Econometrics and Statistics 25 (2023) 93–109

Contents lists available at ScienceDirect

Econometrics and Statistics

journal homepage: www.elsevier.com/locate/ecosta

On The Problem of Relevance in Statistical Inference

Subhadeep Mukhopadhyay[a,*], Kaijun Wang[b]

[a] CEO, United Analytics and Computational Intelligence, Inc., United States
[b] Postdoctoral researcher, Fred Hutchinson Cancer Research Center., United States

ARTICLE INFO

Article history:
Received 14 December 2020
Revised 23 October 2021
Accepted 24 October 2021
Available online 30 October 2021

Keywords:
Customized inference
Empirical Bayes
Global-to-local representation
Heterogeneity
LASER
Relevance paradox
Relevant null
Relevant prior
Reproducibility

ABSTRACT

Given a large cohort of "similar" cases one can construct an efficient statistical inference procedure by learning from the experience of others (also known as "borrowing strength" from the ensemble). But, it is not obvious how to go about gathering strength when each piece of information is fuzzy—part of a massive database of heterogeneous cases. The danger is that borrowing information from irrelevant cases might heavily damage the quality of the inference! This raises some fundamental questions for big data inference: When (not) to borrow? Whom (not) to borrow? How (not) to borrow? These questions are at the heart of the "Problem of Relevance" in statistical inference – a puzzle that has remained too little addressed since its inception nearly half a century ago [Efron and Morris, J. Am. Stat. Assoc, 67, 337 (1972)]. A *new model* of large-scale inference is developed to tackle some of the unsettled issues that surround the relevance problem. Through examples, it is demonstrated how our new statistical perspective answers previously unanswerable questions in a realistic and feasible way.[1]

1. Introduction

We are interested in the following question: Given a large number of summary statistics $z_1, \ldots z_N$ from N cases (genes, voxels, neurons, patients, customers, baseball players, etc.) how to efficiently perform customized inference (testing as well as estimation) for a particular individual case? If we assume that each z_i is *equally* informative or relevant to the case in hand, a precise individualized-inference can be delivered by learning from the experience of others (Efron and Morris, 1972; Mallows and Tukey, 1982; Efron, 2010; 2019). However, this assumption of "uniformity of relevance" breaks down when dealing with large assembly of *heterogeneous* cases, something that is becoming a norm in almost all modern data-science applications including neuroscience, genomics, healthcare, and astronomy.

Origin of the Relevance Problem. To illustrate this point, consider the following example: where for each of the $N = 3,565$ cases we are given a z-score z_i and an extra piece of information in the form of a covariate x_i (e.g., location information of voxels, genomic biomarker of patients, playing position pitcher/nonpitcher of baseball players, etc.) that captures the domain-context. We seek to perform an inference for the target case A (the red dot in Fig. 1) by taking its characteristic feature $x_A = 30$ into account.

* Corresponding author.
E-mail addresses: deep@unitedstatalgo.com (S. Mukhopadhyay), kwang2@fredhutch.org (K. Wang).

[1] The proposed relevance-integrated large-scale inference procedure is implemented using R and `LPRelevance` CRAN package (Mukhopadhyay and Wang, 2021). Available online (including all datasets): https://CRAN.R-project.org/package=LPRelevance.

https://doi.org/10.1016/j.ecosta.2021.10.013

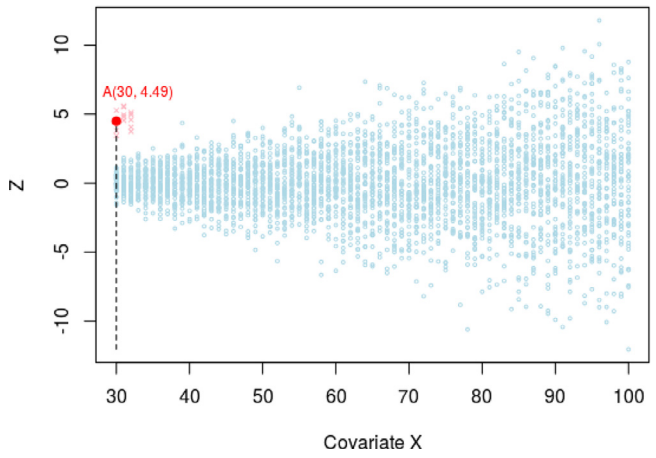

Fig. 1. The `funnel` problem: $z_i \sim \mathcal{N}(\theta_i, \sigma_i^2)$, $i = 1, \ldots, N = 3,565$ where the variability is increasing linearly as a function of x: $\sigma(x_i) = x_i/21 - 0.71$, $30 \leq x_i \leq 100$. For each x between 30 and 100, we have 50 z-values with $\theta_i = 0$. Additional 15 true signals (5 at each locations $x = 30$ to $x = 32$) with $\theta_i = 4.49$ are indicated by the light red color; they are buried in noisy background fluctuations. Obviously, the data generating process (relationship between z and x) will be considered as unknown in our analysis. The red dot is the target case A with $z_A = 4.49$ and $x_A = 30$, for which we like to perform customized inference in a completely *nonparametric* manner. Supplementary A discusses some of the practical motivation behind this `funnel` data problem.

Significant? But, Relative to What? Let's start with the most basic question: whether case A ($x_A = 30, z_A = 4.49$) is statistically significant, or at least intriguing enough to study in detail. However, the word significance only makes sense if we know *relative to what*? A declaration of statistical significance is not an absolute verdict; it's a relativistic concept that depends on what we consider as the reference or baseline. The conventional practice adopts the ensemble of aggregated observations as the 'fixed' relevant comparison set (an "absolute" frame of reference) for each individual case. This global one-size-fits-all strategy leads to troublesome results, as is visible in Figure 2.

> *"The relevance rule of 'all the cases that show up together on my desk' doesn't stand up to scrutiny, but formulating an alternative seems difficult."* (Efron, 2019)

In hindsight, it is no wonder that heterogeneity makes it silly to compare a specific case with the whole population—the comparison has to be done in relation to 'something else' other than the ensemble. But, what is that something else? What other alternatives we have? Should we instead compare with the cases that share exactly same characteristics (i.e, use 55 zs with $x = 30$ for case A)? Clearly, this is not a wise decision, since it produces too little direct data ('$N > 1000$ is necessary,' Efron, 2008b) to deliver any reliable large-scale inference result.

Empirical Bayes and Relevant Prior. The relevance problem also arises when we want to estimate the actual effect-size of a non-null case. Instead of a point estimate, researchers often prefer the whole posterior uncertainty distribution (i.e., the probability distribution of all possible values given the actual data) of the associated parameter. The main hurdle to realizing this goal is the 'prior,' which needs to be estimated in an objective manner before we apply Bayes' theorem. Traditional empirical Bayes *learns* the global prior ('fixed' for all cases) from the full sample $z_1, \ldots, z_N$ (Efron, 2016; Mukhopadhyay and Fletcher, 2018). But the practical concern here is whether the global prior is relevant for the case in hand. Surely it would be if we had one grand mélange of homogeneous observations, which, unfortunately, is not the case in most practical problems. Therefore, it is natural to ask 'which others' carry relevant information for case A. To accurately learn such a customized prior, we need hundreds or even thousands of parallel samples that are related to case A–an impractical expectation. That said, the question remains: how to design a justified recipe for estimating an individualized relevant prior? The answer to this question holds an important key to the practice of empirical Bayes in the era of big heterogeneous datasets.

The Relevance Paradox. It is evident from the discussions so far that big data inference (both simultaneous testing and estimation) poses some unique practical challenges: on the one hand the full-data-based global models are statistically efficient but not contextually relevant; on the other hand, the local inferential models are either uncalculable or absurdly noisy. Figure 3 depicts this bizarre quagmire, which shows that both global and local modes of inferences are unfruitful avenues for harnessing heterogeneous large datasets. So, can we find an algorithmic solution to reconcile this seemingly paradoxical situation emerging from the relevance problem? The "ideal" scenario would be to have a customized-inference framework (in between two extremes: global and local) that is contextually relevant and at the same time sacrifices very little, if any, efficiency. Currently, there exist no such practical theory or implementation protocols that can come close to this much sought-after goal of simultaneously improving the quality and relevance of statistical inferences across the cases.

Goals and Contributions. So, where do we stand now? Over the past several decades, great progress has been made in the field of 'large-scale inference' that helped to create a vast and impressive inventory of global inference methods; see the monographs Efron (2010), Efron and Hastie (2016), and references therein. However, as these methods are primarily useful for large homogeneous problems, the question arises as to how to modify them in order to make them applicable for

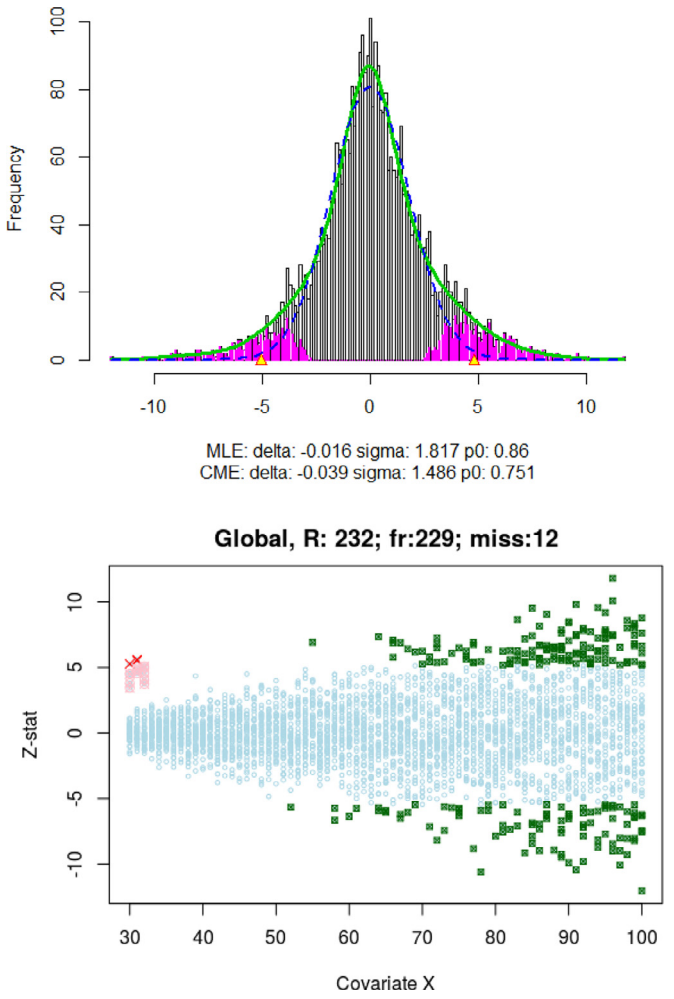

Fig. 2. Heterogeneity blankets the true signals (in light red color), and make them invisible from the global reference frame. The background variability makes the noises look "bigger" than the signals! As a result, all global large-scale inference methods mostly end up selecting loud noises. Here we display the result of local false discovery method (Efron, 2008a) that misses 12 out of 15 true signals and picks 229 false ones (green colored)–acts as a noise amplifier instead of a signal detector. Similar things happen for other methods—see Supplementary Fig..

