Annals of the Institute of Statistical Mathematics manuscript No.
(will be inserted by the editor)

On the Indirect Elicitability of the Mode and Modal Interval

Krisztina Dearborn · Rafael Frongillo

Received: date / Revised: date

Abstract Scoring functions are commonly used to evaluate a point forecast of a particular statistical functional. This scoring function should be consistent, meaning the correct value of the functional is the Bayes act, in which case we say the scoring function elicits the functional. Recent results show that the mode functional is not elicitable. In this work, we ask whether it is at least possible to indirectly elicit the mode, wherein one elicits a low-dimensional functional from which the mode can be computed. We show that this cannot be done: neither the mode nor a modal interval are indirectly elicitable with respect to the class of identifiable functionals.

Keywords Elicitation · Point forecast · Scoring function · Loss function · Mode · Modal interval.

1 Introduction

To evaluate point forecasts, one commonly uses a scoring function, also called a loss function, which measures the inaccuracy of the forecast relative to an observed outcome. Loss functions are also used in estimation, forecast ranking and comparison, model selection, and back-testing (Gneiting and Raftery 2007; Gneiting 2011). In all of these applications, given a target statistical functional, we desire a consistent loss function, meaning that the correct value of the

This work was supported by National Science Foundation Grant CCF-1657598.

K. Dearborn
Department of Mathematics, University of Colorado Boulder
Campus Box 395, Boulder, CO, USA 80309
E-mail: krisztina.dearborn@colorado.edu

R. Frongillo (corresponding author)
Department of Computer Science, University of Colorado Boulder
1111 Engineering Dr, Boulder, CO, USA 80309
E-mail: raf@colorado.edu

functional is the Bayes act with respect to the loss. In this case, we say the loss function *elicits* the functional.

While many common statistics are elicitable, such as the mean, median, and quantiles, it is well-known that the variance is not. This impossibility follows from an observation of Osband (1985) that elicitable functionals have convex level sets, meaning mixtures of distributions with the same functional value must again have the same functional value, an axiom which the variance does not satisfy. (Indeed, mixtures generally have higher variance.) Nonetheless, several authors have pointed out that the variance is *indirectly* elicitable: one may elicit the first and second moment of the distribution, and then combine these values with a link function to obtain the variance. The minimum number of dimensions required in such an indirect elicitation scheme (for the variance, 2) is referred to as the elicitation order, or elicitation complexity, of the functional in question (Lambert et al 2008). Recently, several important non-elicitable functionals, including many risk measures such as conditional value at risk, have been shown to be indirectly elicitable with low elicitation complexity (Lambert et al 2008; Frongillo and Kash 2015; Fissler et al 2016).

Heinrich (2014) recently showed that another common statistic, the mode functional, is not elicitable, despite the fact that its level sets are convex. It is therefore natural to ask whether the mode is indirectly elicitable, and if so, determine its elicitation complexity. Our main result is that the mode has infinite elicitation complexity with respect to identifiable functionals, a relatively weak restriction (see Definitions 3 and 4 and the discussion following). Interestingly, our results also extend to modal intervals, which are elicitable, as we discuss in Section 4.

Our results show that it is impossible to develop a consistent loss function for evaluating point forecasts of the mode, even indirectly. Moreover, they cast doubt on the existence of broadly effective empirical risk minimization schemes for estimating the mode or a modal interval. Our techniques differ from previous work (Frongillo and Kash 2015), and may be applicable to other functionals of interest. We conclude with open questions, including a discussion of other notions of elicitation complexity and other properties.

2 Setting

Let $\mathcal{P}$ be a set of probability measures on a common measurable space $(\mathcal{Y}, \mathcal{F})$. For each probability measure $P \in \mathcal{P}$, denote the expectation of a random variable Y with distribution P by $E_P Y$. We will use the term "property" to refer to a statistical functional taking values in a report space $\mathcal{R}$, often a subset of $\mathbb{R}$ or $\mathbb{R}^k$.

Definition 1 (Property) A property is a functional $\Gamma : \mathcal{P} \to \mathcal{R}$ which assigns a report value to each probability measure in $\mathcal{P}$.

For example, considering probability measures on the measurable space $\mathcal{Y} = \mathbb{R}$, with $\mathcal{F}$ being the Borel σ-algebra on $\mathcal{Y}$, the mean $\Gamma(P) = E_P(Y)$ is a

real-valued property. Similarly, another real-valued property is the variance, $\Gamma(P) = E_P(Y - E_P(Y))^2$. Our focus in this paper will be the mode, which will be defined with care at the end of this section.

We next formalize our notion of consistency, which ensures that the Bayes act for a loss function coincides with the desired property.

Definition 2 (Elicits) A loss function $L : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}$ elicits a property $\Gamma : \mathcal{P} \to \mathcal{R}$ if for every $P \in \mathcal{P}$ we have $\{\Gamma(P)\} = \operatorname{argmin}_r E_P(L(r, Y))$. We say Γ is elicitable if there exists some loss function that elicits Γ. For all $k \in \mathbb{N}$ the set of elicitable properties $\Gamma : \mathcal{P} \to \mathbb{R}^k$ will be denoted $\mathcal{E}_k(\mathcal{P})$.

For example, the mean is elicited by squared loss $L(r, y) = (r - y)^2$.

The following concept of identifiability, due to Osband (1985), has played a central role in the theory of property elicitation. The definition states that each level set of the property, that is, the set of distributions sharing a particular value r of the property, can be described by a linear constraint which depends on r. Note that Steinwart et al. (2014) adopt a weaker notion of identifiability, wherein the condition need only hold for almost every level set, and call the definition below "strong identifiability"; see Section 5.

Definition 3 (Identification) A property $\Gamma : \mathcal{P} \to \mathcal{R} \subseteq \mathbb{R}^k$ is identifiable if there exists an identification function $V : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}^k$ such that for all $P \in \mathcal{P}$ we have $\Gamma(P) = r$ if and only if $E_P(V(r, Y)) = 0$. Let $\mathcal{I}_k(\mathcal{P})$ denote the class of all properties from $\mathcal{P}$ to $\mathbb{R}^k$ which are identifiable.

To illustrate, the mean is identified by the function $V(r, y) = y - r$.

