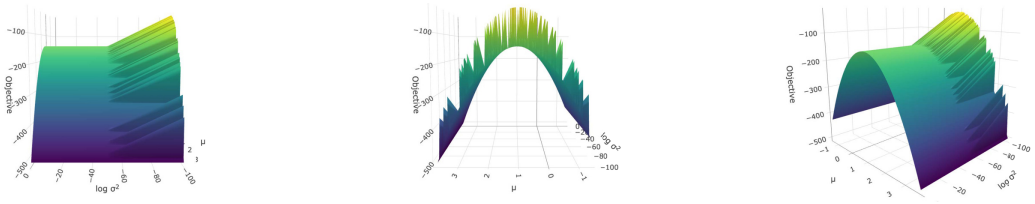
where $\ell_x = \log L_x$ is the usual log-likelihood, $\lambda \geqslant 0$ is a scalar tuning parameter, and r_x^{typ} is a data-dependent, typicality-encouraging penalty. Although λ plays a different role here compared to more familiar settings, namely balancing likelihood-based fitting with typicality-based calibration, standard tuning-parameter selection strategies may still be used. Additional discussion is provided in Section 6. The goal, of course, is to maximize this objective function at a fixed x to obtain a point estimator $\check{\theta}(x)$, which depends on the user's choice of λ and on r_x^{typ}.

Note that the penalty r_x^{typ} depends on data x. Almost exclusively, the penalty functions used in the literature do not depend on data—their goal is to push the maximum likelihood estimator towards where Θ is believed to be based on *a priori* knowledge about the problem. Such penalties do not address the question of whether x is a "typical" sample from P_θ, and nor can they; any measure that is designed to quantify typicality must depend on both x and θ.

Since the focus is on determining if x is atypical relative to P_θ, Tukey's insights on model-checking suggest considering some variation on a goodness-of-fit test in the construction of r_x^{typ}. Indeed, throughout this paper, we consider

$$r_x^{\text{typ}}(\theta) = -\log \mathsf{pval}_x(\theta), \quad \theta \in \mathbb{T}, \tag{3.2}$$

where $\mathsf{pval}_x(\theta)$ is the p-value associated with a test of how well P_θ fits the data x. A fairly general formulation is as follows. Suppose $x = (x_1, \ldots, x_n)$ consists of iid samples where P_θ determines a continuous distribution function F_θ for the components. If x is a "typical" sample from P_θ, then $F_\theta(x_1), \ldots, F_\theta(x_n)$ should resemble an iid $\mathsf{Unif}(0, 1)$ sample, and we can take $\mathsf{pval}_x(\theta)$ to be the p-value associated with, say, the Kolmogorov–Smirnov test of this uniformity. See Liu (2023) and Jiang and Liu (2025) for more on the Kolmogorov–Smirnov p-value for assessing what we refer to here as typicality. In certain special cases, however, other simpler goodness-of-fit assessments may be available. For example, in Gaussian models where "residual sums of squares" are expected to have a suitable chi-square distribution, the p-value associated with such a chi-square test can be used to construct the penalty function r_x^{typ}; see Section 4.2. The examples in Section 4 highlight how the penalty (3.2) leads to desirable regularization, correcting for systematic biases present in the maximum likelihood estimators in sufficiently non-regular models.



(a) Effect of σ^2 (b) Small σ^2 (c) Large σ^2

Figure 1: Log-likelihood surfaces from three viewing angles.

4 Examples: Non-regular Estimation

4.1 Le Cam's Mixture

Following Le Cam (1990), let α be a very small known number, such as 10^{-50}, and consider the mixture model $\mathsf{P}_\theta = (1-\alpha)\,\mathsf{N}(\mu, 1) + \alpha\,\mathsf{N}(\mu, \sigma^2)$, indexed by $\theta = (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$. The goal is to estimate the uncertain true value $\Theta = (M, \Sigma^2)$ of the model parameter.

Given n independent samples $x = (x_1, \ldots, x_n)$, the likelihood function is

$$L_x(\mu, \sigma^2) \propto \prod_{i=1}^{n} \Big[(1-\alpha) \exp\Big\{ -\frac{(x_i-\mu)^2}{2} \Big\} + \alpha\sigma^{-1} \exp\Big\{ -\frac{(x_i-\mu)^2}{2\sigma^2} \Big\} \Big].$$

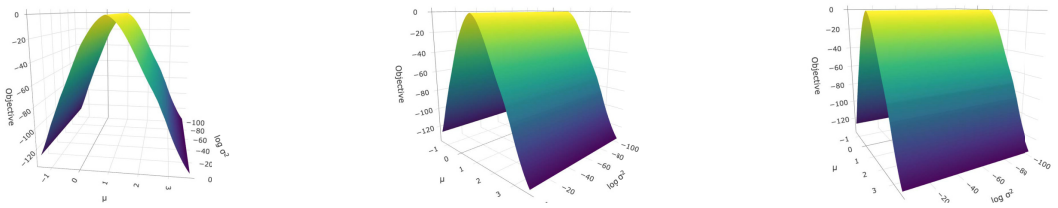
The difficulty with maximum likelihood estimation in this application is that the likelihood function is unbounded; this implies that the maximum likelihood estimator is on the boundary of the parameter space and, hence, does not exist in a practical sense. To see this, set μ equal to any of the observed data points, say, x_1, and consider the likelihood as a function of σ^2 alone:

$$\begin{aligned} L_x(x_1, \sigma^2) &\propto \prod_{i=1}^{n} \Big[(1-\alpha) \exp\Big\{ -\frac{(x_i-x_1)^2}{2} \Big\} + \alpha\sigma^{-1} \exp\Big\{ -\frac{(x_i-x_1)^2}{2\sigma^2} \Big\} \Big] \\ &= \{(1-\alpha) + \alpha\sigma^{-1}\} \prod_{i=2}^{n} \Big[(1-\alpha) \exp\Big\{ -\frac{(x_i-x_1)^2}{2} \Big\} + \alpha\sigma^{-1} \exp\Big\{ -\frac{(x_i-x_1)^2}{2\sigma^2} \Big\} \Big]. \end{aligned}$$

The second factor, i.e., the product over $i = 2, \ldots, n$, is bounded as a function of σ^2; the first factor, $(1-\alpha) + \alpha\sigma^{-1}$, is unbounded as $\sigma^2 \to 0$. This proves the non-existence of the maximum likelihood estimator and highlights a shortcoming in the principle of maximum likelihood.

