European Journal of Operational Research 323 (2025) 1013–1023

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Innovative Applications of O.R.

Expected information of noisy attribute forecasts for probabilistic forecasts

Omid M. Ardakani [a], Robert F. Bordley [b], Ehsan S. Soofi [c,*]

[a] Parker College of Business, Georgia Southern University, Savannah, GA 31419, USA
[b] College of Engineering, University of Michigan, Ann Arbor, 1075 Beal Avenue, MI 48109, USA
[c] Sheldon B. Lubar College of Business, University of Wisconsin-Milwaukee, WI, 53201, USA

ARTICLE INFO

Keywords:
Decision analysis
Expected information
Maximum entropy
Probability forecast
Sensitivity analysis

ABSTRACT

This paper extends the maximum entropy (ME) model to include uncertainty about noisy moment forecasts. In this framework the noise propagates to the ME model through the constrained optimization's Lagrange multipliers. The mutual information and expected Fisher information are included for assessing effects of the noisy moment forecasts on the ME model and its parameters. A new mean–variance decomposition of the mutual information is derived for the normal distribution when the mean and variance are both noisy. A simulation estimator is used to estimate the expected information for noisy ME models on finite support. A family of ensemble of individual level noisy ME forecast models is introduced which includes individual level versions of the conditional logit and multiplicative competitive interaction models as specific cases. To illustrate the implementation and merits of the proposed noisy ME framework, the classic loaded dice problem and discrete choice analysis are examined.

1. Introduction

Probabilistic forecasts have demonstrated their utility in numerous fields, including energy (Hong et al., 2016), electricity load (Ludwig et al., 2023), population studies (Wilson & Bell, 2007), macroeconomic variables (Lahiri et al., 1988; Lahiri & Wang, 2020; Rich & Tracy, 2010; Shoja & Soofi, 2017; Wallis, 2005), among others, and demand forecasting for products (Snyder & Shen, 2011). These forecasts are commonly referred to as probability forecasts for discrete outcomes and density forecasts for continuous outcomes.

Information Theory (IT) provides a methodological approach for constructing models based on partial information about the probability distribution of a stochastic object Q which may be a scalar random variable, a random vector, a categorical variable, or an unobservable parameter. Consider Q that varies according to a probability density or mass function (PDF) π on a support $\mathbb{Q}$. The Shannon entropy of Q (or π) can be represented as

$$\mathcal{H}(Q) = \mathcal{H}(\pi) = -\int_{\mathbb{Q}} \pi(q) \log \pi(q) d\nu(q), \tag{1}$$

where, for the continuous case, π is a PDF and $d\nu(q) = dq$. For the discrete case, the integral becomes a summation and $d\nu(q) = 1$.

The Maximum Entropy (ME) principle (Jaynes, 1957) postulates a probability model that maximizes (1) while maintaining consistency with given partial information. In the absence of any information other than the size of $\mathbb{Q}$, the ME model implies uniform probabilities, in agreement with the Laplace's principle of insufficient reason (Jaynes, 1957). In the presence of partial information, the ME model minimizes the adjustments of the uniform distribution to be consistent with the given information. The ME models optimize (1) while satisfying constraints on various attribute moments, called *moment parameters*, given either as forecasts or expert assessments. Since an expert's skilled intuition relies on the expert's memory (Simon, 1992, p. 150), memory limitations will create error in expert assessments. As a result, forecast errors and memory errors create uncertainty about the moments. It is therefore prudent to consider the given moments as initial guesses about uncertain quantities.

The main contribution of this paper is extension of the ME approach to include uncertainty on the moment parameters. This ME formulation reflects the reality of many applied data analysis problems where moment forecasts are noisy. The noisy ME model is conditional on the given moments which induce noise into the model parameters (Lagrange multipliers). The expectation of the noisy ME distributions with the distribution of the noisy parameters gives a mixture model for the marginal distribution. Our noisy ME moment framework corresponds to the Bayesian model, where the noisy ME is the likelihood model, the distribution of the noisy moment parameters is the prior distribution, and the ME mixture is the prior predictive distribution. Previously, Golan et al. (1996) introduced IT models that incorporate noise for multinomial data. Their model assumes the multinomial indicator variable $y \in \{0, 1\}$ is a linear function of a set of predictors with random

[*] Corresponding author.
E-mail addresses: oardakani@georgiasouthern.edu (O.M. Ardakani), rbordley@umich.edu (R.F. Bordley), esoofi@uwm.edu (E.S. Soofi).

https://doi.org/10.1016/j.ejor.2024.12.024
Received 12 June 2023; Accepted 12 December 2024
Available online 20 December 2024

vector coefficients plus a noise term, both distributed independently on finite sets of points in $[-1, 1]$.