Let us return to the notion of elicitation, and consider the variance $\Gamma(P) = E_P(Y - E_P(Y))^2$. As observed by Osband (1985), for a property to be elicitable it must have convex level sets: the set of distributions having the same property value must be convex. It follows immediately that the variance is not elicitable. As noted in the introduction, however, the variance can be expressed as a function, or link, of elicitable properties, for example the mean and second moment: $\Gamma(P) = E_P(Y^2) - (E_P(Y))^2$. This motivates the notion of indirect elicitability, wherein one elicits an intermediate property, and then computes a link function to obtain the original property. When confronted with non-elicitable properties, it is therefore natural to ask the minimal dimension of such an intermediate elicitable property; this is the notion of elicitation complexity (Lambert et al 2008; Frongillo and Kash 2015; Fissler et al 2016). As we explain following the definition, we further require that these intermediate properties be identifiable.

Definition 4 (Identifiable Elicitation Complexity) Let $\mathcal{I} = \bigcup_{k \in \mathbb{N}} \mathcal{I}_k(\mathcal{P})$ be the class of identifiable properties. For $k \in \mathbb{N}$, a property $\Gamma : \mathcal{P} \to \mathcal{R}$ is k-elicitable with respect to $\mathcal{I}$ if there exists an elicitable property $\hat{\Gamma} \in \mathcal{E}_k(\mathcal{P}) \cap \mathcal{I}$ and a function $f : \mathbb{R}^k \to \mathcal{R}$ such that $\Gamma = f \circ \hat{\Gamma}$. The identifiable elicitation complexity of Γ is then the minimum of all k such that Γ is k-elicitable with respect to $\mathcal{I}$.

Without imposing such a restriction on the class of intermediate properties, the definition of elicitation complexity would be trivial, as noted by Frongillo & Kash (2015): all properties of distributions on $\mathbb{R}$ have complexity 1 by first eliciting the entire distribution via set-theoretic bijections between $\mathbb{R}$ and $\mathbb{R}^{\mathbb{N}}$ (see also the discussion following Corollary 1). To justify the restriction to identifiability in particular, first note that nearly all natural elicitable properties are identifiable, including expectations, ratios of expectations, quantiles, and expectiles. Second, the results of Lambert (2018) and Steinwart et al. (2014) show that continuous non-locally-constant functionals are elicitable if and only if they are weakly identifiable, meaning identifiability is essentially necessary for continuous non-locally-constant properties $\hat{\Gamma}$ in Definition 4. Third, elicitable properties which are not identifiable are often indirectly elicitable via finite-dimensional identifiable properties, as is the case for all finite elicitable properties (those taking values in a finite set); this observation is particularly relevant as we give infinite lower bounds.

Returning to the example of the variance, we see that while it is not elicitable, its identifiable elicitation complexity is at most 2. The variance can be recovered via the function $f(x_1, x_2) = x_2 - x_1^2$ composed with the identifiable and elicitable vector-valued property $\hat{\Gamma}(P) = (E_P(Y), E_P(Y^2)) \in \mathbb{R}^2$. In this case the identification function for $\hat{\Gamma}$ is $V(r, y) = (y - r_1, y^2 - r_2)$ where $r = (r_1, r_2)$. There is a distinction between a property which is elicitable like the mean, $\Gamma(P) = E_P(Y)$, and a property which is 1-elicitable like the mean squared, $\Gamma(P) = (E_P(Y))^2$. While every elicitable real-valued property is trivially 1-elicitable via the identity function, not every 1-elicitable property is elicitable. The mean squared fails to be elicitable, but is 1-elicitable.

Finally, we define identification complexity, which trivially lower bounds identifiable elicitation complexity, a fact we use extensively in our results.

Definition 5 (Identification Complexity) For $k \in \mathbb{N}$, a property $\Gamma : \mathcal{P} \to \mathcal{R}$ is k-identifiable if there exists an identifiable property $\hat{\Gamma} \in \mathcal{I}_k(\mathcal{P})$ and a function $f : \mathbb{R}^k \to \mathcal{R}$ such that $\Gamma = f \circ \hat{\Gamma}$. Furthermore, the identification complexity of Γ is the minimum of all k such that Γ is k-identifiable.

For the remainder of this section, we turn to the mode, which we define as in Gneiting (2011) and Heinrich (2014). Letting $\varepsilon > 0$ and $P \in \mathcal{P}$, consider the cumulative distribution function F associated with P. A modal interval is any interval of the form $[x - \varepsilon, x + \varepsilon]$ to which F assigns maximal probability. Let Γ_ε denote a midpoint of a modal interval, defined as

$$\Gamma_\varepsilon(P) \in \arg\max_x \left(F(x + \varepsilon) - \lim_{z \uparrow x - \varepsilon} F(z) \right) . \tag{1}$$

Regardless of whether the modal interval is unique we can use its midpoint to define the mode of the distribution. Suppose there exists a sequence of real numbers $\{\varepsilon_n\}$ where $\varepsilon_n \to 0$ as $n \to \infty$ and a corresponding choice of midpoints of modal intervals $\{\Gamma_{\varepsilon_n}(P)\}$ converging to a real number, $\Gamma_{\text{mode}}(P)$. Then $\Gamma_{\text{mode}}(P)$ is the mode of the distribution. This definition is careful not

to assume that a probability density exists. In the case where the distribution function F is absolutely continuous and admits a continuous density p, then $\Gamma_{\text{mode}}(P)$ coincides with the global maximum of p. When working with a discrete probability distribution, $\Gamma_{\text{mode}}(P)$ corresponds to the point(s) associated with maximal probability.

We will refer to probability measures which have a well-defined and unique mode as *unimodal*. If a probability measure is unimodal and there exists a probability density associated with it, the density does not necessarily have a unique local maximum, a stronger requirement. For example, a Gaussian density is unimodal in both senses of the term, whereas a mixture of Gaussians with unit variance and strictly distinct weights does not necessarily have a unique local maximum, but does have a well-defined and unique mode and thus is unimodal. (See Section 5 for a discussion of the stronger definition.)