Fig. 3. The relevance paradox: a classic "Catch-22" situation.

real-world large *fuzzy* datasets? Can we develop a *general mechanism* that can "convert" these global inference algorithms into individualized ones? These are some of the questions that motivated us to embark on this research, which has three interconnected parts:

(1) *Diagnosis*: Given $\{(x_i, z_i)\}_{i=1}^N$ how can we check *whether* the global analysis is valid or not for $x = x_0$ cases? Can we develop a nonparametric diagnostic tool?

(2) *Modeling*: If the global assumption is unreasonable, then the next question is *how* the local z-values at x_0 are different from the aggregated $z_1, \ldots, z_N$. For example, as we can see from Fig. 1 the distributional characteristics of the local z-values

at $x = 30$ are very different from the ensemble one. In fact, the marginal variance is almost 3 times that of the variance of the z-values at $x = 30$. The concept of "relevance" function, as we will soon see, critically depends on the difference ("deviance") between the local and global distributions.

(3) *Synthesis*: Finally, the question is: how to "sharpen" the aggregated messy $z_1, \ldots, z_N$ to produce a relevant comparison set? Can we do it in a fully automated and data-driven manner that works even for multivariate $\mathbf{x}$? If our specially-designed dummy z-values faithfully capture the *distributional heterogeneity* (intrinsic uncertainty and fluctuations) at x_0, then we can use them for "borrowing strength." We accomplish this goal by synthesizing LASERs–Artificial RELevant Samples. They provide a direct "one-shot" approach to convert *any* global inferential method into a customized one. A schematic representation of the algorithmic workflow is given below:

$$\text{Global inference engine + } N\text{-laser samples at } x_0 \;=\; \text{Tailor-made inference for } x_0 \text{ cases.}$$

This simple modular architecture hugely simplifies the implementation of our approach. Since, we can now utilize all the existing global inference algorithms (and associated R-routines) to produce its individualized versions. But to get there, we first have to introduce some modern nonparametric concepts and notation, which will lay the basis for a statistical theory of relevance. This is done in Section 2. The key ideas discussed here are: relevance function, global-to-local conditional density representation, and a computational recipe for generating LASERs. In Section 3, we provide a complete picture of LASER-guided customized inference, specifically touching upon the significance of relevance in microinference, empirical Bayes, and reproducible inference. Sections 3.3 and 3.5 deal with two real-applications: DTI neuroscience data and kidney data. We end with some final remarks in the last section. The Supplementary Appendix contains additional details.

Notation. By $F_Z(z) = \Pr(Z \le z)$ we denote the marginal cumulative distribution function (cdf) of the random variable Z, while $Q_Z(u)$, $0 < u < 1$ is the the respective quantile function. We drop the subscripts, whenever it is clear from the context. Note that for Z continuous $Q(u)$ is simply $F^{-1}(u)$ for $0 < u < 1$. We will denote conditional cdf $\Pr(Z \le z|X = x)$ by $F(z|x)$. The marginal and conditional densities (pdfs) are respectively expressed as $f(z)$ and $f(z|x)$. The empirical cdfs are denoted by $\widetilde{F}(z)$ and $\widetilde{F}(z|x)$. The sets $\mathbb{Z}$ and $\mathbb{Z}_x$ contain z-values for the full data and target group with cases $X = x$. Finally, $\text{Rel}(\mathbb{Z}, \mathbb{Z}_x)$ stands for relevance of the full data for the group with $X = x$.

2. A Statistical Theory of Relevance

So far we have handled the issue of relevance in an informal way. However, any serious progress in this direction, will first and foremost, require a mathematically precise statistical description of what we mean by 'relevance.' This section is organized with this goal in mind: to introduce the fundamental modeling principles that are needed to establish a general theory of relevance.

2.1. The Relevance Function

The question of relevance $\text{Rel}(\mathbb{Z}, \mathbb{Z}_x)$ is intimately tied with the question of representativeness: How representative is the full data for a target group with feature $X = x$? If the statistical (distributional) characteristics of the target z-values $\mathbb{Z}_x$ differ significantly from the statistical characteristics of the ensemble $\mathbb{Z}$, then there is a high risk of getting erroneous results using global inferential methods. Therefore, it makes sense to define relevance $\text{Rel}(\mathbb{Z}, \mathbb{Z}_x)$ as "information sharing" between the global marginal Z and local conditional $Z|X = x$, which can be measured by understanding how close (or different) the shape of $f(z|x)$ is to $f(z)$. To formalize the idea, consider the ratio

$$\frac{f(z|x)}{f(z)},$$

which captures the "amount" of information overlap (or relevance) between the combined data and cases with $X = x$. We rewrite this ratio (which is a general function of z) in the quantile-domain by substituting $F(z) = u$ in the previous expression:

$$d_x(u) \;:=\; d(u; Z, Z|X = x) \;=\; \frac{f(Q(u)|x)}{f(Q(u))}, \quad \text{for } 0 \le u \le 1. \tag{2.1}$$

to make it a proper density function over the unit interval, since

$$\int_0^1 d(u; Z, Z|X = x)\, \mathrm{d}u = 1.$$

We now formally define $d_x(u)$ as the relevance function (or kernel) that compares the distribution of the marginal Z with that of $Z|X = x$; this also justifies its notation.

Remark 1 (Shape of. $d_x$under homogeneity) If $Z \perp X$, i.e., when x contains no relevant information for z, $d_x(u)$ reduces to the uniform density

$$d_x(u) = 1, \quad 0 < u < 1 \quad (\text{for all } x).$$

Since under independence: $f(z|x) = f(z)$ for all x. This is further elaborated in Section 2.4.

Remark 2 (Special case when. $X$is binary) Suppose we are dealing with a scenario where we have two groups (baseball batter vs. pitcher or left vs. right-half brain voxels etc.) indicated by $X = 0$ and $X = 1$. In this simplistic scenario, the relevance function $d_x(u)$ reduces to

$$d_x(u) = \frac{f(Q_Z(u)|X=1)}{f(Q_Z(u))} = \frac{\Pr(X=1|Z=Q_Z(u))}{\Pr(X=1)}, \tag{2.2}$$

where the last equality follows from Bayes theorem. This suggests relevance function is proportional to the conditional probability of group assignment given $Z = z$. This is also known as the 'propensity score function,' which is used for controlling selection biases in observational studies. The alternative interpretation of our relevance function as a propensity-weighing function (in the context of binary X) was conjectured by an erudite reviewer, whom we sincerely thank.

However, it is important to keep in mind that for real-world problems, we *don't know* the right groups with comparable cases. The challenge lies in empirically deducing the relevance law $d_x(u)$ in a way that is computationally efficient and works for multivariate features **X**.

2.2. The Global-to-Local Representation

One can recover conditional density through the relevance function using the following universal recipe:

$$f(z|x) \;=\; f(z) \times d\big(F(z); Z, Z|X=x\big). \tag{2.3}$$

Justification of this representation immediately follows from the fact that $d(F_Z(z); Z, Z|X = x)$ is simply $f(z|x)/f(z)$, by virtue of the definition (2.1). For brevity's sake, we will refer $d(F_Z(z); Z, Z|X = x)$ as $d_x(z)$ throughout the article.

Interpretations. Our two-component conditional density decomposition formula (2.3) can be interpreted from several angles:

- We call it 'global-to-local' since it admits the following decomposition:

  local $f(z|x)$ = global $f(z)$ × "relevance correction" as a function of x and z.

Hence, local distributions can be created by *warping* the shape of the global distribution via $d_x(z)$. This allows us to "borrow strength from the ensemble" for efficient modeling.

- By its construction, the relevance function extracts all the 'fine details' that are *exclusive* to $Z|X = x$, i.e., different from the marginal Z. Accordingly, the shape of $d_x(z)$ contains important clues about the degree of *required customization* to go from $f(z)$ to $f(z|x)$. We will elaborate more on this in Section 2.4.

- Another noteworthy aspect of the global-to-local representation lies in expressing the relevance function $d \circ F(z)$ in the quantile or rank-transform domain, i.e., expressing it as a function of the probability integral transform $F(Z)$. This allows a *robust* way to construct the local conditional distribution from the global marginal.

Polynomials of Ranks. Perform a robust and efficient nonparametric estimation of the relevance function $d \circ F(z)$ by expressing it as a linear combination of polynomial of rank transformation $F(z)$. For Z *continuous*, one can easily construct such polynomials of rank transformation $F(z)$, according to the following recipe: standardize $F(z)$ by its mean $\mathbb{E}[F(Z)] = 1/2$ and variance $\mathrm{Var}[F(Z)] = 1/12$ to get

$$T_1(Z; F) = \sqrt{12}\big(F(Z) - 1/2\big).$$

Construct an orthonormal basis $\{T_j(Z; F)\}_{j\geq 1}$ for $\mathscr{L}^2(F)$ by applying Gram-Schmidt orthonormalization on the set of functions $\{T_1, T_1^2, \ldots\}$.

Remark 3. The polynomial construction is discussed in the continuous case since we are dealing with two-sample z-statistic in our context. The general case where Z is *mixed* (either discrete or continuous) is slightly more involved, details can be found in Mukhopadhyay and Wang (2020), Mukhopadhyay and Parzen (2020) and Mukhopadhyay (2020).