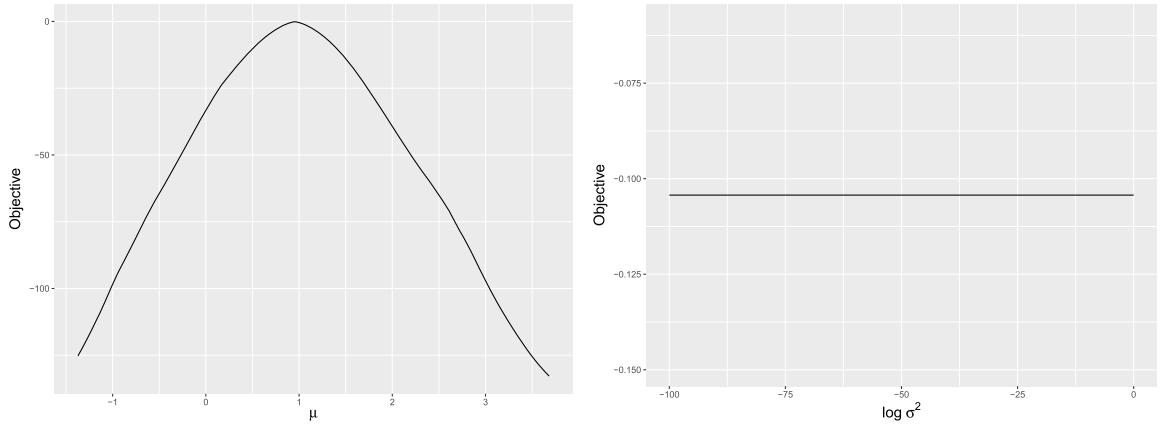
To illustrate this shortcoming, we simulate data of size $n = 100$ from the distribution P_Θ, where $\Theta = (M, \Sigma^2) = (1, 2)$. Then the log-likelihood is evaluated over a range of different $(\mu, \log\sigma^2)$ values and the results are plotted in Figure 1. As the theoretical analysis suggests, the log-likelihood increases monotonically as $\sigma^2 \to 0$. When σ is small enough, as shown in Figure 1(b), the choices of μ affect the log-likelihood and create the jumps when they are equal to the values of the samples. On the other hand, the surface is quadratic when σ^2 is large, as shown in Figure 1(c), while only the first quadratic plays a role with small α. In summary, using a "maximum likelihood estimator" $\hat\theta(x) = (x_1, 0)$ amounts to ignoring virtually all of the relevant information in the data—bad practice! When making predictions, the estimated model $\mathsf{P}_{\hat\theta(x)}$ will be inefficient, since it is effectively the same fitted model as that based on x_1 alone.

To overcome the pitfalls of maximum likelihood, we apply the typicality principle as suggested in Section 3 using the proposed typicality-based regularization (3.2). That is, we take



(a) Effect of μ (b) Effects of μ and σ^2 (c) Effect of σ^2

Figure 2: Log p-value surfaces for the Kolmogorov–Smirnov test.



(a) Effect of μ (b) Effect of σ^2

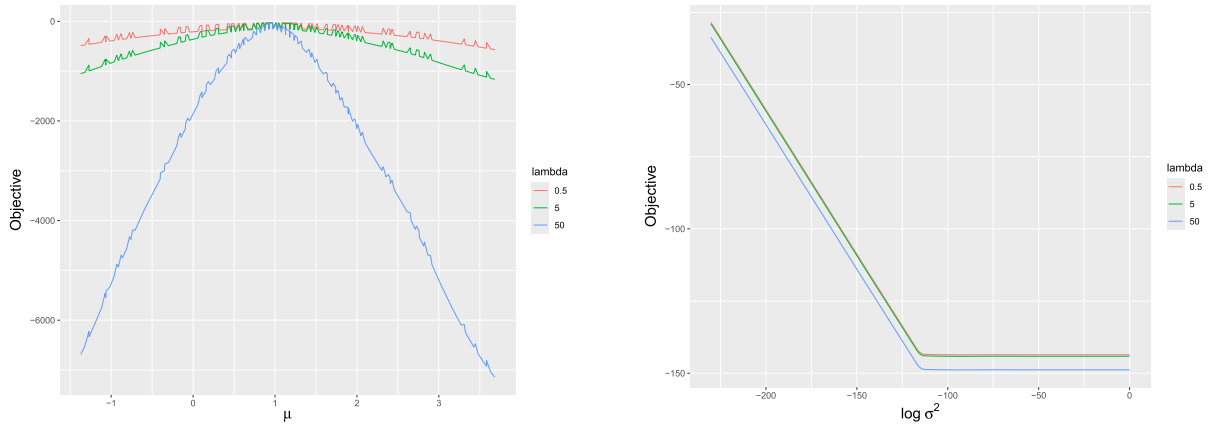
Figure 3: Profiled log p-values for the Kolmogorov–Smirnov test.

$r_x^{\text{typ}}(\theta)$ as the negative log p-value of the Kolmogorov–Smirnov test assessing the fit of P_θ to the observed x. Clearly from Figure 2, we find that μ plays a more important role than σ^2. The choice of σ^2, however, does not affect the p-values at all. The p-value profiles, i.e., fixing one parameter and maximizing over the other, are shown in Figure 3. Interestingly, the value of μ suggested by the goodness-of-fit test is around 1, which coincides with the true parameter value, $M = 1$.

Lastly, we illustrate the effect of the tuning parameter λ. Figure 4 presents two plots of the objective function based on different λ. We postpone investigation into data-driven choices of λ for a future paper, but make the observation here that certain values of the tuning parameter make it easier to optimize over μ. There are, however, still challenges associated with optimizing over σ^2, which is to be expected: since α is very small, at most a few samples out of n will come from the second component of the mixture, hence, the data is not informative about Σ^2.

4.2 Neyman–Scott Problem

Next we consider "a more disturbing example" (Le Cam, 1990), due to Neyman and Scott (1948). In this case, the model P_θ, indexed by $\theta = (\xi_1, \ldots, \xi_n, \sigma^2) \in \mathbb{R}^n \times (0, \infty)$, posits that the data consists of independent pairs, $X_i = (X_{i1}, X_{i2})$, where $X_{i1}, X_{i2} \overset{\text{ind}}{\sim} \mathsf{N}(\xi_i, \sigma^2)$, for $i = 1, \ldots, n$. All of $\Theta = (\Xi_1, \ldots, \Xi_n, \Sigma^2)$ are unknown, but the primary goal is estimation of Σ^2. It is easy to



(a) Effect of μ (b) Effect of σ^2

Figure 4: Profiles of the objective function $\rho_\lambda^{\text{typ}}(\mu, \sigma^2)$ for different λ.

show that the maximum likelihood estimator of Ξ_i is $\hat{\xi}_i = \frac{1}{2}(X_{i1} + X_{i2})$, and the corresponding maximum likelihood estimator of Σ^2 is

$$\hat{\sigma}^2 = \frac{1}{2n} \sum_{i=1}^{n} \{(X_{i1} - \hat{\xi}_i)^2 + (X_{i2} - \hat{\xi}_i)^2\}.$$

But $\hat{\sigma}^2$ is a biased estimator—its expected value is $\Sigma^2/2 \neq \Sigma^2$—which further implies that the maximum likelihood estimator is inconsistent. Hence, the Neyman–Scott problem sheds light on further issues with the principle of maximum likelihood.

We address this problem by applying the typicality principle, in particular, by employing the typicality-based regularization strategy outlined in Section 3. With a slight abuse of the notation in Section 3, we focus on the interest parameter Σ^2. Having eliminated the nuisance parameters $\Xi_1, \ldots, \Xi_n$, a simple goodness-of-fit strategy suggests itself. That is, we take $\mathsf{pval}_x(\sigma^2)$ to be the (two-tailed) p-value associated with the test statistic

$$(x, \sigma^2) \mapsto \frac{1}{\sigma^2} \sum_{i=1}^{n} \{(X_{i1} - \hat{\xi}_i)^2 + (X_{i2} - \hat{\xi}_i)^2\}$$

and the corresponding $\mathsf{ChiSq}(n)$ or χ^2_n distribution expected when σ^2 is the true variance. Had we chosen not to eliminate $\xi_1, \ldots, \xi_n$ directly, we could have employed the same Kolmogorov–Smirnov test procedure above, which amounts to using how closely $\{(x_{ij} - \xi_i)/\sigma : i = 1, \ldots, n;\ j = 1, 2\}$ resembles a sample from $\mathsf{N}(0, 1)$ as a measure of how typical x is relative to $(\xi_1, \ldots, \xi_n, \sigma^2)$.