3 Impossibility

Heinrich (2014) demonstrates that the mode is not directly elicitable with respect to several classes of unimodal probability measures. We proceed by studying the identifiable elicitation complexity of the mode. Our main results Theorems 1 and 2 both show that the identifiable elicitation complexity of the mode is infinite with respect to two classes of probability measures. These results imply that, when restricting to identifiable intermediate properties, the mode is not even indirectly elicitable.

To begin, let $\mathcal{P}$ denote the class of unimodal probability measures defined on the real line which admit a smooth and bounded density. Below we will define a class $\mathcal{P}_{\psi,\varepsilon}$ of probability measures within $\mathcal{P}$ consisting of (finite) mixtures of normalized bump functions which will be the class of probability measures employed in Lemma 1, Theorem 1, and Corollary 1. We will denote by $\mathcal{Q} \subset \mathcal{P}$ the class of probability measures which can be expressed as a (finite) mixture of Gaussians, the focus of Theorem 2. Since each $P \in \mathcal{P}$ admits a unique probability density p, we will identify the probability measure P with its density p, and use the two interchangeably. Hence, when we choose an element $p \in \mathcal{P}$, we mean the probability density p associated with a probability measure $P \in \mathcal{P}$. Finally, $\Gamma_{\text{mode}}(P)$ and $\Gamma_{\text{mode}}(p)$ both denote the mode of the distribution P as defined in Section 2 which corresponds to the global maximum of p.

We define the bump function centered at 0 of width $2\varepsilon > 0$ as follows,

$$\psi_{0,\varepsilon}(x) = \begin{cases} \frac{1}{c_\varepsilon} \exp\left(-\frac{1}{\varepsilon^2 - x^2}\right) & |x| < \varepsilon \\ 0 & |x| \geq \varepsilon \ , \end{cases} \tag{2}$$

where $c_\varepsilon = \int_{-\varepsilon}^{\varepsilon} \exp(-1/(\varepsilon^2 - x^2))\ dx$. We then define the bump centered at x_0 to be the function $\psi_{x_0,\varepsilon}(x) = \psi_{0,\varepsilon}(x - x_0)$. Note that $\psi_{x_0,\varepsilon} \in \mathcal{P}$ and $\Gamma_{\text{mode}}(\psi_{x_0,\varepsilon}) = x_0$. Let $\mathcal{P}_{\psi,\varepsilon}$ denote the class of distributions in $\mathcal{P}$ which are finite mixtures of bump functions in the set $\{\psi_{4t\varepsilon,\varepsilon} : t \in \mathbb{N}\}$, i.e., of width 2ε centered at $\{0, 4\varepsilon, \ldots, 4(t-1)\varepsilon, 4t\varepsilon, \ldots\}$.

To build intuition, let us first see why the mode itself is not identifiable. In fact, we will establish the stronger statement that the mode is not identifiable with respect to $\mathcal{P}_{\psi,\varepsilon} \subset \mathcal{P}$. (See also (Fissler and Ziegel 2017, Lemma 2.4).)

Lemma 1 *The mode, $\Gamma_{mode} : \mathcal{P} \to \mathcal{R}$, is not identifiable with respect to $\mathcal{P}$, the class of unimodal probability measures defined on the real line which admit a smooth and bounded density.*

Proof For a contradiction, suppose there exists $V : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ such that Γ_{mode} is identified by V. For $h = 2/3$ define the density $p_h = h\psi_{0,1} + (1-h)\psi_{4,1}$ in $\mathcal{P}_{\psi,1}$. Clearly, $\Gamma_{\text{mode}}(p_h) = \Gamma_{\text{mode}}(\psi_{0,1}) = 0$, and since V identifies Γ_{mode}, we thus have $E_{p_h} V(0, Y) = 0$ and $E_{\psi_{0,1}} V(0, Y) = 0$. Combining,

$$0 = E_{p_h} V(0,Y) = hE_{\psi_{0,1}} V(0,Y) + (1-h)E_{\psi_{4,1}} V(0,Y) = (1-h)E_{\psi_{4,1}} V(0,Y)\,,$$

from which we conclude $E_{\psi_{4,1}} V(0, Y) = 0$ and thus $\Gamma_{\text{mode}}(\psi_{4,1}) = 0$, a contradiction. □

We now see that the mode is not identifiable, but it remains to understand its identifiable elicitation complexity. Theorem 1 generalizes the argument of Lemma 1, showing that the mode is not *indirectly* identifiable with respect to $\mathcal{P}$, the class of unimodal probability measures defined on the real line which admit a smooth and bounded density. In other words, for this class $\mathcal{P}$, there is no way to express the mode as a function of a finite-dimensional identifiable property. We conclude that the identifiable elicitation complexity of the mode is infinite with respect to $\mathcal{P}$.

Theorem 1 *The mode, Γ_{mode}, has infinite identifiable elicitation complexity with respect to $\mathcal{P}$, the class of unimodal probability measures defined on the real line which admit a smooth and bounded density.*

We briefly outline the proof of Theorem 1; the full proof appears in Appendix A. Let V be an identification function, which identifies some intermediate property $\hat{\Gamma} : \mathcal{P} \to \mathbb{R}^k$ for some finite dimension k. Taking $t > k$, we construct a probability density $p \in \mathcal{P}_{\psi,\varepsilon}$ with bump heights specified by a vector $h \in \mathbb{R}_+^{t+1}$, chosen so that the gap between any two bump heights is smaller than the minimum height. We observe that the expected value of V is linear in the bump heights, and moreover this linear transformation is rank deficient, giving us a nontrivial vector $h' \in \mathbb{R}^{t+1}$ in its kernel. By our initial choice of h, for any such h' we can find a suitable choice of coefficient $\alpha \in \mathbb{R}$ so that $h + \alpha h' \in \mathbb{R}_+^{t+1}$ while changing the mode. After normalization, this gives us a valid density $p' \in \mathcal{P}_{\psi,\varepsilon}$ yielding zero expectation of V, and thus residing in the same level set of $\hat{\Gamma}$ as p, yet with a different mode. This contradicts the existence of a function f satisfying $\Gamma_{\text{mode}} = f \circ \hat{\Gamma}$, as f would need to map the same $\hat{\Gamma}$ value to two different Γ_{mode} values. Figure 1 illustrates this construction, showing the density p along with a hypothetical choice of p'.