Empirical Rank-Polynomials. However, we cannot directly construct these polynomials, as they depend on the unknown F_Z. Given $z_1, \ldots, z_N$ we construct empirical polynomials $\{T_j(Z; \widetilde{F})\}_{j\geq 1}$ which by design obey the following property:

$$\int_z \widetilde{T}_j(z)\, \mathrm{d}\widetilde{F}(z) \;=\; 0 \;\text{ and }\; \int_z \widetilde{T}_j(z)\widetilde{T}_k(z)\, \mathrm{d}\widetilde{F}(z) \;=\; \delta_{jk}, \quad \text{for all } j, k.$$

where, by a slight abuse of notation, $\widetilde{T}_j(z)$ denotes $T_j(z; \widetilde{F})$. This orthonormality feature will be very useful for obtaining a neat computational algorithm that estimates the relevance function $d_x(z)$.

Remark 4. Note the our empirical LP-polynomials $\widetilde{T}_j(z)$ are functions of ranks, since $\widetilde{F}(z_i)$ is equal to rank of z_i divided by the sample size N. For that reason, we call them 'LP-polynomials', where the the letter L denotes it is rank-based and P stands for polynomial.

2.3. Estimation of Relevance Function

The relevance function admits the following LP-orthogonal series representation:

$$d_x(z) \;:=\; d(F(z); Z, Z|X=x) \;=\; 1 \;+\; \sum_j \mathrm{LP}_{j|x}\, T_j(z; F). \tag{2.4}$$

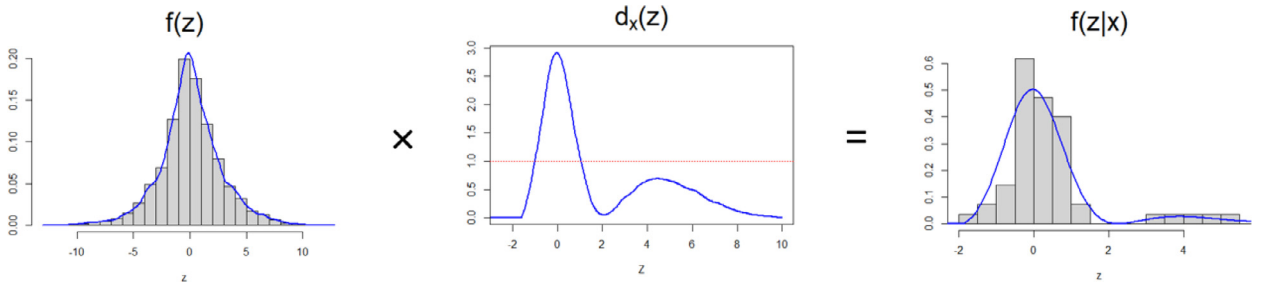

Fig. 4. It shows the mechanics of global-to-local modeling for `funnel` data. The estimated relevance function at $x = 30$ is shown in the middle plot. The conditional density estimate, shown in the last panel, seems to be 'data-consistent' in the sense that it fits the observed data excellently.

The goal is to estimate the unknown orthogonal LP-Fourier coefficients $\text{LP}_{j|x}$, which determine the shape of the relevance function. To get a compact expression for these parameters, first note that

$$\text{LP}_{j|x} = \int_z d_x(z) T_j(z; F)\, \mathrm{d}F(z). \tag{2.5}$$

since T_j's are orthonormal with respect to probability distribution F. Next, recall that $d_x(z)$ is actually $f(z|x)/f(z)$, by virtue of (2.1). Substituting this into (2.5), we immediately get the following important result.

Theorem 1. *The LP-Fourier coefficients* $\text{LP}_{j|x}$ *admit the following conditional mean representation:*

$$\text{LP}_{j|x} = \int_z T_j(z; F) f(z|x)\, \mathrm{d}z = \mathbb{E}\big[T_j(Z; F)|X = x\big]. \tag{2.6}$$

Remark 5. The practical significance of Theorem 1 resides in the fact that we can now use the whole machinery of non-parametric and machine learning regression techniques to estimate the unknown LP-coefficients by regressing $T_j(z; F)$ on the multivariate feature X. Another important point to emphasize is that Theorem 1 allows us to estimate the point-wise relevance function $d_x(z)$ by borrowing strength from the ensemble.

Robust LP-regression. However, instead of directly regressing $T_j(z; F)$ on the covariates X, we approach it in a slightly matured way to inject robustness in the procedure. Robustness is critically important for reliable and reproducible inference from noisy heterogeneous data; see Sec. 3.6. Our robust learning theory starts with the following important result:

Theorem 2. *For mixed (discrete or continuous) X we have the following important result:*

$$\mathbb{E}[T_j(Z; F_Z)|X] = \mathbb{E}[T_j(Z; F_Z)|F_X(X)], \quad \text{with probability 1}. \tag{2.7}$$

This holds, owing to the fundamental fact of the quantile function: For a general (discrete or continuous) random variable we have $X = Q_X(F_X(X))$ with probability one (Parzen, 1979). The practical consequence of this result is that we can now approximate $\mathbb{E}[T_j(Z; F_Z)|X = x]$ by projecting onto the span of LP-bases $\{T_k(x; F_X)\}_{k\geq 1}$ of X.

LP-regression: Computational algorithm. From a computational standpoint, it amounts to simply running linear regression of $\widetilde{T}_j(z)$ on the LP-basis functions of X, which takes just one line in R by calling the

$$\texttt{ML}\big(\widetilde{T}_j(z) \sim \widetilde{T}_X\big), \quad j = 1, 2, \ldots, m.$$

where `ML` can user-specified regression routine—e.g., simple linear regression (`lm`), lasso regression (`glmnet`), k-nearest neighbor (`knn`) regression, etc. Since our style of nonparametric regression allows for easy integration with stepwise variable selection or other penalized methods (e.g., AIC, BIC, or even LASSO), it produces smooth nonlinear regression function with inbuilt robustness. A non-parametric bootstrap method to quantify the uncertainty of the estimated relevance function $\widehat{d}_x$ is discussed in Supp. Appendix H.

Back to Funnel Example. Figure 4 shows the mechanics of global-to-local modeling for the `funnel` data. First thing to note is that the estimated relevance function $\widehat{d}(z; Z, Z|X = 30)$ deviates from uniformity, which means the conventional large-scale global analysis is not appropriate for cases with $x = 30$. It also reveals how the distribution of the local z-values $\mathbb{Z}_x$ differ from the aggregated z-values $\mathbb{Z}$. In other words, much information supplied by the full ensemble is *irrelevant* for the cases with $x = 30$. We need some way of distilling the relevant information by refining the mélange of heterogeneous cases. This is done by our d-perturbative scheme (Eq. 2.3), as demonstrated in Fig. 4. The relevance function $d_x(z)$ modulates the global marginal $f(z)$ to yield the local $f(z|x)$ —which has reduced the variability and an additional bump at the right tail, where those five true signals reside.

2.4. Relevance Equitability Index

For a given target case with $X = x$, should we perform a combined or a customized analysis? To satisfactorily answer this question we have to define a 'relevance equitability index' (`REI`) to measure the degree relevance between the cases with

$X = x$ and the full data. Interestingly, this information can be extracted from the shape of the relevance function $d_x(z)$. If the estimated relevance function is "flat" then it indicates that all the observations are equally relevant for the case in hand– i.e., "uniformity of relevance" is a valid assumption. In that scenario, one can safely go with the usual global inferential methods. However, deviation from uniformity (as in Fig. 4) suggests that customization is needed. In particular, one can define an appropriate REI by measuring the the unevenness of d_x:

$$\text{CUST}(x) = \int_0^1 \left\{d_x(u) - 1\right\}^2 \, \mathrm{d}u = \sum_j \left| \text{LP}_{j|x} \right|^2. \tag{2.8}$$

A higher value of CUST-statistic indicates weaker relevance (i.e., higher degree of customization required) between the cases with feature $X = x$ and the overall sample.

Remark 6. The CUST-statistic can also be interpreted as a *fairness-index*, which says how much it is fair to compare a given target case with the full ensembles of cases. Relevance and fairness are the two interrelated principles that underpin modern-day statistical inference. This is especially important for high-stakes decision making in applications such as health care, finance, hiring, criminal justice, etc.

Interestingly, one can even go to the extent of calculating the number of effective relevant samples available at each x:

$$N_{\text{rel}}(x) = N \times \text{rel}(x), \tag{2.9}$$

where

$$\text{rel}(x) = \frac{1}{1 + \sum_j |\, \text{LP}_{j|x} \,|^2} = \frac{1}{1 + \text{CUST}(x)}.$$

When $d_x(u) \equiv 1$, i.e, all LP-Fourier coefficients are zero, we have $N_{\text{rel}}(x) = N$, otherwise the effective sample size gets *dampened* by the factor $\text{rel}(x)$; see Supplementary C.

Remark 7. In summary, $\widehat{d_x}$ serves three purposes in customization: (1) quantification (measure of comparability), (2) characterization (nature of individualization required), and (3) synthesis of relevant samples. The last point is discussed in the next section.

2.5. Synthesizing LASER

The relevance function provides an easy way to generate samples from the conditional distribution of Z given $X = x$. The key is to use these samples as synthetic relevant cases that permit one to "zoom in" on a specific target case. We call these simulated cases LASERs–they are specially-designed Artificial RELevant Samples.

Learning whom to learn from. Next we provide the algorithm to generate targeted LASER samples from the full aggregated data $z_1, \ldots, z_N$. Our global-to-local representation (2.3) allows us to perform accept-reject-style sampling through $\widehat{d_x}$ to generate LASERs.

Algorithm 1 Relevance Sampler: Construction of LASER$(N; x)$

Step 0. Input: The global $\mathbb{Z} = \{z_1, \ldots, z_N\}$; Target $X = x_0$, and $\widehat{d_{x_0}}(u)$.

Step 1. If the estimated $\widehat{d_x}(u)$ is "flat" uniform density (i.e., no customization warranted), then return the full data $\{z_1, \ldots, z_N\}$, else perform steps 2-5.

Step 2. Sample z' from the global empirical cdf $\widetilde{F}$; In R perform:

$$z' \leftarrow \texttt{sample}(z_1, \ldots, z_N, \texttt{size} = 1, \texttt{replace} = \text{TRUE}).$$

Step 3. Define $u' = \widetilde{F}(z')$. Generate $U \sim \text{Uniform}(0, 1)$.

Step 4. Accept and set $z^* = z'$ if

$$\widehat{d_{x_0}}(u') > U \max_u \{\widehat{d_{x_0}}(u)\}.$$

otherwise, discard z' and return to Step 1.

Step 5. Repeat until we have obtained N samples $\{z_1^*, z_2^*, \cdots, z_N^*\}$. We denote them by LASER$(N; x_0)$, which are samples from conditional distribution $\widehat{f}(z|x_0)$.