For illustration, we simulate data of size $n = 100$, with $\Xi_i \overset{\text{iid}}{\sim} \mathsf{N}(0, 1)$ for $i = 1, \ldots, n$ and $\Sigma = 1$. The effect of our proposed regularization is visualized in Figure 5. First note that $\lambda = 0$ corresponds to the usual negative log-likelihood function, which reaches its maximum at approximately $\Sigma^2/2 = 0.5$, consistent with the theoretical analysis. Second, as λ increases, the minimizer of the typicality-based objective function gradually shifts from $\Sigma^2/2 = 0.5$ to $\Sigma^2 = 1$, suggesting that the proposed estimator is consistent, at least for sufficiently large λ. While there is surely room for further investigation, we think it is safe to conclude that our application of the typicality principle resolves the Neyman–Scott paradox and, more general, the shortcomings of the principle of maximum likelihood.

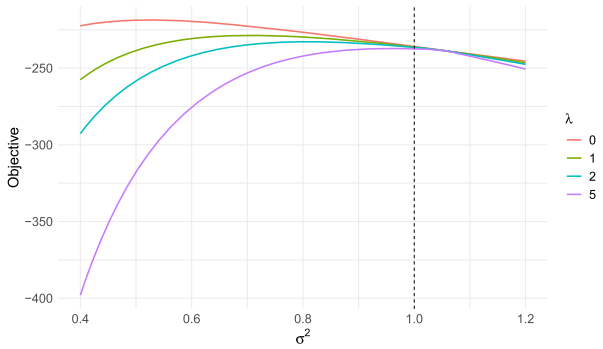


Figure 5: The objective function $\sigma^2 \mapsto \rho_\lambda^{\text{typ}}(x, \sigma^2)$ with varying choices of λ. The true Σ^2 value is denoted as the vertical dashed line.

4.3 Stein's Mean Vector Length

Let X be a n-dimensional normal random vector with unknown mean vector Θ and identity covariance matrix. Suppose that the quantity of interest is $\Phi = \|\Theta\|$, the Euclidean length of the mean vector. Inference on Φ is non-trivial, as pointed out by Stein (1956, 1959), also listed in Fraser et al. (2018) as one of the "challenge problems" advanced by the late Sir D. R. Cox.

Start with a reparametrization of the model parameter Θ, an n-dimensional mean vector, as the pair (Φ, Δ), where Φ is the length and $\Delta = \Theta/\Phi$ is the unit vector direction. From this perspective, we see that Δ is a $(n-1)$-dimensional nuisance parameter so, if n is relatively large, then we can expect challenges with marginalization, similar to those encountered in the Neyman–Scott problem above. Indeed, Stein demonstrated that the maximum likelihood estimator, $\hat{\phi}_X = \|X\|$ has non-trivial and non-vanishing upward bias, so the maximum likelihood estimator systematically overestimates Φ. Two standard strategies for eliminating nuisance parameters involve using log-profile and log-marginal likelihoods which, in this case, are given by

$$\ell_x^{\text{PRO}}(\phi) = -\tfrac{1}{2}(\|x\| - \phi)^2 \quad \text{and} \quad \ell_x^{\text{MAR}}(\phi) = \log q_{\phi^2}(\|x\|^2),$$

where q_{ϕ^2} is the density function for a non-central chi-square distribution with n degrees of freedom and non-centrality parameter ϕ^2. It is easy to see from the log-profile likelihood function above, and it follows immediately from the familiar invariance property of maximum likelihood estimators, that the maximum profile likelihood estimator of Φ is the same as the full maximum likelihood estimator, $\hat{\phi}_X = \|X\|$, which is unsatisfactory as described above. The log-marginal likelihood maximizer has no closed-form expression, but it closely resembles the simple method of moments estimator $\tilde{\phi}_X = (\|X\|^2 - n)^{1/2}$, which has a built-in upward bias correction.

As an alternative, after eliminating the nuisance parameters via profiling, we employ our typicality-motivated regularization with objective function

$$\rho_\lambda^{\text{typ}}(x, \phi) = \ell_x^{\text{PRO}}(\phi) - \lambda \log\bigl[\min\{Q_{\phi^2}(\|x\|^2),\, 1 - Q_{\phi^2}(\|x\|^2)\}\bigr], \tag{4.1}$$

where Q_{ϕ^2} is the non-central chi-square distribution function corresponding to the density q_{ϕ^2} defined above. Figure 6 shows plots of this function for several different values of λ, based on simulated data of size n with true $\Phi = 4\sqrt{10} \approx 12.65$. When $\lambda = 0$, so that the objective function is the log-profile likelihood, the upward bias is clear, and the maximizer is $\|X\| \approx 15.9$. As λ increases, the peak of the objective function gets sharper and shifts down towards the

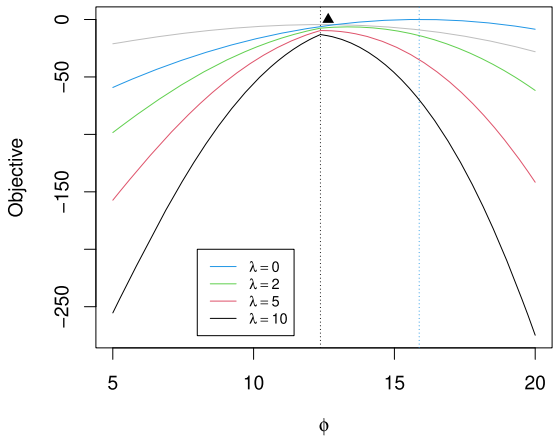


Figure 6: Plots of the objective function (3.1), specialized for inference on the normal mean vector length in (4.1), for several different values of the tuning parameter $\lambda \geqslant 0$. Gray line is the log marginal likelihood function and the black triangle marks the true value of Φ in this simulation. The dotted blue and black lines correspond to the objective function maximizers with $\lambda = 0$—corresponding to the maximum likelihood estimator—and $\lambda = 10$, respectively.

true Φ; indeed, the objective function maximizer with $\lambda = 10$ is ≈ 12.4. For comparison, the log-marginal likelihood is shown in gray and, while its maximizer is a satisfactory estimator, it has a rather dull peak indicating a potential loss of efficiency.

We also carried out a brief simulation study to compare different estimators. With $\lambda = 10$, our estimator has significantly smaller mean squared error than the naive maximum likelihood estimator but very similar mean square error compared to the maximum marginal likelihood estimator. The Bayesian literature offers a non-informative prior for Θ designed specifically for inference on Φ (e.g., Tibshirani, 1989). This penalizes large ϕ values, which makes the corresponding maximum *a posteriori* estimator appealing. But this prior has a strong built-in "λ," which is $O(n)$. So, this approach penalizes too much, pushing the estimates too close to 0, hence is not competitive with our estimator or the usual marginal maximum likelihood estimator.