The impossibility result of Theorem 1 is strengthened in Theorem 2, which shows that the mode has infinite identifiable elicitation complexity even after

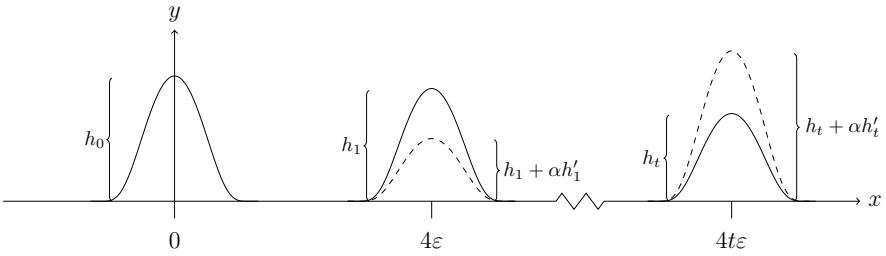

Fig. 1 The initial density p of Theorem 1 depicted with a solid line, and alternate density p' (before normalization) with a dashed. Here $t > k$, where k is the dimension of the intermediate property $\hat{\Gamma}$.

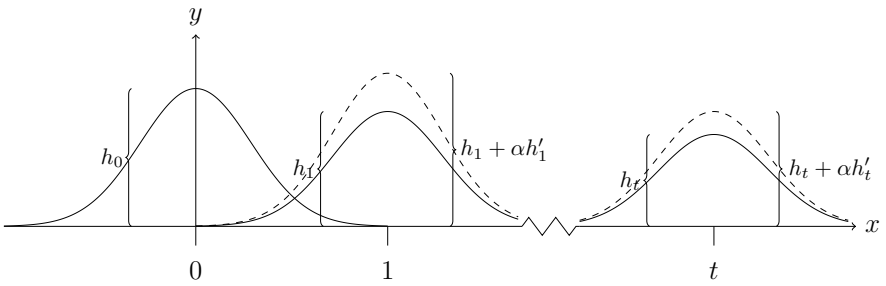

Fig. 2 The initial density q in Theorem 2 depicted as a mixture of Gaussians with solid lines, and the alternate density q' with dashed, before normalization.

restricting to the family $\mathcal{Q}$ of probability measures in $\mathcal{P}$ which can be expressed as a mixture of Gaussians. While the general outline of the proof is similar, the bump functions used in Theorem 1 were supported on disjoint intervals, which is clearly not true of Gaussians. In particular, changing the heights of distant Gaussians will now alter the mode.

Theorem 2 *The mode, $\Gamma_{mode} : \mathcal{Q} \to \mathcal{R}$, has infinite identifiable elicitation complexity with respect to $\mathcal{Q}$, the class of probability measures in $\mathcal{P}$ which can be expressed as a mixture of Gaussians.*

See Appendix A for the proof, which shows that the statement holds even when the mixture is over Gaussians with the same variance. As in Theorem 1, we assume $\Gamma_{\text{mode}} = f \circ \hat{\Gamma}$ for some finite-dimensional identifiable $\hat{\Gamma}$, construct an initial density $q \in \mathcal{Q}$, and show that there must exist another $q' \in \mathcal{Q}$ in the same level set as q but with a different mode (see Figure 2). While the broad outline remains the same, several technical challenges arise from the overlapping supports of Gaussians. To address these issues, we bound the potential contribution of one Gaussian to the density value at another to show that the mode changes from q to q', and use these bounds again to set the height vector h so that a coefficient α still exists for all possible vectors h'.

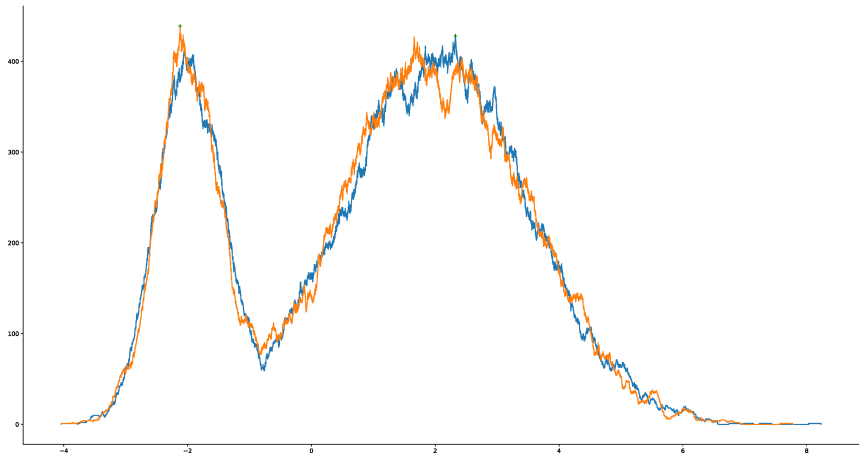

Fig. 3 The number $c(x)$ of the $n = 10,000$ data points within $\varepsilon = 0.1$ of a given point x, plotted here for two typical samples drawn independently from density p_{mix}. The empirical loss $\hat{L}(r) = \frac{1}{n}\sum_{i=1}^{n} L_\varepsilon(r, y_i)$ can be written $\hat{L}(r) = 1 - \frac{1}{n}c(r)$, and thus the maximum of each $c(\cdot)$, marked '+', corresponds to the estimated modal midpoint. In orange, the estimated mode and modal interval are both reasonably accurate, and in blue, neither are accurate.