Remark 8. LASERs can alternatively be viewed as samples from the *relevance-weighted z*-population (see step 2 and 4 above). The covariate-adaptive relevance weights $\widehat{d_x}(z_i)$ for $i = 1, \ldots, N$ act as a *"data sharpening"* tool to create tailor-made LASERs from the big messy dataset.

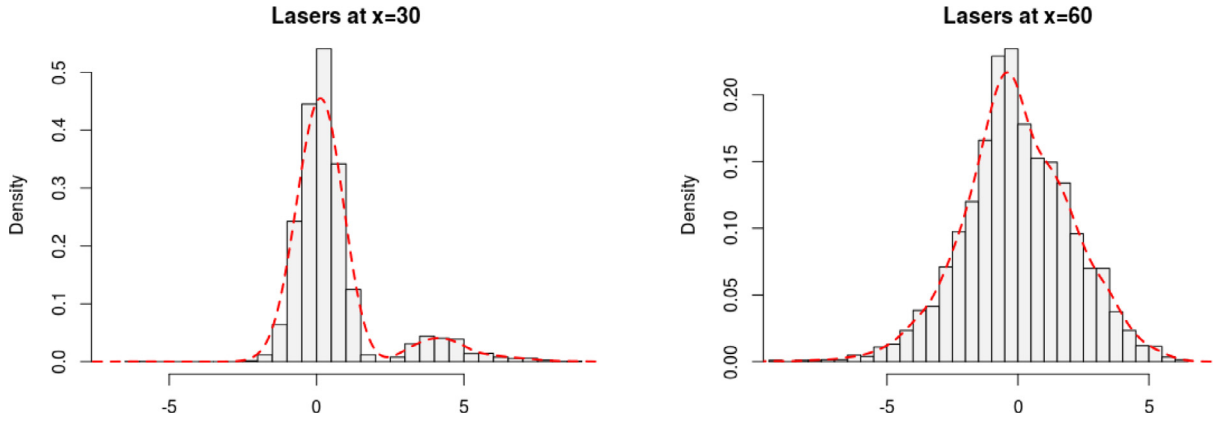

Fig. 5. Histograms of the lasers at $x = 30$ and $x = 60$ for the funnel data. Notice the contrasting shapes, in particular, the difference in the width of the two histograms.

3. Customized Statistical Inference

To deal with big messy datasets, the principle of relevance has to be an integral part of the laws of inference—which means we have to switch our attention from a one-size-fits-all global scheme to a more tailor-made one that takes into account the individual characteristics of the target case. The question becomes, how to individualize a global inference method? Here we describe the principles and protocols of LASER-guided customized inference to answer this key question. The core idea is extremely simple: feed LASERs into your favorite global inference model to make it contextually adapted.

Algorithm 2 LASER-guided customized inference: Algorithm in Pseudo-code

`Step 1.` Given $\{(x_i, z_i)\}_{i=1}^N$ the goal is to perform inference for cases with $X = x_0$.
`Step 2.` Generate $\texttt{LASER}(N; x_0)$ using the recipe ofAlgorithm 1, given in section 2.5.
`Step 3.` Perform inference at x_0 by plugging-in lasers into global algorithm:

$$\texttt{global}\big(x_0\,;\,\texttt{LASER}(N; x_0)\big).$$

Instead of $\texttt{LASER}(N; x_0)$, classical global methods use $\mathbb{Z} = \{z_1, \ldots, z_N\}$ as the fixed comparison set for *all* cases. This section, through several examples, attempts to demonstrate that learning an appropriate relevant comparison set is often a prerequisite for a valid large-scale inference method, especially when we are dealing with dissimilar cases.

The most attractive part of this algorithm is its simplicity and generalizability–and the most crucial part of this algorithm is to properly synthesize the lasers, which we feed into (any user-preferred) global inference machine. The above histograms show the $\texttt{LASER}(N; x)$ for the funnel example at $x = 30$ and 60.

As seen in Fig. 5, lasers capture the "natural variation" that is present at $x = 30$ and 60. For example, $z = 4$ can be considered as "large" for the population with $x = 30$, but is a typical occurrence for $x = 60$ population. The same argument holds for the effect-size estimation problem: $z = 4$ should be shrinked (towards zero) much more aggressively for $x = 60$ cases, compare to $x = 30$. The bottom line is context matters. Lasers allow us to contextualize any global inference method, in one step.

3.1. MicroInference

Given the z-values $z_1, \ldots, z_N$, the local-false discovery rate (Efron, 2010) is defined as

$$\text{fdr}(z) = \Pr(\text{null}|Z = z) = \frac{\pi_0 f_0(z)}{f(z)}, \tag{3.1}$$

where the last equality follows from the Bayes rule, $\pi_0 = \Pr(\text{null})$, $f_0(z)$ is the null density, and $f(z)$ denotes the marginal distribution of all the z's. We seek to develop a computable theoretical framework for covariate-adaptive false-discovery method. The following theorem shed light into the practical challenges that need to be dealt with before developing a mathematically precise solution.

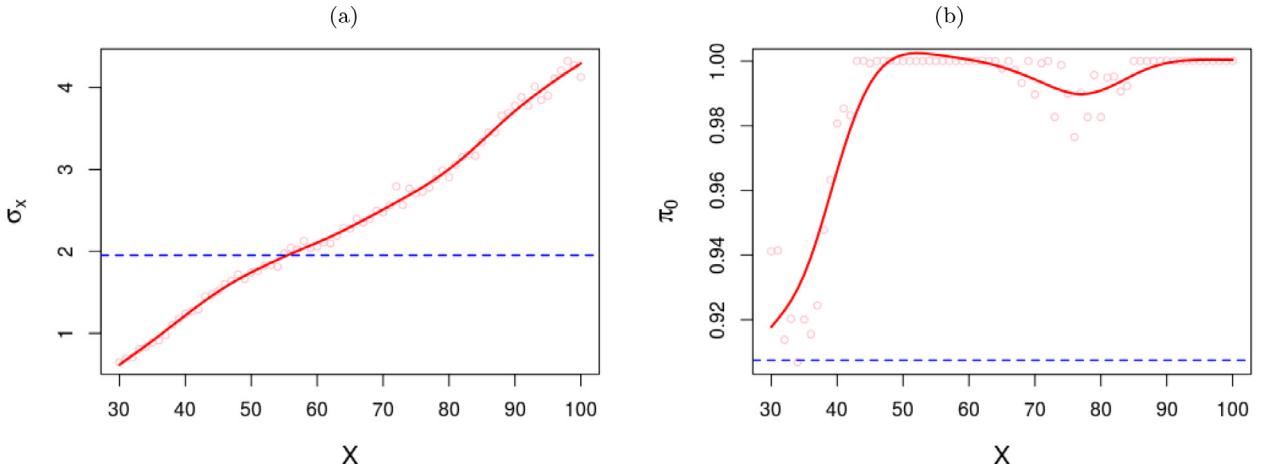

Fig. 6. Funnel data: How do the parameters of the relevant empirical nulls $\mathcal{N}(\mu_0(x), \sigma_0(x))$ and $\pi_0(x)$ change with x? As the data is already centered (i.e., $\mu_0(x) = 0$, for all x), we only focus on the $\sigma_0(x)$ and $\pi_0(x)$ which is shown respectively in the panel (a) and (b). They are calculated by applying global empirical null estimation algorithm (locfdr) on $\texttt{LASER}(N; x)$, at each x. The blue dotted lines denote the global empirical null parameter values in each plot.

Theorem 3. *The conditional false discovery rate (fdr) function* $\text{fdr}(z|x)$ *admits the following 'global-to-local' representation (close in spirit to equation 2.3) in terms of the marginal* $\text{fdr}(z)$*:*

$$\text{fdr}(z|x) \;=\; \text{fdr}(z)\left[\frac{\pi_0(x)}{\pi_0} \times \frac{f_0(z|x)}{f_0(z)} \times \frac{1}{d(F_Z(z); Z, Z|X = x)}\right], \tag{3.2}$$

where $f_0(z|x)$ *is the null distribution of* $Z|X = x$*, and* $\pi_0(x)$ *is* $\Pr(\text{null}|X = x)$*.*

Proof. The trick lies in expressing the conditional fdr function as

$$\text{fdr}(z|x) \;=\; \Pr(\text{null}|Z = z, X = x) \;=\; \frac{\pi_0(x) f_0(z|x)}{f(z|x)}, \tag{3.3}$$

which by virtue of eq. (3.1) can be re-written as

$$\text{fdr}(z)\left[\frac{\pi_0(x)}{\pi_0} \times \frac{f_0(z|x)}{f_0(z)} \times \frac{f(z)}{f(z|x)}\right].$$

Finally, apply eq. (2.3) and substitute $1/d_x(z)$ for the ratio $f(z)/f(z|x)$ to finish the proof.

The derived theory is undoubtedly beautiful but it contains uncalculable parameters! Let's focus on the "relevance correction" part inside the square brackets of (3.2): (i) The first factor $\pi_0(x)/\pi_0 \approx 1$ for most practical problems; (ii) the last factor is the "well-behaved" $d_x(z)$ function, whose estimation is already discussed in Section 2.3. (iii) Finally, we are left with the factor in the middle: ratio of relevant null $f_0(z|x)$ to the global null $f_0(z)$. How to empirically estimate the relevant null? Efficiently estimating the parameters of $f_0(z|x)$ is difficult (if not impossible), as we have too little direct data available at $X = x$. The current literature bypasses this problem by *assuming* that X is independent of Z under the null hypothesis, i.e., $f_0(z|x) = f_0(z)$. But is this a sensible assumption?

It is actually a dangerous assumption, especially when we are dealing with large heterogeneous data. Fig. 6 shows how the different parameters of the relevant null are changing as a function of x for the `funnel` example. The most dramatic one, among these two, is the leftmost one, Fig. 6(a), which shows how different the standard deviations are between the global and the relevant null. This makes the ratio of these two nulls

$$f_0(z|x)/f_0(z) \;\approx\; e^{-\frac{z^2}{2}(1/\sigma_0^2(x) - 1/\sigma_0^2)},$$

since the mean parameters are all practically zero. As seen from Fig. 6(b), $\sigma_0(x) < \sigma_0$ for all x less than 50, which means the ratio of the nulls will exponentially decrease for large $|z|$. This explains why the estimated conditional fdr function $\text{fdr}(z|x = 30)$ in Fig. 7 sharply bends inward and successfully detects all the true signals at the extreme right.