5 Reliable Uncertainty Quantification

5.1 Typicality Principle, Revisited

Of course, there is more to statistics and data science than point estimation, and it turns out that the typicality principle has implications beyond that in Section 3. When prior information about Θ is vacuous, it is mathematically impossible to derive a posterior distribution for Θ, given x, in a rigorous, justifiable way. With only data x and model $\{\mathsf{P}_\theta : \theta \in \mathbb{T}\}$, all we can possibly say about a hypothesis H is if x is typical or atypical relative to P_θ for $\theta \in H$. Therefore, any legitimate attempt to quantify uncertainty about Θ given x is simply choosing to interpret typicality—a characteristic of data given a value of the parameter—as a statement about that value of the parameter given data. This is what's at the heart of falsificationism and Mayo (2014)'s

> Frequentist Principle of Evidence: Drawing inferences from data requires considering the relevant error probabilities associated with the data-generating process.

Neyman's frequentism stops with the intuitive version of the typicality principle:

- specify a test statistic, a significance level, and a rejection region such that, if hypothesis H is true, then the event "test statistic falls in the rejection region" has (error) probability no more than the stated significance level; and
- say data x is sufficiently atypical relative to H if the test statistic based on x falls in the stated rejection region.

That is, frequentists make no attempt to quantify uncertainty about the truthfulness of H given x—they are satisfied with an error-probability-controlling decision rule. Attempts to find the Bayesian–frequentist "Holy Grail," namely a framework that provides data-dependent measures of support for or degrees of belief in the truthfulness of H while remaining consistent with Mayo's frequentist principle of evidence, go further. This includes default-prior Bayesian inference (e.g., Jeffreys, 1946; Datta and Ghosh, 1995; Berger et al., 2024), Fisher's fiducial inference (e.g., Fisher, 1933, 1935a; Zabell, 1992) and its generalizations (e.g., Fraser, 1968; Hannig et al., 2016; Xie and Singh, 2013), Dempster–Shafer theory (e.g., Dempster, 1966, 2008; Shafer, 1976, 1982), and inferential models (e.g., Martin and Liu, 2013, 2015; Martin, 2025b). One thing these approaches have in common is that, at least at face value, they fail to recognize that typicality as the sole fundamental notion to work with, which creates confusion and prevents progress. By putting typicality front and center, we can clear up this confusion. In particular, we show what kinds of properties are consistent with a measure of typicality and, in turn, demonstrate what it takes to find the Holy Grail of reliable and principled uncertainty quantification.

In order to help orient our pursuit, we start by giving a formal counterpart to the informally stated typicality principle in Section 3. We follow this up with a clarification of several terms and concepts introduced in the formal statement. (Note that $2^{\mathbb{T}}$ denotes the power set of $\mathbb{T}$, the class of all possible hypotheses about Θ.)

Typicality Principle (formal version)**.** *Given a level $\alpha \in (0,1)$ and typicality measure $\tau_x : 2^{\mathbb{T}} \to [0,1]$ that is large when x is consistent with H and small otherwise, if data x is sufficiently atypical relative to the posited model and hypothesis H, in the sense that $\tau_x(H) \leqslant \alpha$, then hypothesis H is unwarranted at level α.*

There are two key features inherent in a measure of typicality that will assist in making the abstract principle above more concrete. First, typicality of x relative to a general, non-singleton H has no direct meaning—the typicality of x relative to H can only be assessed indirectly via the typicality of x relative to θ for individual θ in H. That is, one must first define typicality of x relative to simple, singleton hypotheses $H = \{\theta\}$, for $\theta \in \mathbb{T}$, and then derive the typicality $\tau_x(H)$ of x relative to general H from the primitive typicality $\tau_x(\theta) = \tau_x(\{\theta\})$. An important observation is that, unlike a Bayesian posterior for a continuous parameter that assigns zero posterior probability to "$\Theta = \theta$" for all θ, the typicality $\tau_x(\theta)$ of x relative to θ is generally non-zero. To remain consistent with our falsificationist motivations, it is essential that a hypothesis be falsified if and only if all those hypotheses that entail it are falsified. This means data is sufficiently atypical relative to "$\Theta \in H$" if and only if it's sufficiently atypical relative to "$\Theta = \theta$" for all $\theta \in H$. Mathematically, it means that $\tau_x(H)$ cannot exceed $\sup_{\theta \in H} \tau_x(\theta)$ since, otherwise, it would be possible for x to be typical relative to H but sufficiently atypical relative to all θ in H. Since $\tau_x(H)$ for non-singleton hypotheses H is derived from the primitive $\tau_x(\theta)$, and since there is no practical advantage to taking $\tau_x(H)$ smaller than necessary, we take

$$\tau_x(H) = \sup_{\theta \in H} \tau_x(\theta), \quad H \subseteq \mathbb{T}. \tag{5.1}$$

So, $H \mapsto \tau_x(H)$ is a maxitive set function (Molchanov, 2005, Ch. 1.2), just like a consonant plausibility function (Shafer, 1976, Ch. 10) or a possibility measure (Dubois and Prade, 1988).

Second, typicality suggests a suitable calibration property. If $\Theta = \theta$, then we expect the random variable $\tau_X(\theta)$ to be moderate to large but, of course, it is possible that $\tau_X(\theta)$ is small, and this latter event should have small probability. So, when we say that "x is atypical relative to θ," so that $\tau_x(\theta)$ is small, we mean that a P_θ-rare event has occurred. Then it makes sense to require that $\tau_X(\theta)$ be stochastically no smaller than $\mathsf{Unif}(0,1)$, as a function of $X \sim \mathsf{P}_\theta$, i.e.,

$$\sup_{\theta \in \mathbb{T}} \mathsf{P}_\theta\{\tau_X(\theta) \leqslant \alpha\} \leqslant \alpha, \quad \alpha \in [0,1]. \tag{5.2}$$

See Martin (2025b) for more discussion of this calibration property.

Putting everything together, under the typicality principle with a typicality measure satisfying the above properties, there are a number of desirable inference-related consequences. Two specific consequences follow immediately from the discussion above.

Proposition 1. *A test that rejects the hypothesis H if and only if x is sufficiently atypical relative to H at level α, i.e., rejects if and only if $\tau_x(H) \leqslant \alpha$, controls the frequentist Type I error:*

$$\sup_{\theta \in H} \mathsf{P}_\theta\{\tau_X(H) \leqslant \alpha\} \leqslant \alpha, \quad \alpha \in [0,1], \quad H \subseteq \mathbb{T}. \tag{5.3}$$

Proposition 2. *The set $C_\alpha(x) = \{\theta \in \mathbb{T} : \tau_x(\theta) \geqslant \alpha\}$, i.e., the collection of all θ's such that x is not sufficiently atypical relative to θ at level α, is a $100(1-\alpha)\%$ confidence set for Θ:*

$$\sup_{\theta \in \mathbb{T}} \mathsf{P}_\theta\{C_\alpha(X) \not\ni \theta\} \leqslant \alpha, \quad \alpha \in [0,1].$$

These are exactly the kind of frequentist error rate control guarantees that ensure basic reliability of inference. The property in (5.3) holds for all fixed $H \subseteq \mathbb{T}$, but reliable uncertainty quantification requires that H can vary and that similar reliability properties are maintained. For this reason, the uniform-in-H version of (5.3) established in Cella and Martin (2023) is relevant. Finally, since τ_x has at least some loose resemblance to a Bayesian or fiducial posterior distribution, it is natural to consider formal decision-making wherein the "optimal" action is defined as that which minimizes a suitable upper expected loss—a Choquet integral (e.g., Troffaes and de Cooman, 2014, App. C)—with respect to τ_x. These details are beyond our present scope, but the interested reader can consult Martin et al. (2025).