4 Implications for the Modal Interval

While the mode is not elicitable, it is well-known that the midpoint of the modal interval Γ_ε defined in eq. (1), which we will refer to as the *modal midpoint*, is elicitable, via the simple loss function $L_\varepsilon(r, y) = \mathbb{1}\{|r - y| > \varepsilon\}$. (Note that as we restrict to single-valued functionals in this paper, in the technical results that follow, we will only consider distributions with a unique solution to eq. (1).) Recalling that the mode is the limit of the modal midpoint Γ_ε as the radius ε approaches 0, it is often suggested to estimate the mode by Γ_ε for a sufficiently small ε. Heinrich (2014) argues that this practice is ill-advised, given the non-elicitability of the mode, and further demonstrates this argument empirically. For a particular Gaussian mixture with density p_{mix}, with two local maxima m_0 and m_1, $p_{\text{mix}}(m_0) < p_{\text{mix}}(m_1)$, Heinrich shows that given a fixed number of samples, even small values of ε result in a modal midpoint $\hat{x}_\varepsilon$ which is more often closer to m_0 than the mode m_1. More precisely, for $\varepsilon \in \{0.5, 0.25, 0.1, 0.05, 0.025, 0.001\}$, out of 1000 trials each, Heinrich finds that $|\hat{x}_\varepsilon - m_1| < |\hat{x}_\varepsilon - m_0|$ in no more than 438 trials. Moreover, this success rate drops when $\varepsilon < 0.1$.

Yet we observe that, as the midpoint $x_\varepsilon = \Gamma_\varepsilon(p_{\text{mix}})$ of the true modal interval is very close to m_1 for sufficiently small ε, this simulation study also shows that the sample modal midpoint $\hat{x}_\varepsilon$ is an ineffective estimate of the true modal midpoint x_ε. When m_1 is replaced by x_ε in the preceding paragraph, we obtain qualitatively similar results: the majority of the time, $\hat{x}_\varepsilon$ is closer to m_0

than the true modal midpoint x_ε, and the situation worsens for $\varepsilon < 0.1$. (See Figure 3 and Appendix B for details.) In summary, not only does the modal midpoint fail to estimate the mode, it fails to estimate the modal midpoint.

These empirical findings suggest the difficulty of eliciting modal midpoints in practice, despite the fact that they are elicitable. This sentiment is confirmed by the following Corollary, which extends our argument on the elicitation complexity of the mode to modal midpoints. The result essentially follows from the following observation. For a distribution consisting of disjoint bump functions in $\mathcal{P}_{\psi,\varepsilon}$, as defined after eq. (2) and used in the argument of Theorem 1, the mode and modal midpoint Γ_ε coincide. While this equivalence does not hold anymore for mixtures of Gaussians, we remark that the proof of Theorem 2 could be directly modified, by enlarging the width of the balls to $B_{2\sigma}(x_i)$, and choosing sufficiently small γ and large C, so that the same logic would hold for the modal interval when ε is sufficiently small.

Corollary 1 *For any $\varepsilon > 0$, the modal midpoint, $\Gamma_\varepsilon : \mathcal{P}_\varepsilon \to \mathcal{R}$, has infinite identifiable elicitation complexity with respect to $\mathcal{P}_\varepsilon$, the class of probability measures defined on the real line which admit a smooth and bounded density, and have a unique mode and ε-modal midpoint.*

Proof Let $\varepsilon > 0$ be given. Observe that for any $p \in \mathcal{P}_{\psi,\varepsilon}$, the disjoint bump functions comprising p are spaced far enough apart so that an interval of width 2ε can only intersect the support of at most one bump function. Moreover, if the interval intersects the ith such function, it maximizes the contained mass by centering the interval at exactly $4\varepsilon i$. The global maximum mass is therefore achieved by centering the interval to capture the mass of the bump function with the largest weight, whose midpoint coincides with the mode. From these observations, we conclude $\Gamma_{\text{mode}}(p) = \Gamma_\varepsilon(p)$ for all $p \in \mathcal{P}_{\psi,\varepsilon}$. In other words, Γ_ε and Γ_{mode} are the same functional with respect to $\mathcal{P}_{\psi,\varepsilon} \subseteq \mathcal{P}_\varepsilon$. Hence, the identification complexity and identifiable elicitation complexity of the modal midpoint Γ_ε with respect to $\mathcal{P}_\varepsilon$ are at least that of the mode. □

The fact that modal midpoints are elicitable yet have infinite identifiable elicitation complexity illustrates the subtlety of our definitions. This subtlety is important; as pointed out by Frongillo and Kash (2015), one can construct pathological yet elicitable properties, such as a bijective $\Gamma : \Delta(\mathcal{Y}) \to [0,1]$ for finite $\mathcal{Y}$ via any strictly proper scoring rule (Gneiting and Raftery 2007). Hence, the restriction to identifiable intermediate properties, or some other class of properties ruling out such pathologies (see Section 5), is necessary for practical estimation schemes. In this light, our results are in line with the observation that both the mode and modal midpoint fail to be continuous in even weak senses: for certain distributions p_1, p_2, $\Gamma_{\text{mode}}(\lambda p_1 + (1-\lambda)p_2)$ is not continuous in λ, and the same is true of Γ_ε.

To close, it is interesting to contrast the above demonstration and negative result with the existing positive results in the literature on the estimation of the mode and modal midpoints. Some positive results, showing favorable error bounds, assume that the true density is not only unimodal but has a

unique local maximum, i.e., the density increases before the mode and decreases afterwards; see for example Robertson and Cryer (1974, Section 2) and Lee (1989, Assumption 2). Moreover, many proposed estimators are expressed as sequences of estimators which depend on the sample size (Parzen 1962; Chernoff 1964; Grenander 1965; Venter 1967); we may roughly view these estimators as intermediate properties of countably infinite dimension, consistent with our results.

5 Discussion

Several interesting open questions remain. One could further ask for the identifiable elicitation complexity of the mode with respect to other classes of probability distributions. One interesting class would be distributions with densities having a unique local maximum, though note that the elicitability of the mode is still open in this case. The method of perturbing the heights of (in this case, heavily overlapping) bumps as in Lemma 1 and Theorems 1 and 2 does not seem sufficient for this class.

Another set of questions arises when stepping away from the class of identifiable properties and considering other classes, such as weakly identifiable properties; negative results with respect to this class would show infinite complexity with respect to continuous, non-locally-constant, component-wise elicitable properties (Lambert 2018; Steinwart et al 2014). Another interesting class of properties in this context would be those elicited by convex loss functions, as these properties are of practical interest yet need not be identifiable (Frongillo and Kash 2015). Finally, we suspect that our techniques could be applied to other properties whose elicitation complexity is not known, such as the width of the smallest confidence interval.