Remark 9. It is our view that insufficient regard for the relevant (empirical) null $f_0(z|x)$ is the root cause why current large-scale inference methods fail so miserably. Supp. Appendix D reviews this issue of estimating 'relevant empirical null' in more details.

Remark 10. Theorem 3 provides an indirect two-step estimation recipe for $\text{fdr}(z|x)$ by going through the marginal $\text{fdr}(z)$ function. In practice, we can do a quick approximation in a much simpler and more direct manner by simply feeding pointwise $\texttt{LASER}(N; x)$ into the global inference engine, as implemented in Fig. 7. In that sense, $\text{fdr}(z|x)$ can be considered as a "synthetic model". See Supplementary G for a step-by-step description of our microinference procedure.

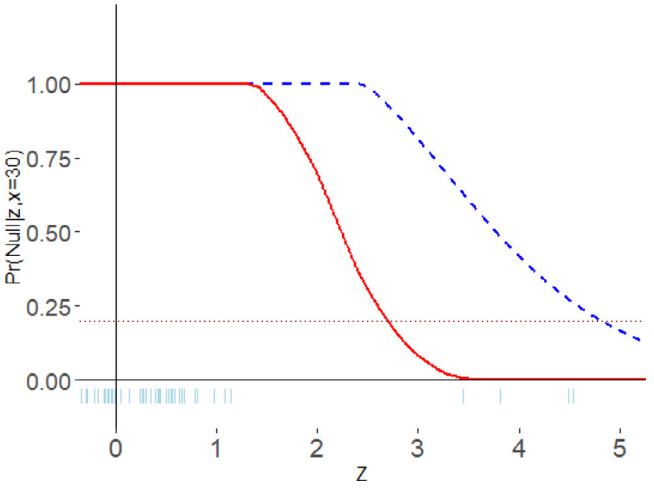

Fig. 7. The shape of $\text{fdr}(z|x=30)$ for the `funnel` example. The red is our customized fdr and blue is the global one. The horizontal red dotted line denotes 0.2 threshold.

Remark 11. There is an alternative way to express $\text{fdr}(z|x)$ by conditioning on z (instead of x as done in Theorem 3):

$$\Pr(\text{null}|Z=z, X=x) \;=\; \frac{\pi_0 f_0(z)}{f_Z(z)} \times \frac{f_0(x|z)}{f(x|z)}. \tag{3.4}$$

The first factor is simply the marginal $\text{fdr}(z)$. For X *discrete*, the second factor can be written as a conditional probability of $X=x$ given Z:

$$\Pr(\text{null}|Z=z, X=x) = \text{fdr}(z) \cdot \frac{\Pr_{\text{null}}(X=x|Z=z)}{\Pr(X=x|Z=z)} = \text{fdr}(z) \cdot \frac{\varpi^0(x|z)}{\varpi(x|z)},$$

which matches with Eq. (2.16) of Efron (2008b), after substituting $\varpi^0(x|z)/\varpi(x|z)$ by $R_x(z)$; also compare with Theorem 10.3 of Efron (2010). Note that we prefer conditioning by x to retain the applicability of our formula for mixed (either discrete or continuous) multivariate X.

3.2. MacroInference

Our goal is to perform a full-scale search and combing operation to locate the hidden signals, by retaining the individuality of each case.

Stage 1: Triage. This is a process of prioritizing or sorting cases based on their discovery proneness. For case_i, define 'Discovery Propensity Score'

$$\text{DPS}_i \;=\; -\log_{10}\left\{\Pr(\text{null}|z_i, x_i)\right\} = -\log_{10}\left\{\text{fdr}(z_i|x_i)\right\}, \quad \text{for } i=1,\ldots,N. \tag{3.5}$$

DPS-values act as an index to rank the cases. Supplementary Fig. shows the DPS-plot for `funnel` data, which correctly separates (notice the "gap") the 15 signals from the rest of the null cases. Investigators can use this ordered list for more detailed follow-up studies.

Remark 12. Unlike p-values, the DPS-values can be used as summaries of statistical evidence, since they provide a direct assessment of the probability that a finding is spurious.

Stage 2: Select. Here we "select" a small number of the most promising results by applying the false discovery threshold. Fig. 8 shows the remarkable performance of the LASER-guided multiple testing procedures; contrast this with Fig. 2. We have used both local false discovery (locfdr) and Benjamini and Hochberg (BH) as the choice of global large-scale testing methods.

3.3. DTI Neuroscience Data Analysis

Here we apply our customized large-scale testing procedure to a data set obtained from diffusion tensor imaging (DTI). This study compares brain activity of $n_1=6$ dyslexic children with $n_2=6$ normal controls. We are given two-sample z-values $z_1,\ldots,z_N$ of $N=15,443$ voxels, along with the location information: X_{1i} (distance from the back of the brain) and X_{2i} (distance from the right of the brain). Our primary interest is in: (i) microinference: to estimate the customized local fdr curve for voxels A,B, and C; and (ii) macroinference: to locate the significant voxels; see Supp. Fig..

Microinference. The blue dotted curve in Fig. 9 denotes the full-data based (global) fdr function. We are interested in individualizing this global fdr function at a voxel-level. Let's start with the case A, which has the z-value 3.95 at the location

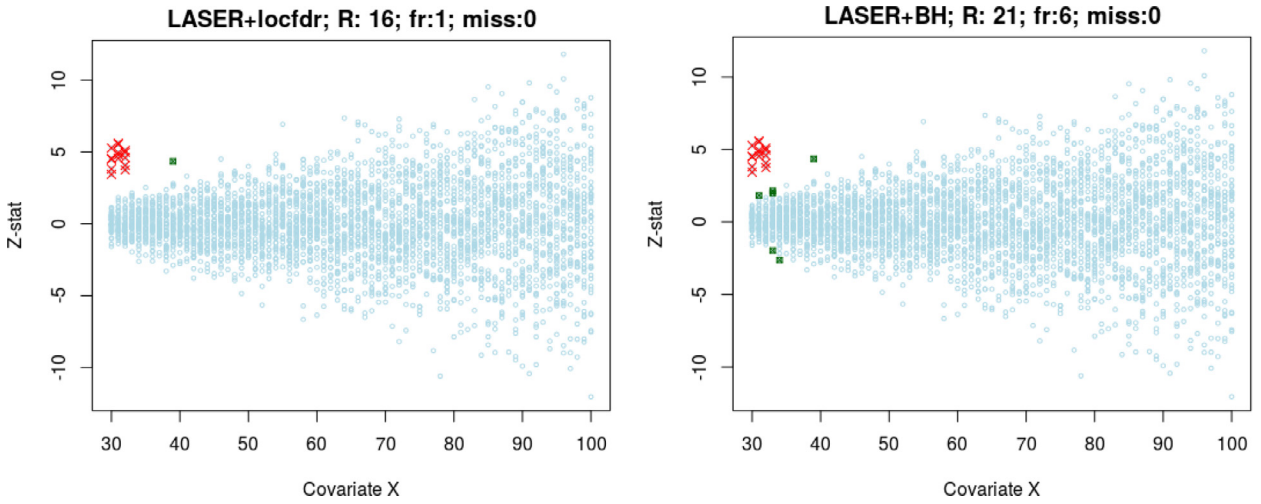

Fig. 8. Output of LASER-guided customized multiple testing result at level $\alpha = 0.05$, using local false discovery rate and BenjaminiHochberg as the global procedures. 'R' stands for number of rejections, 'fr' means number of falsely declared signals, and 'miss' denotes number of true signals missed.

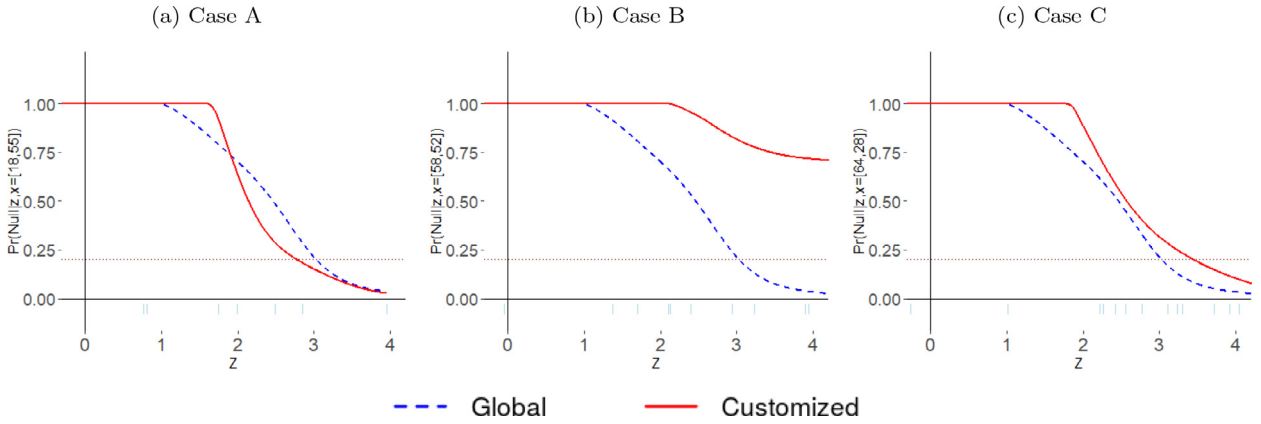

Fig. 9. The fixed global fdr function is denoted by the blue dotted line. Our customized fdr curves for voxels A, B and C are shown in red. Note their individually adaptive shapes. All the computations are done using the R-package `LPRelevance`.

$(x_{1A} = 20, x_{2A} = 55)$. At exactly the same location, we have only nine other relevant voxels! This minuscule size makes it impossible to estimate the customized-fdr function from direct relevant samples (this weird phenomenon was illustrated in Fig. 3). To tackle this problem, we generate $\texttt{LASER}(N; x_{1A}, x_{2A})$ and plug them into the locfdr function to generate the red curve in Fig. 9 (a). We follow the same procedure for the other two cases, B and C. What's most interesting about these plots is the contextually-adaptive *shape* of the $\text{fdr}(z|x)$ functions, even when the z-values are almost same!