5.2 Putting the Principle into Practice

In light of the developments in Section 3, there is a conceptually straightforward path to getting from the typicality principle to a practical typicality measure that has the properties discussed in the previous section. Let $\check{\theta}_x = \check{\theta}(x)$ denote the minimizer of the objective function $\theta \mapsto \rho_\lambda^{\text{typ}}(x,\theta)$ defined in (3.1), which depends implicitly on the choice of $\lambda \geqslant 0$, with $\lambda = 0$ leading to the usual maximum likelihood estimator. Now define

$$\tau_x(\theta) = \mathsf{P}_\theta\big\{R_\lambda(X,\theta) \geqslant R_\lambda(x,\theta)\big\}, \quad \theta \in \mathbb{T}, \tag{5.4}$$

where $R_\lambda(x,\theta) = \rho_\lambda^{\text{typ}}(x,\theta) - \rho_\lambda^{\text{typ}}(x,\check{\theta}_x)$. While the ingredients might be unfamiliar, the operation is a familiar one: it's the p-value associated with a test of the hypothesis "$\Theta = \theta$" using data x and test statistic $R_\lambda(x,\theta)$. This is also a version of the probability-to-possibility transform

(e.g., Hose, 2022; Dubois et al., 2004), critical to the developments in Martin (2022b). Below we discuss the properties of τ_x in (5.4).

Since the right-hand side of (5.4) is a probability, the left-hand side clearly cannot exceed 1. But it actually attains the value 1 at $\theta = \check{\theta}_x$, the estimator from Section 3. Therefore, $\check{\theta}_x$ can legitimately be called the "maximum typicality estimator," and is further distinguished as being the point contained in all of the confidence sets $C_\alpha(x)$ as α ranges from 0 to 1. That $\theta \mapsto \tau_x(\theta)$ attains the value 1 further implies that the typicality measure $H \mapsto \tau_x(H)$ as in (5.1) is a genuine possibility measure and, therefore, can be interpreted as a coherent, fully conditional upper probability. This is in addition to the desirable frequentist properties in Section 5.1.

Implementation of the proposed framework requires that we can evaluate the contour defined in (5.4). We do not have sufficient space here to address computation, but we direct the interested reader to Jiang et al. (2023) and Martin (2025a) for details; see, also, Martin (2025b).

5.3 Relation to Other Statistical Principles

To place the foregoing discussion in context, it is helpful to revisit the *likelihood principle* (e.g., Birnbaum, 1962; Basu, 1975; Berger and Wolpert, 1984), which states that all of what is relevant in the data for inference on Θ is captured by the shape of the likelihood function. This might seem intuitive and harmless, since the commonly used maximum likelihood estimators and likelihood ratio statistics only depend on the shape of the likelihood function. On closer inspection, however, the things we commonly do with these summaries, e.g., p-value calculations, rely on sampling distributions under the posited model and, since sampling distributions are not determined by the observed likelihood, inference based on these violates the likelihood principle. On its own, this violation is of no concern, but becomes potentially problematic in light of Birnbaum's theorem stating that the likelihood principle is equivalent to conjunction of the more commonsense sufficiency and conditionality principles; accordingly, a violation of the likelihood principle implies a violation of at least one commonsense principle, hence the controversy. Doubts about the reach of Birnbaum's theorem, first back in Durbin (1970) and more recently in Evans (2013) and Mayo (2014), have only fueled the controversy.

We emphasize that the typicality principle is unapologetically orthogonal to the likelihood principle. It is clear, for example, that unless $\lambda = 0$, the minimizer $\check{\theta}(x)$ of the objective function (3.1) is not determined by the shape of the likelihood alone. The point is that data being "(a)typical" can only be assessed relative to the posited sampling model, which, again, is not determined by the observed likelihood function. If one expects their inferences to favor models that adequately predict the observed data, then observed likelihood alone is not enough.

From the perspective of Sections 5.1–5.2 above, there is more that can be said even if we ignore the typicality-motivated penalty term r_λ^{typ} or, equivalently, if we take $\lambda = 0$ in (3.1). In that case, $R_0(x, \theta)$ is just the log-relative likelihood which, itself, only depends on the shape of the observed likelihood function. It is easy to verify that inference based on the typicality measure (5.4) satisfies the sufficiency principle. It can also readily facilitate conditioning on an ancillary statistic, if the user so desires, but we refer the reader to Martin (2022b, Sec. 6) for these details. But despite the close connection to the relative likelihood, inference based on the typicality measure (5.4) do still violate the likelihood principle. The outer P_θ-probability calculation is the culprit, as this brings in aspects of the model not included in the relative likelihood itself. As Martin (2024) argues, however, the modified version of the typicality measure (5.4),

$$\tau_x^{\text{LP}}(\theta) := \sup \mathsf{P}_\theta\{R_0(X, \theta) \geqslant R_0(x, \theta)\}, \quad \theta \in \mathbb{T}, \tag{5.5}$$

where the outer supremum is over all those probabilities P_θ whose relative likelihood has the same shape as $R_0(x, \theta)$, retains the above frequentist properties and simultaneously satisfies the likelihood principle. The above modification makes crystal clear the cost of satisfying the likelihood principle: since $\tau_x(\theta) \leqslant \tau_x^{\text{LP}}(\theta)$ for all x and θ, usually with strict inequality "$\ll$," and since smaller typicality measure means more efficient inference, the cost of satisfying the likelihood principle is a potentially significant loss of efficiency. If we genuinely don't know the sampling model, then the modification in (5.5) is justified. If we are confident in the posited sampling model, then there is no need to sacrifice efficiency in order to satisfy the likelihood principle; this is why we're "unapologetic" about our proposal violating the likelihood principle.

Another statistical principle relevant to our proposal is the *prediction principle* advanced in Martin and Liu (2014, 2015). Roughly, this principle states that the sampling model should be expressed in terms of a "predictable quantity" and that inference should be based on (reliable) prediction of that predictable quantity. This turns out to be closely related to the proposed typicality principle, as we explain next. The unobserved realization of a random variable U with known distribution Q supported on $\mathbb{U}$ is an example of a predictable quantity. Martin and Liu suggest predicting the unobserved realization of U as follows: define a set-valued function $\mathcal{S} : \mathbb{U} \to 2^{\mathbb{U}}$ such that the hitting probability function $h_{\mathcal{S}}(u) = \mathsf{Q}\{\mathcal{S}(U) \ni u\}$ satisfies

$$\mathsf{Q}\{h_{\mathcal{S}}(U) \leqslant \alpha\} \leqslant \alpha, \quad \alpha \in [0, 1]. \tag{5.6}$$

A relevant example is where $\mathbb{U} = [0, 1]$, $\mathsf{Q} = \mathsf{Unif}(0, 1)$, and $\mathcal{S}(u) = [0, u]$, so that $h_{\mathcal{S}}(u) = 1 - u$ and the property (5.6) holds trivially. If the goal is to guess an unobserved realization of a random variable drawn from Q, then we have no hope but to assume that it is a "typical" realization, and we measure the typicality of a candidate u via $h_{\mathcal{S}}(u)$. The intuition is that, by (5.6), the random set $\mathcal{S}(U)$, with $U \sim \mathsf{Q}$, contains typical realizations from Q in the sense that the random hitting probability $h_{\mathcal{S}}(U)$ tends to be not-small; hence we are warranted to conclude that u values with $h_{\mathcal{S}}(u)$ small are atypical and, hence, poor predictions of the unobserved realization.