A Omitted Proofs

Proof (of Theorem 1) Let $\varepsilon > 0$ be given. Since the identification complexity lower bounds the identifiable elicitation complexity of the mode it suffices to show that the mode is not k-identifiable for arbitrary $k \in \mathbb{N}$. Suppose, by way of contradiction, that the mode is k-identifiable. Hence, there exists a property $\hat{\Gamma} : \mathcal{P} \to \hat{\mathcal{R}} \subseteq \mathbb{R}^k$ identified by $V : \hat{\mathcal{R}} \times \mathbb{R} \to \mathbb{R}^k$ and function $f : \hat{\mathcal{R}} \to \mathcal{R}$ such that $\Gamma_{\text{mode}} = f \circ \hat{\Gamma}$. Our goal will be to specify two densities $p, p' \in \mathcal{P}_{\psi,\varepsilon} \subseteq \mathcal{P}$ with $\hat{\Gamma}(p) = \hat{\Gamma}(p')$ and $\Gamma_{\text{mode}}(p) \neq \Gamma_{\text{mode}}(p')$, contradicting the existence of f.

Let $t > k$ and consider the following density $p = \sum_{i=0}^{t} h_i \psi_{4i\varepsilon,\varepsilon}$ in $\mathcal{P}_{\psi,\varepsilon}$ with strictly decreasing heights $h_0 > h_1 > \ldots h_t > h_0/2 > 0$ and $\sum_{i=0}^{t} h_i = 1$. Observe that $\Gamma_{\text{mode}}(p) = 0$ and denote $\hat{\Gamma}(p) = r$. Consider the $k \times t$ matrix

$$M = \left[E_{\psi_{4\varepsilon,\varepsilon}}(V(r, Y)), \ldots, E_{\psi_{4t\varepsilon,\varepsilon}}(V(r, Y)) \right]. \tag{3}$$

Let $h' = (h'_1, \ldots, h'_t)$ denote a nontrivial vector in the kernel of M. To complete the proof, we will demonstrate that for any h' there exist real numbers $\alpha, \beta \in \mathbb{R}$ so that $p' = \beta \left(p + \alpha \left(\sum_{i=1}^{t} h'_i \psi_{4i\varepsilon,\varepsilon} \right) \right)$ is a density satisfying $\hat{\Gamma}(p') = r$ and $\Gamma(p') \neq 0$. We proceed by considering all cases of h' and showing the existence of α in each case.

First, considering $h'_1, \dots, h'_t \geq 0$, let $h'_{i(\max)}$ denote the entry of h' with greatest magnitude (if not unique, choose the entry associated with the maximal initial height $h_{i(\max)}$), and take $\alpha > (h_0 - h_{i(\max)})/h'_{i(\max)}$. Second, if $h'_1, \dots, h'_t \leq 0$, then take $-h'$ and treat as above. In the final case, at least one pair of entries of h' have opposite sign. Let $h'_{i(\max)}$ denote an entry of h' with the greatest magnitude (if not unique, choose the entry associated with the maximal initial height $h_{i(\max)}$) and assume $h'_{i(\max)} > 0$; otherwise take $-h'$. Choose α such that $(h_0 - h_{i(\max)})/h'_{i(\max)} < \alpha \leq \min_{\{i:h'_i<0\}} h_i/|h'_i|$ satisfying $\alpha \neq |h_{i(\max)} - h_i|/(h'_{i(\max)} - h'_i)$ for any i with $h'_{i(\max)} > h'_i > 0$. Note this interval is nonempty because $h_0 - h_{i(\max)} < \frac{h_0}{2} < h_i$ and $|h'_i| < h'_{i(\max)}$ for all i such that $h'_i < 0$. In each of the above cases, there are finitely many α which do not yield a unimodal p'. If the α chosen yields a p' which is not unimodal, then discard this particular α from the interval and choose again.

With the appropriate normalization constant β, we now have a density given by $p' = \beta\left(p + \alpha\left(\sum_{i=1}^t h'_i \psi_{4i\varepsilon,\varepsilon}\right)\right)$. As h' is contained in the kernel of M, linearity of expectation and the definition of V now guarantee that $\hat{\Gamma}(p') = \hat{\Gamma}(p) = r$, and the method with which we showed α exists ensures that p' is unimodal with $\Gamma_{\text{mode}}(p') \neq \Gamma_{\text{mode}}(p) = 0$. These two statements together contradict the existence of the link function f satisfying $\Gamma_{\text{mode}} = f \circ \hat{\Gamma}$. □

Proof (of Theorem 2) As in the proof of Theorem 1, we assume the mode is k-identifiable and arrive at a contradiction. Hence, we assume there exists a property $\hat{\Gamma} : \mathcal{Q} \to \hat{\mathcal{R}} \subseteq \mathbb{R}^k$ identified by $V : \hat{\mathcal{R}} \times \mathbb{R} \to \mathbb{R}^k$ and function $f : \hat{\mathcal{R}} \to \mathcal{R}$ such that $\Gamma_{\text{mode}} = f \circ \hat{\Gamma}$. We will again specify two densities from $\mathcal{Q}$ in the same level set of $\hat{\Gamma}$, but different modes which contradicts the existence of f.

Let $t > k$, and let $q_0, q_1, \dots, q_t$ be Gaussian densities with unit height ($\sigma^2 = \frac{1}{2\pi}$) centered at $x_i = Ci$ for some C to be determined. For any mixture parameters $h = (h_0, h_1, \dots, h_t) \in \mathbb{R}_+^{t+1}$, we will denote the Gaussian mixture density as follows,

$$q[h](x) = \sum_{i=0}^{t} h_i q_i(x) \in \mathcal{Q}' \ ,$$

where we define $\mathcal{Q}'$ to be all positive scalings of densities in $\mathcal{Q}$. As we are interested in the mode, we can always renormalize to obtain a distribution in $\mathcal{Q}$ with the same mode. In the following, we extend $\Gamma_{\text{mode}}(p)$ for unnormalized densities in the natural way.