Macroinference. Next, we move to the question of macroinference: can we find a few interesting, differentially expressed voxels? Our result is summarized in Fig. 10. Global locfdr method compares each voxel with all the remaining $N - 1$ voxels for assessing significance, which is clearly inadvisable due to heterogeneity. If we define 'signal' as the exceptional cases among their own tribe, it makes complete sense to compare each voxel with its own specially-designed LASERs for a fair comparison. Our customized locfdr declares 33 voxels to be significant at $\alpha = 0.1$. The global version, meanwhile, finds 190 discoveries. Supplementary Fig. takes a closer look at the cluster of 111 voxels around the left frontal area of the brain, which were declared significant by the global locfdr but avoided tactfully by the customized one. Surprisingly, all of these additional cases clump together at the top of the heterogeneity wave, near $x_1 = 60$ and $x_2 = 55$. This makes us suspect that they look "big" because of the unaccounted heterogeneity. The main point here is: raw magnitude of a case does not matter; what matters is how big a specific case is *with respect to* its own LASERs. It's all relative!

3.4. Empirical Bayes Inference

We now shift our focus from testing to estimation. Given a large number of sample z-statistic $z_i \sim \mathcal{N}(\theta_i, \sigma_i^2)$, the goal is to estimate the unknown mean parameters (also called effect sizes) θ_i, especially for the non-null cases. By now it is well-known that empirical Bayes provides a simple and elegant approach to effect-size estimation (Efron, 2011) by enabling

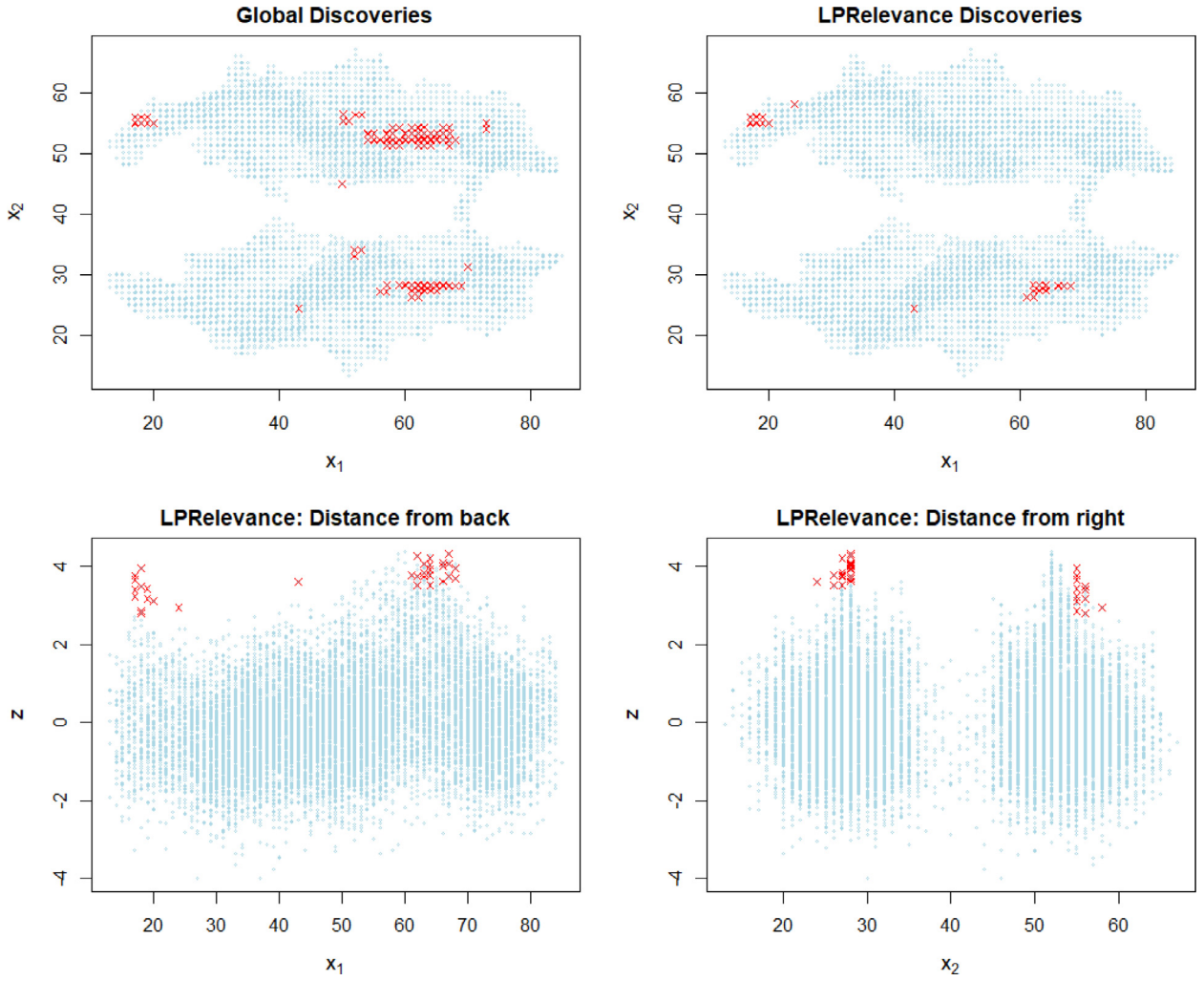

Fig. 10. Top row: Comparison of macro inference results using global and our LASER-guided customized locfdr. Bottom row: Displays our findings separately as a function of x_1 and x_2 for easy interpretation and visualization. The red crosses are 'significant' voxels.

Table 1
Effect-size estimates (posterior mean) for cases A and B: Comparing global empirical Bayes (gEB) with contextually-tailored relevant empirical Bayes (rEB) analysis.

	$\hat{E}[\Theta\|x,z]$	80% HPD Interval
global-EB:	2.42	(0.23, 4.20)
rEB:A	3.78	(3.23, 4.58)
rEB:B	0.29	(−0.41, 0.93)

one to 'learn from the experience of others.' However, the basic premise of empirical Bayes relies on the assumption that we are given a bag of samples that are relevant to each other—which, of course, is questionable for most real-world practical problems. Stuck in such a predicament, how should we proceed?

Global to Individualized Relevant Prior. The core idea is remarkably simple: rather than lumping heterogeneous, unrelated cases all together, what if we used LASERs to estimate the context-aware "personal" prior? Fig. 11 shows two cases: A ($x = 30, z = 4.49$) and B ($x = 60, z = 4.49$). They have the same z-value but in two different contexts, captured by the covariate x. The global empirical Bayes prior is shown in the blue dotted line which, by design, does not change with x. On the contrary, LASER-guided empirical Bayes priors show interesting differences: $\pi_A(\theta)$ has a longer tail with a slight bump around 2.5, and $\pi_B(\theta)$ has a much sharper peak around zero. It's impact is clearly visible on the posterior distributions. Table 1 summarizes the effect-size estimates for global as well as relevant empirical Bayes (rEB), which shows the adaptive shrinkage property of our rEB method.

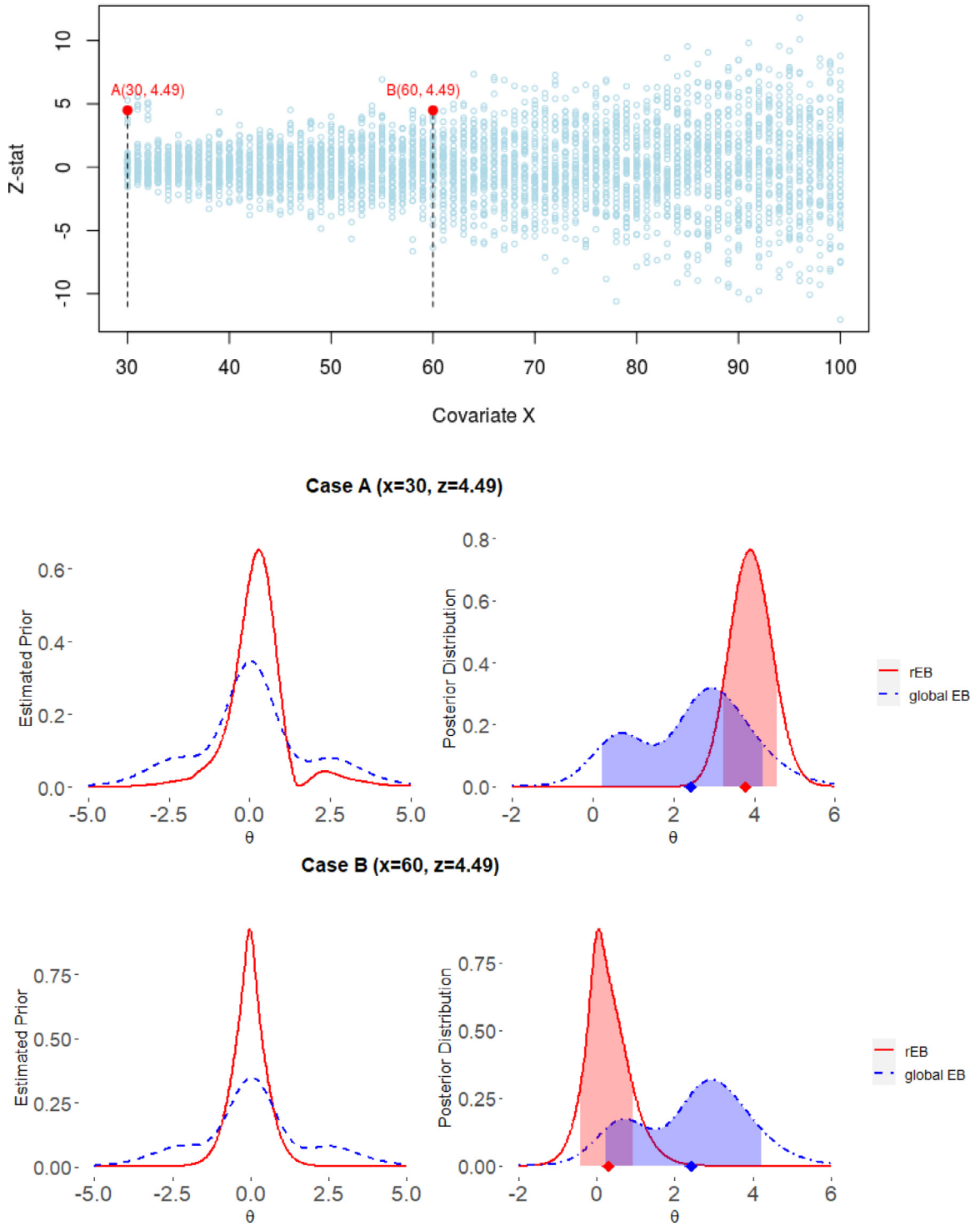

Fig. 11. Estimated prior and posteriors for two cases, A and B. Both have identical z-value, yet one is signal and the other is noise. The global empirical Bayes (gEB) prior and posterior are denoted by blue dotted curves, which are unchanging. The red curves are rEB results, changing with the contextual variable; contrast the sharpness of the rEB priors around zero. The gEB estimate (mean): 2.42, denoted as blue dot. The rEB estimates (mean): for case A (true signal) is 3.78 and for case B (noise) 0.29 are denoted as red dots. The shaded areas denote 80% highest posterior density (HPD) intervals. The rEB produces much more precise (properly centered narrower HPD) estimates—addressing both the selection and relevance bias. For a step-by-step guide on the rEB algorithm, see Supplementary F.