The connection to our discussion of typicality in the context of statistical inference is as follows. Martin (2018) suggests a generalized association to relate observed data x, uncertain parameter θ, and predictable quantity u. Following Section 5.2, a natural choice of generalized association is $G_\theta\{R_\lambda(x, \theta)\} = u$, where G_θ is the distribution function of the random variable $R_\lambda(X, \theta)$, as a function of $X \sim \mathsf{P}_\theta$, and $\mathbb{U} = [0, 1]$. Using the same set-valued map in the illustration above, the typicality measure in (5.4) can be re-expressed as $\tau_x(\theta) = h_{\mathcal{S}}\big[G_\theta\{R_\lambda(x, \theta)\}\big]$, for $\theta \in \mathbb{T}$, so that our proposed notion of typicality here can be directly tied to the prediction of predictable quantities emphasized in Martin and Liu (2014).

5.4 Stein's Mean Vector Length, Again

To illustrate the broader, typicality-motivated uncertainty quantification strategy described in Section 5, we revisit the example from Section 4.3. Here, with a slight abuse of notation and terminology, our goal is to construct the typicality contour $\phi \mapsto \tau_x(\phi)$ for the length Φ of the mean vector Θ. Towards this, a key observation is that the proposed objective function $\phi \mapsto \rho_\lambda^{\text{typ}}(X, \phi)$ and, hence, the corresponding $R_\lambda(X, \phi)$ depends on data X only through the squared length $\|X\|^2$. Consequently, the probability calculation in the general formula (5.4) here only depends on ϕ, since the distribution of $\|X\|^2$ under the posited model is the non-central chi-square with non-centrality parameter ϕ^2, so dependence on the nuisance parameter—the mean vector direction—drops out automatically. While elimination of the nuisance parameter is relatively straightforward in this case, it will not be so straightforward in other cases.

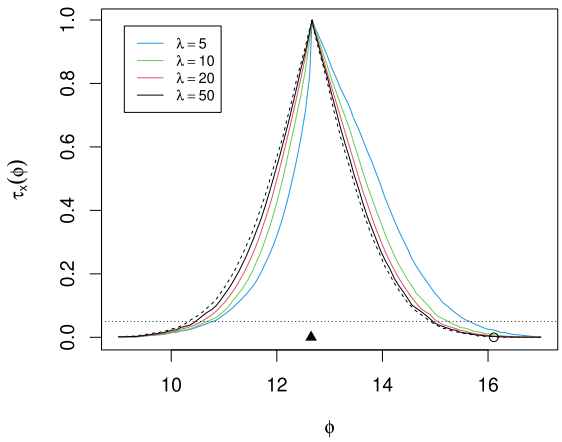


Figure 7: Plots of the typicality contour $\phi \mapsto \tau_x(\phi)$ for the mean vector length Φ based on the simulated data of size $n = 100$ and different λ values. Dashed line corresponds to the marginal likelihood-based contour suggested in Martin (2023). Open circle corresponds to the maximum likelihood estimator $\|x\|$ and the triangle is the true Φ.

For illustration, we simulated data with $n = 100$ and Θ such that $\Phi = 4\sqrt{10} \approx 12.65$; the generated data X is such that $\|X\| \approx 16.10$. Figure 7 shows a plot of the typicality contour $\tau_x(\phi)$ as defined in (5.4) for several different values of λ. The first observation is that the contours indexed by λ are all peaked at effectively the same point, which is the typicality-motivated estimator of Φ proposed in Section 3. As is common of these contour functions, they vanish away from the peak, often asymmetrically. The dotted horizontal line at 0.05 determines the 95% confidence interval for Φ as in Proposition 2; note that the maximum likelihood estimator does not belong to these confidence sets, but the true Φ does. For comparison, we also show the contour corresponding to the recommended marginal likelihood-based construction in Martin (2023), where $R_\lambda(x, \phi)$ in (5.4) is replaced by the negative log-marginal likelihood, based on the non-central chi-square distribution of $\|X\|^2$. It is difficult to imagine how another solution can be substantially better than that based on the marginal likelihood: some information about Φ is lost when the data X is reduced to $\|X\|^2$, but profiling or integrating with respect to a prior distribution will introduce bias that needs to be corrected for. An interesting observation, although not unexpected given the penalty structure in (4.1), is that the proposed typicality contour merges with the high-quality marginal likelihood-based solution in Martin (2023) as λ increases.

6 Conclusion

Motivated by the scientific perspectives of Fisher, Popper, and Tukey, here we advanced a new *typicality principle* that has a number of methodological and foundational implications for data science. First, on the methodological side, the typicality principle immediately suggests a novel strategy for regularization in the context of parameter estimation. Indeed, instead of shrinking estimators toward assumed structure (e.g., "sparsity"), our typicality-based focus encourages goodness-of-fit, ensuring that the observed data looks "typical" under the fitted model. Three non-trivial examples are presented where the method of maximum likelihood fails miserably

while our proposed typicality-focused regularization strategy is shown to be more than satisfactory. Second, on the foundational side, a more formal version of the typicality principle is readily accommodated by the general inferential model framework for provably reliable uncertainty quantification beyond point estimation, hypothesis testing, etc. This broader form of uncertainty quantification can readily accommodate von Neumann and Morganstern-style decision-making and other formal inferences, similar to Bayes, but without requiring a prior distribution and without sacrificing on error-rate control guarantees. More generally, we believe that various implementations of the proposed typicality principle will be beneficial to data science, as automated applications are expected to play a key role in the advancement of artificial intelligence. Although our focus in the present paper was on model-based inference, the notion of typicality is tied tightly to goodness-of-fit considerations, so we fully expect the typicality principle and the methodology derived from it to have an impact on modeling as well as inference.

The typicality principle and the various methodological advancements derived from it are subject to further investigations. First, modern applications involve complex models where, without the guidance of one regularization strategy or another, the tendency will be to overfit; it is precisely this tendency to overfit that leads to the poor performance of maximum likelihood estimators in the examples presented in Section 4. Then a natural follow-up to the present paper is an investigation into the performance of the proposed typicality-based regularization strategy in a class of modern, data science-related problems involving complex and over-parameterized models such as deepnets and transformers (Vaswani et al., 2017). It would also be interesting to compare our proposed with other advanced techniques, such as knowledge distillation (e.g., Hinton et al., 2015; Jiang and Liu, 2025). Second, there is also the important practical question of how to set the tuning parameter λ in (3.1). While there are so many now-standard tuning parameter selection strategies available, a relevant question is if the data-dependence inherent in our typicality-based penalty warrants new tuning parameter selection considerations. After all, compared to the usual sparsity-encouraging penalties, p-values have a meaningful scale, so new considerations concerning how to balance the contribution of the latter penalty with the likelihood may be needed. Third, the finite- and large-sample efficiency properties of the proposed typicality-based regularized maximum likelihood estimators, and of the broader uncertainty quantification developed in Section 5, are completely open for investigation.