Observe that for any mixture h, we have $\Gamma_{\text{mode}}(q[h]) \in \cup_{i=0}^t B_\sigma(x_i)$, for any $C > 0$. This follows from second-order optimality conditions: as the inflection point of a Gaussian density $N(\mu, \sigma)$ is at $\mu \pm \sigma$, we have $\frac{d^2}{dx^2} q_i(x) < 0 \iff |x - x_i| < \sigma$, and thus $\frac{d^2}{dx^2} q[h](x) < 0 \implies |x - x_i| < \sigma$ for some i. Let $\gamma := q_1(\sigma) = e^{-\pi(\sigma - C)^2}$. We will want $\gamma < \frac{1}{4(t+1)}$, and thus we choose any $C > \sigma + \sqrt{\frac{\log(4(t+1))}{\pi}}$.

We will additionally use the following claims in our proof.

Claim 1 *For all h, i, $h_i \leq q[h](x_i) \leq \max_{x \in B_\sigma(x_i)} q[h](x) \leq h_i + \gamma \sum_{j \neq i} h_j$.*

Claim 2 *If $h_i > \max_{j \neq i} h_j + \gamma \sum_k h_k$, then $\Gamma_{mode}(q[h]) \in B_\sigma(x_i)$.*

Claim 3 *If $h_i < h_j - \gamma \sum_{k \neq i} h_k$, then $\Gamma_{mode}(q[h]) \notin B_\sigma(x_i)$.*

In Claim 1, the first two inequalities are trivial, and the third follows from the observation that the contribution of q_j to $q[h](x)$ is upper bounded by $h_j \gamma$ for all $x \in B_\sigma(x_i)$. Claim 2 then follows from Claim 1: for all j we have $q[h](x_i) \geq h_i > h_j + \gamma \sum_k h_k \geq h_j + \gamma \sum_{k \neq j} h_k \geq \max_{x \in B_\sigma(x_j)} q[h](x)$. Similarly, for Claim 3, $\max_{x \in B_\sigma(x_i)} q[h](x) \leq h_i + \gamma \sum_{k \neq i} h_k < h_j \leq q[h](x_j)$.

Finally, we construct our initial mixture h so that $\sum_i h_i = 1$ and the following condition holds,

$$h_0 - \gamma > h_1 > h_2 > \cdots > h_t > \frac{3}{4}h_0 \ . \tag{4}$$

By Claim 2, we therefore would have $\Gamma_{\text{mode}}(q[h]) \in B_\sigma(x_0)$. Condition (4) can be satisfied for $t > 5$ (and smaller if C is larger); we give one explicit construction here. Letting $c = 1/(t+1)$ for ease of notation, we may take $h_0 = (5/4)c$ and $h_1 = c$. Enforcing $\sum_i h_i = 1$, the average of the remaining elements is then $c - (1/4)c/(t-1) = (1 - 1/4(t-1))c$ which is strictly less than h_1 but strictly greater than $c(1 - 1/16) = (3/4)h_0$, as desired. We may therefore choose the remaining elements to be any decreasing sequence in the interval $(3h_0/4, h_1)$ whose average is $c(1 - 1/4(t-1)) \in (3h_0/4, h_1)$.

Now let $\hat{\Gamma}(q[h]) = r$. Consider the $k \times t$ matrix

$$M = \left[E_{q_1}(V(r, Y)), \ldots, E_{q_t}(V(r, Y))\right] . \tag{5}$$

Let $h' = (h'_1, \ldots, h'_t)$ denote a nontrivial vector in the kernel of M. To complete the proof, we will demonstrate that for any such h' there exists a real number $\alpha \in \mathbb{R}$ so that $q[h + \alpha h'] = q[h] + \alpha \sum_{i=1}^t h'_i q_i$ (after normalization to obtain the corresponding element in $\mathcal{Q}$) is the desired density. We proceed by cases on the entries of h'.

First, if $h'_1, \ldots, h'_t \geq 0$, then let $h'_{i(\max)}$ denote the entry of h' with greatest magnitude. If $h'_{i(\max)}$ is not unique, then choose the entry associated with the maximal initial height $h_{i(\max)}$. Choose α such that

$$\frac{h_0 - h_{i(\max)} + \gamma}{h'_{i(\max)} - \gamma\left(\sum_{k \neq i(\max)} h'_k\right)} < \alpha.$$

This ensures that $h_0 < (h_{i(\max)} + \alpha h'_{i(\max)}) - \gamma\left(1 + \alpha \sum_{k \neq i(\max)} h'_k\right)$ so that $\Gamma_{\text{mode}}(q[h + \alpha h']) \notin B_\sigma(x_0)$ by Claim 3. Second, if $h'_1, \ldots, h'_t \leq 0$, then take $-h'$ and treat as above.

In the final case, at least one pair of entries of h' have opposite sign. Let $h'_{i(\max)}$ denote the entry of h' with the greatest magnitude and assume $h'_{i(\max)} > 0$; otherwise take $-h'$. If $h'_{i(\max)}$ is not unique, then choose the entry associated with the maximal initial height $h_{i(\max)}$. Choose α such that

$$\frac{h_0 - h_{i(\max)} + \gamma}{h'_{i(\max)} - \gamma\left(\sum_{k \neq i(\max)} h'_k\right)} < \alpha \leq \min_{i : h'_i < 0} \frac{h_i}{|h'_i|}.$$

Once again, the lower bound ensures that $h_0 < (h_{i(\max)} + \alpha h'_{i(\max)}) - \gamma\left(1 + \alpha \sum_{k \neq i(\max)} h'_k\right)$ so that $\Gamma_{\text{mode}}(q[h + \alpha h']) \notin B_\sigma(x_0)$ by Claim 3. We bound α from above in this case to ensure that $q[h + \alpha h'] \geq 0$, meaning we have a valid density.

It thus remains to verify that this interval is nonempty. Take an index i such that $h'_i < 0$. Note that $h'_{i(\max)} \geq \frac{\sum_{k \neq i(\max)} h'_k}{t} > \gamma \sum_{k \neq i(\max)} h'_k$, so that $h'_{i(\max)} - \gamma\left(\sum_{k \neq i(\max)} h'_k\right) > h'_{i(\max)}(1 - \gamma t) > \frac{3h'_{i(\max)}}{4} \geq \frac{3|h'_i|}{4}$. Also note that $\frac{h_0}{4} + \gamma < \frac{h_0}{4} + \frac{1}{4(t+1)} < \frac{h_0}{2} < \frac{3h_0}{4} \cdot \frac{3}{4}$.