Remark 13. For estimating EB prior, we have used the DS-prior algorithm prescribed in Mukhopadhyay and Fletcher (2018), which is implemented in the R-package `BayesGOF`; see Supp. Appendix F. However, one can use any other method–e.g., Efron's deconvolution (Efron, 2016) or Koenker's NPMLE (Gu and Koenker, 2016). The important point here is *not* how to estimate an EB prior, but how to design 'right' relevant samples to learn from.

Remark 14. A few things are evident from this study:

1. Our rEB framework *simultaneously* balances two kinds of errors: selection bias (Efron, 2011), and relevance bias (Efron and Morris, 1972). Without relevance-correction, the usual global EB analysis would produce faulty effect-size estimates. The reason being, *where* to shrink and *how much* to shrink is directly related of the shape for the relevant EB prior, which is often very different from the global EB prior, especially in the tails; see Fig. 11. To the best of our knowledge, no previous research has achieved this dual goal of balancing the selection as well as relevance bias; rEB made it possible by selectively borrowing information from other *relevant* cases through LASERs.
2. Redeeming the curse of a real winner: From Table 1, note that the global EB estimate of $\widehat{\theta}_A$ is 2.42 and the relevance-corrected estimate is 3.78. Our rEB-adjustment prevents over-shrinkage—unfairly pulling $z_A = 4.49$ towards the zero effect for cases with $x = x_A$. As a result, rEB produces a more *fairer* effect-size estimate, by taking context into account.
3. Our style of empirical Bayes analysis is completely data-driven, which avoids the appearance of arbitrariness resulting from guessing the different parametric forms of the relevance functions (Efron, 2011, p.14).
4. One other important point is that LASERs reduce the direct contact of the prior with the actually observed data, and hence alleviate the "double-dipping" problem.

Variance Reduction by Model-averaging. For a more precise answer perform the following model averaging: (i) generate B ($B = 10$ is often enough) bags of parametric bootstrapped `LASERs`$(N; x)$ from the estimated $\widehat{d}_x(z)$; (ii) for each bag, compute the locfdr curve and the empirical Bayes posterior distribution; (iii) finally, return the "averaged" estimated curve (averaging over B runs). This "bagged estimate" reduces the variability of a single LASER-based model without affecting the bias.

Connection with Regression-adjusted Empirical Bayes. The regression-adjusted empirical-Bayes approach starts with the following model: $z_i \sim \mathcal{N}(\mu_i, \sigma_0^2)$, and $\mu_i \sim \mathcal{N}(\alpha + \beta x_i, \tau^2)$. Surely, instead of simple linear regression, one can use any nonparametric method, but the basic idea is simple: take out the mean heterogeneity by fitting a curve

$$y_i = z_i - (\widehat{\alpha} + \widehat{\beta} x_i), \;\; i = 1, \ldots, N \tag{3.6}$$

and then apply global empirical Bayes method on y_i's to make a prediction for a specific case. This is a completely justified model, provided we assume the heterogeneity is affecting only the mean of $Z|X = x$, which we call the "first-order" covariate adjusted model. Unfortunately, a practical statistician might find it an overly simplified and unrealistic assumption. Consider for example the `funnel` data, where the conditional mean is not even changing, but there exists substantial higher-order heterogeneity. Thus the real question is whether we can develop a general covariate-adjusted empirical Bayes framework that *includes* the first-order regression-adjusted model as a *special case*. Remarkably, the answer is yes, as illustrated in the next section.

3.5. Kidney Data Analysis

Fig. 12 displays the age and kidney function of $N = 157$ volunteers. Higher scores indicate better function. We are interested in the following question (Efron, 2010, Ch. 1.4): What is the empirical Bayes shrinkage estimate for the case A ($x = 55, z = 1$), denoted by the red '*' sign in Fig. 12 (a)? The main steps of our analysis are summarized below:

`Step 1.` Flattening: Looking at the data, the first obvious thing to do is to take out the mean-heterogeneity by fitting a regression-smoother. The fitted least-square line is shown in Fig. 12 (a). Next, we construct y_i's by subtracting z_i from $\widehat{\alpha} + \widehat{\beta} x_i$; see the panel (b). We call this process "flattening" of scatter.

`Step 2.` Estimation of the Relevance Function: Can we make a better inference for the target case A (see panel B) by learning from the experience of other 156 volunteers? Yes, if they are comparable—i.e., if the y-values around $x = 55$ have the same statistical (distributional) properties as the full data. This equatability information is stored in the relevance function $d_{x=55}(y)$, which contrasts the conditional density $f(y|x = 55)$ with the marginal $f(y)$. We apply the theory of section 2.3 to estimate the unknown coefficients of d_x:

$$\hat{d}_x(y) \;=\; 1 + \sum_{j=1}^{m} \widehat{\mathrm{LP}}_{j|x}\, T_j(y; F_Y), \quad \text{at } x = 55. \tag{3.7}$$

The BIC-selected coefficients all turn out to be zero: $\widehat{\mathrm{LP}}_{j|x=55} = 0$ for $j = 1, \ldots, m = 6$. That means the relevance function is "flat": $\widehat{d}_{x=55}(y) = 1$; see Fig. 12(c).

`Step 3.` "Uniformity of Relevance" Test: The flat shape of the estimated relevance function indicates that there is no "excess" heterogeneity in $f(y|x = 55)$ relative to the ensemble, and therefore the other $N = 156$ cases can be taken as admissible comparison set for borrowing information. Relevance function acts as a formal nonparametric exploratory test to validate this assumption. Accordingly, the relevance sampler (see Section 2.5) returns the full observed data $\{y_1, \ldots, y_{157}\}$ as `LASERs` for case A.

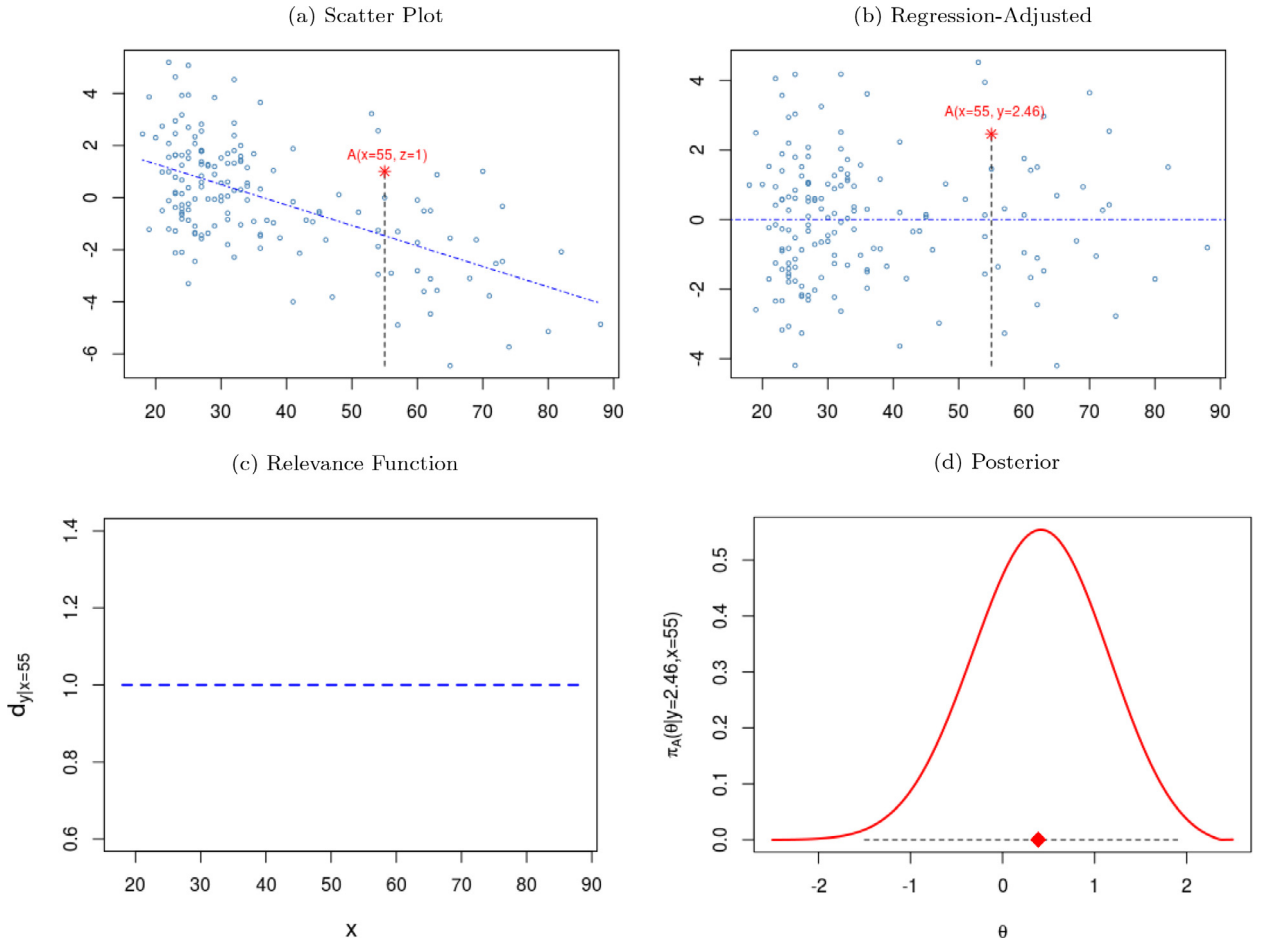

Fig. 12. Steps of kidney data analysis: (a) scatter plot of age verses kidney function; The blue dotted line denotes the least-square regression function $2.86 - 0.0786x$; The red '*' is the target case A for which inference is sought. The regression estimate at $x = 55$ is $2.86 - 0.0786 \times 55 = -1.463$. (b) flattening step $z \to y$: the regression-adjusted (x, y) plot, where $y_i = z_i - (2.86 - 0.078x_i)$; (c) The estimated relevance function $\widehat{d}_x(y)$, which interestingly takes the flat uniform shape; (d) The estimated posterior distribution $\pi_A(\theta|x = 55, y = 2.46)$ with posterior mean 0.385.