Given the exploratory nature of this paper, the examples in Section 4 focus on non-regular models. Nevertheless, the typicality principle applies equally to regular models, correctly specified or otherwise. In regular, correctly specified parametric models, typicality is largely inactive in large samples. Specifically, for θ in an $n^{-1/2}$-neighborhood of the true parameter Θ, standard goodness-of-fit diagnostics have p-values that are bounded away from 0 under P_Θ, so $r_x^{\text{typ}}(\theta) = -\log \mathsf{pval}_x(\theta) = O_p(1)$ and varies weakly across this neighborhood. Consequently, for any fixed λ, maximizing $\ell_x(\theta) - \lambda r_x^{\text{typ}}(\theta)$ yields the same first-order behavior as the maximum likelihood estimator (e.g., $n^{1/2}$-consistency and asymptotic normality), with the penalty contributing at most a second-order perturbation. In contrast, under a misspecified model, the likelihood concentrates on parameter values that optimize Kullback–Leibler fit while producing systematic departures in aspects targeted by the goodness-of-fit statistic; then $\mathsf{pval}_x(\theta)$ becomes small near the likelihood maximizer, making $r_x^{\text{typ}}(\theta)$ large and allowing the typicality term to intervene. From this perspective, λ controls how aggressively lack-of-fit moderates likelihood-based fitting: when the model is adequate, λ is effectively inactive; when diagnostics indicate misspecification, larger λ trades some likelihood for improved typicality (better calibrated fit in the chosen diagnostic sense). Standard tuning approaches (e.g., minimizing predictive loss subject to a typicality constraint) can be used, with the understanding that λ mainly matters

when the likelihood optimum is diagnostically implausible.

The proposed framework relies on the choice of a goodness-of-fit statistic to ensure calibration, and this choice can affect statistical efficiency and power. In regular, correctly specified models, this impact is typically mild, since a wide class of goodness-of-fit statistics yield p-values that are bounded away from zero in neighborhoods of the true parameter, as discussed in the previous paragraph, rendering the typicality penalty inactive. In contrast, under misspecification or in non-regular settings, different goodness-of-fit choices emphasize different aspects of lack-of-fit, leading to trade-offs between sensitivity to particular alternatives and statistical efficiency. From the perspective of trustworthy inference, this behavior is intentional: rather than optimizing power against a specific alternative, the framework allows practitioners to select diagnostics aligned with the aspects of model adequacy most relevant to the scientific or decision-making context. Exploring optimal or adaptive choices of goodness-of-fit statistics for balancing calibration and efficiency is an important direction for future work.

Aligned with our proposed statistical principles, and motivated by other philosophical considerations, recent advancements have been made to enhance the creativity and trustworthiness of artificial intelligence (e.g., Eschker and Liu, 2024). A major challenge lies in understanding how these and other philosophical advances can help to refine today's cutting-edge methods and inspire new developments that push modern boundaries. It is also important that these philosophical contributions be of Tukey's hands-dirty, applications-oriented, "bottom-up" style, as opposed to hands-tying, "top-down" style protocol dictated in an ivory tower.

Supplementary Material

Code to reproduce all figures in this paper is included in the supplementary materials.

Acknowledgments

The authors thank the editor and an associate editor for their thoughtful and helpful comments.

Funding

Liu and Zhang are supported by the U.S. National Science Foundation grant DMS-2412629. Martin is supported by the U.S. National Science Foundation grant DMS–2412628.

References

Aldrich J (1997). R. A. Fisher and the making of maximum likelihood 1912–1922. *Statistical Science*, 12(3): 162–176. https://doi.org/10.1214/ss/1030037906

Basu D (1975). Statistical information and likelihood. Sankhyā: The Indian. *Journal of Statistics, Series A*, 37(1): 1–71. Discussion and correspondance between Barnard and Basu. https://doi.org/10.1111/j.2517-6161.1975.tb01024.x

Berger J, Bernardo J, Sun D (2024). *Objective Bayesian Inference*. World Scientific Publishing Co.

Berger JO, Wolpert RL (1984). *The Likelihood Principle*, volume 6 of *Institute of Mathematical Statistics Lecture Notes—Monograph Series*. Institute of Mathematical Statistics, Hayward, CA.

Birnbaum A (1962). On the foundations of statistical inference. *Journal of the American Statistical Association*, 57: 269–326. https://doi.org/10.1080/01621459.1962.10480660

Breiman L (2001). Statistical modeling: The two cultures. *Statistical Science*, 16(3): 199–231. https://doi.org/10.1214/ss/1009213726

Carnap R (1962). *Logical Foundations of Probability*. [2]Second edition. The University of Chicago Press, Chicago, Ill.

Cella L, Martin R (2023). Possibility-theoretic statistical inference offers performance and probativeness assurances. *International Journal of Approximate Reasoning*, 163: 109060. https://doi.org/10.1016/j.ijar.2023.109060

Datta GS, Ghosh JK (1995). On priors providing frequentist validity for Bayesian inference. *Biometrika*, 82(1): 37–45. https://doi.org/10.2307/2337625

Dempster AP (1966). New methods for reasoning towards posterior distributions based on sample data. *The Annals of Mathematical Statistics*, 37: 355–374. https://doi.org/10.1214/aoms/1177699517

Dempster AP (2002). John W. Tukey as "philosopher". The Annals of Statistics, 30(6): 1619–1628. In memory of John W. Tukey.

Dempster AP (2008). The Dempster–Shafer calculus for statisticians. *International Journal of Approximate Reasoning*, 48(2): 365–377. https://doi.org/10.1016/j.ijar.2007.03.004

Deuschel J, Foltyn A, Roscher K, Scheele S (2024). The role of uncertainty quantification for trustworthy AI. In: *Unlocking Artificial Intelligence* (C Mutschler, C Kunzmann, N Ullmann, A Martin, eds.), 95–115. Springer, Cham. Open access.

Donoho D (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4): 745–766. https://doi.org/10.1080/10618600.2017.1384734

Dubois D, Foulloy L, Mauris G, Prade H (2004). Probability-possibility transformations, triangular fuzzy sets, and probabilistic inequalities. *Reliable Computing*, 10(4): 273–297. https://doi.org/10.1023/B:REOM.0000032115.22510.b5

Dubois D, Prade H (1988). *Possibility Theory*. Plenum Press, New York.

Durbin J (1970). On Birnbaum's theorem on the relation between sufficiency, conditionality and likelihood. *Journal of the American Statistical Association*, 65(329): 395–398. https://doi.org/10.1080/01621459.1970.10481088

Edwards AWF (1992). *Likelihood*. Johns Hopkins University Press, Baltimore, MD, expanded edition. Revised reprint of the 1972 original.

Eschker SJ, Liu C (2024). Towards strong AI: Transformational beliefs and scientific creativity. arXiv preprint: https://arxiv.org/abs/2412.19938

Evans M (2013). What does the proof of Birnbaum's theorem prove? *Electronic Journal of Statistics*, 7: 2645–2655.

Fisher RA (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A*, 222: 309–368. https://doi.org/10.1098/rsta.1922.0009

Fisher RA (1925). Theory of statistical estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, 22: 200–225. https://doi.org/10.1017/S0305004100009580

Fisher RA (1933). The concepts of inverse probability and fiducial probability referring to unknown parameters. *Proceedings of the Royal Society of London. Series A*, 139: 343–348.

Fisher RA (1935a). The fiducial argument in statistical inference. *Annals of Eugenics*, 6: 391–398. https://doi.org/10.1111/j.1469-1809.1935.tb02120.x

Fisher RA (1935b). The logic of inductive inference. *Journal of the Royal Statistical Society*, 98:

39–82. https://doi.org/10.2307/2342435

Fraser DAS (1968). *The Structure of Inference*. John Wiley & Sons Inc., New York.