Chaining these inequalities together,

$$\begin{aligned}\frac{h_0 - h_{i(\max)} + \gamma}{h'_{i(\max)} - \gamma\left(\sum_{k \neq i(\max)} h'_k\right)} &< \frac{h_0 - h_{i(\max)} + \gamma}{h'_{i(\max)}(1-\gamma t)} \\ &< \frac{\frac{h_0}{4} + \gamma}{h'_{i(\max)}(1-\gamma t)} \\ &< \frac{\frac{3h_0}{4} \cdot \frac{3}{4}}{h'_{i(\max)}(1-\gamma t)} \\ &\leq \frac{\frac{3h_0}{4} \cdot \frac{3}{4}}{\frac{3|h'_i|}{4}} = \frac{\frac{3h_0}{4}}{|h'_i|} < \frac{h_i}{|h'_i|} \, .\end{aligned}$$

As this inequality holds for all such i, it holds for the minimum over i.

In each of the above cases, there are finitely many α which fail to yield a unimodal density, $q[h + \alpha h']$. If the α chosen yields such a $q[h + \alpha h']$, discard this particular α and choose again.

Similar to the conclusion of Theorem 1, the density $q[h + \alpha h']$ (after normalization to obtain the corresponding element in $\mathcal{Q}$) gives the desired contradiction. □

B Experimental Details

So as to allow comparison with Heinrich (2014), we consider a density p_{mix} which is a mixture of two Gaussians; letting $p_1 = N(2, 1.5)$ and $p_2 = N(-2, 0.5)$, where $N(\mu, \sigma)$ denotes a Gaussian density with mean μ and standard deviation σ, we set $p_{\text{mix}} = 0.75p_1 + 0.25p_2$. The true mode of p_{mix} is $m_0 = \Gamma_{\text{mode}}(p_{\text{mix}}) \approx -1.987047$, with the other local maximum occuring at $m_1 \approx 2.000000$. The experiment performed is analogous to Heinrich (2014): for each value of ε as shown in Table 1, and in each of 1000 trials, we collect $n = 10,000$ independent samples from p_{mix}, and measure the performance of the empirical modal midpiont $\hat{x}_\varepsilon$ relative to the true mode m_0 and true modal midpoint $x_\varepsilon = \Gamma_\varepsilon(p_{\text{mix}})$. In the case of a tie for $\hat{x}_\varepsilon$, we take the lowest value (which the reader will note should favor the correct value). In sum, our results are qualitatively similar to Heinrich (2014), in that the modal midpoint $\hat{x}_\varepsilon$ fails to estimate the mode, but we can also confirm that it fails to estimate the modal midpoint x_ε as well. Note in particular that the two "Versus local max" columns are identical.

Table 1 The ineffectiveness of the modal midpoint as an estimate of the mode, or even of the modal midpoint itself. The table headings denote the following. x_ε: the true modal midpoint; MSE: mean squared error with respect to the true mode and true modal midpoint x_ε; Versus local max: the number of trials (out of 1000) where the estimate $\hat{x}_\varepsilon$ was closer to the true mode m_0, or true modal midpoint x_ε, than the other local maximum m_1; Minimal loss: the best empirical average loss observed in the 1000 trials.

		MSE		Versus local max		
ε	x_ε	Mode	Modal	Mode	Modal	Minimal loss
0.5	−1.976 691	15.88	15.80	0	0	0.791
0.25	−1.984 999	11.06	11.05	302	302	0.890
0.1	−1.986 739	8.75	8.75	447	447	0.952
0.05	−1.986 970	9.00	9.00	433	433	0.972
0.025	−1.987 028	9.12	9.12	424	424	0.985
0.001	−1.987 047	8.84	8.84	431	431	0.998

Acknowledgements We thank Tobias Fissler and Jessie Finocchiaro for helpful suggestions, Jonas Brehmer for simplifying the proof of Lemma 1, and Nicole Woytarowicz for her initial work on this project, including a proof of Lemma 1 in her B.S. thesis.

References

Chernoff H (1964) Estimation of the mode. Annals of the Institute of Statistical Mathematics 16(1):31–41

Fissler T, Ziegel JF (2017) Order-Sensitivity and Equivariance of Scoring Functions. arXiv:171109628 [math, stat] URL `http://arxiv.org/abs/1711.09628`, arXiv: 1711.09628

Fissler T, Ziegel JF, et al (2016) Higher order elicitability and osbands principle. The Annals of Statistics 44(4):1680–1707

Frongillo R, Kash I (2015) On elicitation complexity. In: Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R (eds) Advances in Neural Information Processing Systems 28, Curran Associates, Inc., pp 3258–3266

Gneiting T (2011) Making and evaluating point forecasts. Journal of the American Statistical Association 106(494):746–762

Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association 102(477):359–378

Grenander U (1965) Some direct estimates of the mode. The Annals of Mathematical Statistics 36(1):131–138

Heinrich C (2014) The mode functional is not elicitable. Biometrika 101(1):245–251

Lambert NS (2018) Elicitation and evaluation of statistical forecasts. Preprint

Lambert NS, Pennock DM, Shoham Y (2008) Eliciting properties of probability distributions. In: Proceedings of the 9th ACM Conference on Electronic Commerce, ACM, pp 129–138

Lee Mj (1989) Mode regression. Journal of Econometrics 42(3):337 – 349

Osband K (1985) Providing incentives for better cost forecasting. PhD thesis, University of California, Berkeley

Parzen E (1962) On estimation of a probability density function and mode. The Annals of Mathematical Statistics 33(3):1065–1076

Robertson T, Cryer JD (1974) An iterative procedure for estimating the mode. Journal of the American Statistical Association 69(348):1012–1016

Steinwart I, Pasin C, Williamson R, Zhang S (2014) Elicitation and identification of properties. In: Balcan MF, Feldman V, Szepesvri C (eds) Proceedings of The 27th Conference on Learning Theory, PMLR, Barcelona, Spain, Proceedings of Machine Learning Research, vol 35, pp 482–526

Venter J (1967) On estimation of the mode. The Annals of Mathematical Statistics 38(5):1446–1455