`Step 4.` Posterior Analysis of A: We borrow strength from the y-ensemble, to perform microinference for donor A $(x = 55, z = 1)$. Model: $y_i|\theta_i \sim \mathcal{N}(\theta_i, \sigma_0^2)$ where we have used

$$\sigma_0 = \texttt{IQR}(y)/1.3489 = 1.79,$$

`IQR` stands for interquartile range. Fig. 12(d) shows the estimated (empirical Bayes) posterior distribution $\pi_A(\theta|y = 2.46, x = 55)$ with posterior mean 0.385.

`Step 5.` Empirical Bayes Correction: Finally, we transform the y-domain answer in the original z-domain: $0.3845 - 1.46 = -1.0755$. Note that this "corrects" the frequentists' regression estimate -1.46 by a factor of 0.385.

This whole rEB analysis can be implemented in a few lines using `LPRelevance` R-package. See Supplementary F for more details.

Remark 15. The purpose of this example is to convey the message that our general theory of covariate-adjusted rEB *reduces* to conventional regression-adjusted empirical Bayes model ("first-order" EB model) when the relevance function $d_x(y) \equiv 1$, and customizes non-parametrically through $d_x(y)$, otherwise.

3.6. Robust Reproducible Inference

It is shown here that ignoring relevance, not only reduces the performance of an inference algorithm, but also drastically exacerbates the replicability crisis. Consider Fig. 13, which is based on two independent replications (i.e, we have used two different random seeds to generate the datasets from the same model or phenomena) of the `funnel` problem. For each one of the datasets, we applied both global and customized large-scale testing procedures. The top panel shows the results of the global locfdr method, where we have very few discoveries in common. But the most alarming part is that among these

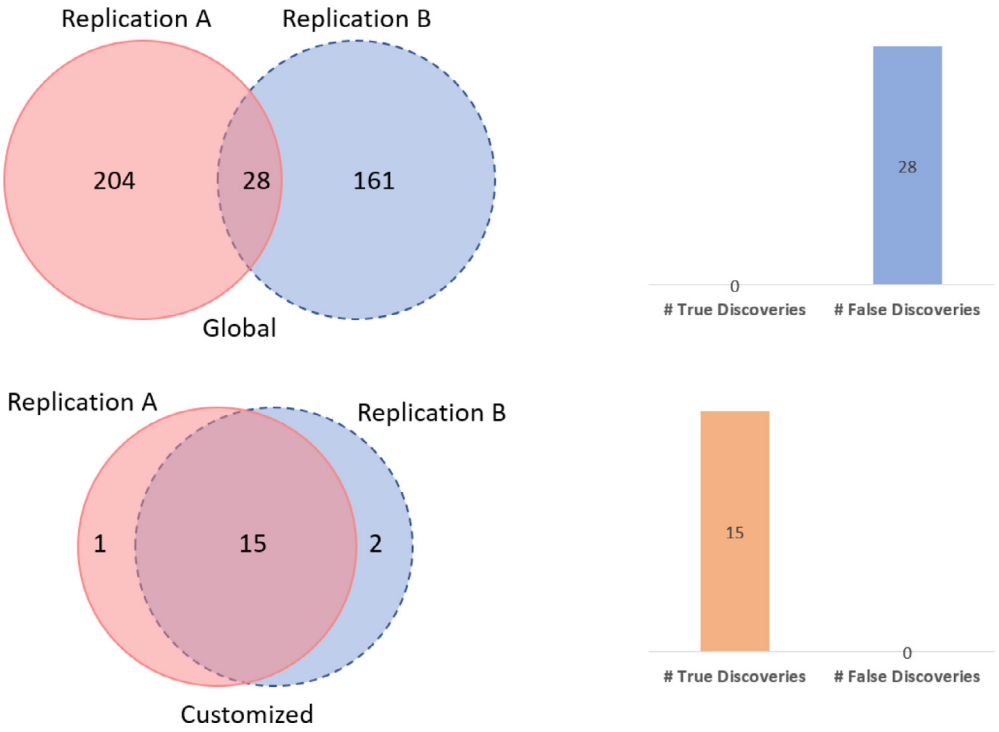

Fig. 13. (color online) Two independent replications were done (they were generated independently from the same model using two different random seeds). Top panel: global methods have only 28 discoveries in common, and shockingly all of them are false! Lesson learned: reproducible discovery $\not\Rightarrow$ correct inference. Bottom panel: the relevance-integrated customized methods are more reproducible; all the common 15 discoveries are true signals.

28 reproduced discoveries, not a single one is a true signal! This is shocking, to say the least. By contrast, the LASER-guided locfdr method shows exceptional performance: it finds all the 15 effects that are "reproducibly significant."

Remark 16. The moral of the story is this:

1. *Relevance is underrated*: When replicated experiments do not yield the same results, should we panic and call it a "crisis" or figure out why? Modern experiments are complex and delicate, with several unknown moving parts. To ensure reproducibility, we must apply the principle of relevance to design *robust large-scale inference* methods that can efficiently withstand unknown heterogeneity shocks. Robustness is the key to reproducibility.

2. *Reproducibility is overrated*: Reproducibility by its own does not serve as a "stamp of approval" to a correct inference, especially in the presence of heterogeneity —commonplace in genomics and neuroscience. The dual objective of "Relevance + Reproducibility" seems to be a better goal to strive for.

3.7. A Universal Converter

Here we discuss some practical benefits of our proposed customized-inference framework, which proceeds as follows: (i) Choose an appropriate global inference model (*any* large-scale testing or estimation method); (ii) generate $\texttt{LASER}(N; x)$ by estimating the relevance function $d_x(z)$; (iii) feed those LASERs into the selected global model to individualize the inference based on case-specific characteristics.

This modular architecture makes the computational interface extremely simple and robust. If, in the future, we want to change, upgrade, or add new global inferential procedures, none of these changes will affect the LASER-based individualization process — since we do not have to redesign the customization principle every time separately for each algorithm. This makes the whole implementation pipeline easily adaptable and specializable, which could be helpful for applied researchers and data scientists.

4. Discussions

> "*We are only beginning to recognize the many roles of borrowing strength. We need to do this more rapidly, more widely and in more diverse situations.*" (Mallows and Tukey, 1982)

Statistical inference is a problem of 'learning by comparison.' To tackle real-life modern statistical inference problems, we have to face the question: *how to compare a large number of heterogeneous parameters in a meaningful way*? The key

obstacle to addressing this question lies in the difficulty of resolving the "relevance paradox," without proper consideration of which, even a prudent statistical inference method can go awry. This paper offers the first practical theory of relevance (with precisely describable statistical formulation and algorithm) to extract individual-level customized inference from increasingly massive and heterogeneous data. The advocated robust large-scale inference technology offers a simple mantra: "personalize your inference by feeding LASERs into your global full-data-based models." It is our hope that this simple and general principle will take us close to the ultimate goal of building an inference machine with contextual adaptation, which could be a powerful tool for precision medicine, healthcare, recommendation system, defense, and national security-related applications.

Acknowledgement

This paper is dedicated to the "50 Years of the Relevance Problem." The authors would like to thank the editor, associate editor, and the two anonymous reviewers for their clear and concise suggestions. This research was inspired by a question raised by Brad Efron to S.M. We are grateful to Brad Efron and Jerry Friedman for valuable comments.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ecosta.2021.10.013

References

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Statist Soc Ser B. 57, 289–300.
Efron, B., 2008a. Microarrays, empirical Bayes, and the two-groups model. Statistical Science 23, 1–22..
Efron, B., 2008b. Simultaneous inference: When should hypothesis testing problems be combined? The Annals of Applied Statistics 197–223.
Efron, B., 2010. Large-scale inference: empirical Bayes methods for estimation, testing, and prediction, vol. 1. Cambridge; New York: Cambridge University Press..
Efron, B., 2011. Tweedies formula and selection bias. Journal of the American Statistical Association 106 (496), 1602–1614.
Efron, B., 2016. Empirical Bayes deconvolution estimates. Biometrika 103 (1), 1–20.
Efron, B., 2019. Bayes, oracle Bayes, and empirical Bayes (with discussion). Statistical Science 2, 177–201.
Efron, B., Hastie, T., 2016. Computer Age Statistical Inference, vol. 5. Cambridge University Press.
Efron, B., Morris, C., 1972. Limiting the risk of Bayes and empirical Bayes estimatorspart II: The empirical Bayes case. Journal of the American Statistical Association 67 (337), 130–139.
Gu, J., Koenker, R., 2016. On a problem of Robbins. International Statistical Review 84 (2), 224–244.
Mallows, C.L., Tukey, J.W., 1982. An overview of techniques of data analysis, emphasizing its exploratory aspects. Some Recent Advances in Statistics 33, 111–172.
Mukhopadhyay, S., 2020. United statistical algorithms and data science: An introduction to the principles. In: La rocca m., liseo b., salmaso l. (eds) nonparametric statistics, vol 339. Springer Proceedings in Mathematics & Statistics, pp. 367–377.
Mukhopadhyay, S., Fletcher, D., 2018. Generalized empirical Bayes modeling via frequentist goodness-of-fit. Nature: Scientific Reports 8 (9983), 1–15.
Mukhopadhyay, S., Parzen, E., 2020. Nonparametric universal copula modeling. Applied Stochastic Models in Business and Industry, special issue on "Data Science" 36 (1), 77–94.
Mukhopadhyay, S., Wang, K., 2020. A nonparametric approach to high-dimensional k-sample comparison problem. Biometrika 107 (3), 555–572.
Mukhopadhyay, S., & Wang, K. (2021). Lprelevance: Relevance-integrated statistical inference engine. R package version 3.2 https://CRAN.R-project.org/package=LPRelevance.
Parzen, E., 1979. Nonparametric statistical data modeling (with discussion). Journal of the American Statistical Association 74, 105–131.