Fraser DAS, Reid N, Lin W (2018). When should modes of inference disagree? Some simple but challenging examples. *The Annals of Applied Statistics*, 12(2): 750–770. https://doi.org/10.1214/18-AOAS1160SF

Hacking I (1976). *Logic of Statistical Inference*. Cambridge University Press, Cambridge-New York-Melbourne.

Hannig J, Iyer H, Lai RCS, Lee TCM (2016). Generalized fiducial inference: A review and new results. *Journal of the American Statistical Association*, 111(515): 1346–1361. https://doi.org/10.1080/01621459.2016.1165102

Hinton G, Vinyals O, Dean J (2015). Distilling the knowledge in a neural network. arXiv preprint: https://arxiv.org/abs/1503.02531

Hose D (2022). Possibilistic Reasoning with Imprecise Probabilities: Statistical Inference and Dynamic Filtering, Ph.D. thesis, University of Stuttgart. https://dominikhose.github.io/dissertation/diss_dhose.pdf

Jaynes ET (2003). *Probability Theory*. Cambridge University Press, Cambridge.

Jeffreys H (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A*, 186: 453–461.

Jeffreys H (1998). *Theory of Probability. Oxford Classic Texts in the Physical Sciences*. The Clarendon Press, Oxford University Press, New York. Reprint of the 1983 edition.

Jiang Y, Liu C (2025). Estimation of over-parameterized models from an auto-modeling perspective. *Journal of the American Statistical Association*, 120(552): 2422–2434. https://doi.org/10.1080/01621459.2025.2455192

Jiang Y, Liu C, Zhang H (2023). Finite sample valid inference via calibrated bootstrap. https://arxiv.org/abs/2408.16763

Kyburg HE Jr (1987). Bayesian and non-Bayesian evidential updating. *Artificial Intelligence*, 31(3): 271–293. https://doi.org/10.1016/0004-3702(87)90068-3

Le Cam L (1986). *Asymptotic Methods in Statistical Decision Theory Springer Series in Statistics*. Springer-Verlag, New York.

Le Cam L (1990). Maximum likelihood: An introduction. *International Statistical Review*, 58(2): 153–171. https://doi.org/10.2307/1403464

Lehmann E (1983). *Theory of Point Estimation*. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, California.

Liu C (2023). Reweighted and circularised Anderson-Darling tests of goodness-of-fit. *Journal of Nonparametric Statistics*, 35(4): 869–904. https://doi.org/10.1080/10485252.2023.2213782

Liu Y, Yao Y, Ton JF, Zhang X, Guo R, Cheng H, et al. (2024). Trustworthy LLMs: A survey and guideline for evaluating large language models' alignment. https://arxiv.org/abs/2308.05374

Martin R (2018). On an inferential model construction using generalized associations. *Journal of Statistical Planning and Inference*, 195: 105–115. https://doi.org/10.1016/j.jspi.2016.11.006

Martin R (2022a). Valid and efficient imprecise-probabilistic inference with partial priors, I. First results. https://arxiv.org/abs/2203.06703

Martin R (2022b). Valid and efficient imprecise-probabilistic inference with partial priors, II. General framework. https://arxiv.org/abs/2211.14567

Martin R (2023). Valid and efficient imprecise-probabilistic inference with partial priors. *III. Marginalization*. https://arxiv.org/abs/2309.13454

Martin R (2024). A possibility-theoretic solution to Basu's Bayesian–frequentist via media.

Sankhya A, 86: 43–70. https://doi.org/10.1007/s13171-023-00323-9

Martin R (2025a). A new Monte Carlo method for valid prior-free possibilistic statistical inference. https://arxiv.org/abs/2501.10585

Martin R (2025b). Possibilistic inferential models: a review. *Journal of the American Statistical Association*. To appear: https://arxiv.org/abs/2507.09007

Martin R, Liu C (2013). Inferential models: A framework for prior-free posterior probabilistic inference. *Journal of the American Statistical Association*, 108(501): 301–313. https://doi.org/10.1080/01621459.2012.747960

Martin R, Liu C (2014). Discussion: Foundations of statistical inference, revisited. *Statistical Science*, 29: 247–251.

Martin R, Liu C (2015). *Inferential Models*, volume 147 of *Monographs on Statistics and Applied Probability*. CRC Press, Boca Raton, FL.

Martin R, Prim SN, Williams J (2025). Decision-making with possibilistic inferential models. https://arxiv.org/abs/2112.13247

Mayo D (2014). On the Birnbaum argument for the strong likelihood principle. *Statistical Science*, 29(2): 227–239. https://doi.org/10.1214/13-STS457

Molchanov I (2005). *Theory of Random Sets*. Probability and Its Applications, (New York). Springer-Verlag London Ltd., London.

Neyman J, Scott EL (1948). Consistent estimates based on partially consistent observations. *Econometrica*, 16: 1–32. https://doi.org/10.2307/1914288

Pardo L (2018). *Statistical Inference Based on Divergence Measures*. Chapman and Hall/CRC.

Popper KR (1959a). *The Logic of Scientific Discovery*. Basic Books, New York.

Popper KR (1959b). *The Logic of Scientific Discovery*. Hutchinson and Co., Ltd., London.

Shafer G (1976). *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, N.J.

Shafer G (1982). Belief functions and parametric models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 44(3): 322–352. With discussion. https://doi.org/10.1111/j.2517-6161.1982.tb01211.x

Stein C (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In: *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1954–1955, vol. I, Neyman J (Ed.), 197–206. University of California Press, Berkeley and Los Angeles.

Stein C (1959). An example of wide discrepancy between fiducial and confidence intervals. *The Annals of Mathematical Statistics*, 30: 877–880. https://doi.org/10.1214/aoms/1177706072

Stigler SM (2007). The epic story of maximum likelihood. *Statistical Science*, 22(4): 598–620. https://doi.org/10.1214/07-STS249

Tibshirani R (1989). Noninformative priors for one parameter of many. *Biometrika*, 76(3): 604–608. https://doi.org/10.1093/biomet/76.3.604

Troffaes MCM, de Cooman G (2014). *Lower Previsions Wiley Series in Probability and Statistics*. John Wiley & Sons, Ltd., Chichester.

Tukey JW (1962). The future of data analysis. In: *Breakthroughs in Statistics: Methodology and Distribution*, Kotz S, Johnson N.L. (Eds.), 408–452. Springer.

Tukey JW (1977). *Exploratory Data Analysis*. Pearson.

Tukey JW (1986). *The Collected Works of John W. Tukey. Vol. III. The Wadsworth & Brooks/Cole Statistics/Probability Series*. Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, CA. Philosophy and principles of data analysis: 1949–1964, Edited and with

comments by Lyle V. Jones, With a biography of Tukey by Frederick Mosteller.

van der Vaart A (2002). The statistical work of Lucien Le Cam. *The Annals of Statistics*, 30(3): 631–682. Dedicated to the memory of Lucien Le Cam.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. (2017). Attention is all you need. In: *Advances in Neural Information Processing Systems* (I Guyon, UV Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, R Garnett, eds.), volume 30 of *Curran Associates*. Inc.

Xie M, Singh K (2013). Confidence distribution, the frequentist distribution estimator of a parameter: A review. *International Statistical Review*, 81(1): 3–39. https://doi.org/10.1111/insr.12000

Zabell SL (1992). R. A. Fisher and the fiducial argument. *Statistical Science*, 7(3): 369–387. https://doi.org/10.1214/ss/1177011233