The proposed noisy moments ME framework facilitates assessing the impacts of variations of the moment forecasts on the ME forecast model. Our focal tool for the assessment is the information measure known as the mutual information between two variables or vectors. This measure provides the *expected information* in terms of the entropy reduction and Kullback–Leibler divergence. The expected Fisher information is included for assessing the sensitivity of the ME model parameters to variations of the noisy moments. These measures enable implementing (Borgonovo et al., 2021) recommendation that a forecasting report be accompanied by a sensitivity analysis describing the level of uncertainty in the analysis.

The proposed noisy moment ME framework leads to three contributions to the IT literature. Firstly, the normal distribution is the principal model for statistics, econometrics, and many other fields. It is known that with given mean and variance parameters the ME models on the real line $\Re$ support is the normal distribution. The mutual information for the mean of the normal distribution with the normal prior, given the variance, was developed by Lindley (1956) and is used, for example, by Stone (1959), Tinsley et al. (1980), Theil and Chung (1988), Granger and Lin (1994), Pourahmadi and Soofi (2000), Retzer et al. (2009), and Ebrahimi et al. (2010a), among others. However, the mutual information for the more general and practical case where both mean and variance are unknown has not appeared in the literature. We fill this gap and derive the mutual information for the more realistic case when both the mean and variance are noisy with the normal-gamma conjugate prior distribution widely used in the Bayesian modeling. This result provides a new mean–variance decomposition of information.

Secondly, the ME probability models on finite supports are used in many fields for various applications. These models are in logit form. The mutual information for assessing probabilistic sensitivity of the noisy ME logit models requires the marginal distribution and the conditional distribution given the moment parameters. The integral for computing continuous mixtures of the noisy ME logit models is intractable. Harding and Hausman (2007) and Bordley (2011) developed approximations for computing continuous mixtures of the random coefficient logit model. Pakes and Pollard (1989) developed a theory for uniformly convergent simulation estimates without imposing the smoothness condition on the mixing distribution. We utilize this theory to introduce an empirical mutual information for the random sample of N moment parameters. This approach extends the existing entropy reduction information indices of the ME models to the noisy ME models (Soofi, 1992, 1994) with the decision-theoretic interpretations of the mutual information (Bernardo, 1979).

Thirdly, the existing ME models are derived based on moment constraints that are formulated by combining relevant information for random samples of N observations. These formulations have led to derivations of ME models that are in the form of the multinomial logit and Multiplicative Competitive Interaction (MCI) models with their parameters being the same for the entire sample of observations (Brockett et al., 1995; Soofi, 1992, 1994). These ME models, which we call the *pooled ME* models, are the IT counterparts of the traditional logit models in the literature. Thus far the IT counterpart of the traditional random parameters logit model used in various applications (Train, 2009) have not been developed. We fill this gap by considering the relevant information drawn from each individual observation as a random draw from the distributions of the noisy moments across the sample of N observations. This new formulation produces a family of ensemble of N models with varying parameters. The special cases include individual level versions of the conditional logit (McFadden, 1973) and MCI (Cooper, 1993; Huff, 1962) models. The differences between the ME approach and the traditional approaches, e.g., the maximum likelihood estimate (MLE) logit and the random parameter logit, are the derivations and data requirements. The ME requires the moment parameters for the constraints and traditional methods require actual $\{0, 1\}$ data on outcomes that already have occurred such as choices already made by N decision makers (DMs).

For the choice probability forecasting problems, future choices are yet to be made, so $\{0, 1\}$ data is not available. Traditional approaches use historical data to estimate parameters of a pre-specified choice probability model. It has been shown that the MLE of the logit model can be derived as a special case of the pooled ME logit (Soofi, 1992). Plugging the future attribute scores in the past model provides probability forecasts and the pooled ME/MLE attribute forecasts. We utilize these attribute forecasts to derive the ensemble of ME forecast models for N DMs. An illustrative application example examines the accuracy of the individual level ME forecast model with the pooled ME/MLE probability forecast model and three versions of the traditional random parameters logit model.

The paper is organized as follows. Section 2 defines the uncertainty and divergence measures, and outlines the information properties of ME models. Section 3 introduces the noisy moment ME modeling framework, the information measures for the normal ME model when the mean and variance are both noisy, and the empirical information measures for the ME models on finite supports. Section 4 presents the individual level ME logit model for the choices of N DMs and the differences between the new and the existing ME models. This section also presents the individual level ME forecast model consistent with the attribute forecasts of existing models. Section 5 gives some concluding remarks. An online Supplementary Document provides R codes for implementation of the new ME models.

2. Preliminaries

In many problems, a stochastic object Q is considered under two distributions $\pi_Q \in \Omega$ and $\pi_{Q|\theta} \in \Omega_\theta \subseteq \Omega$ on a support $\mathbb{Q}$, where θ represents a set of parameters. Q may be unobservable (e.g., a parameter of a probability distribution) or observable (e.g., a future outcome of a random variable). The parameter θ can be scalar, vector, or a function of a set of predictors X. It can be non-stochastic or stochastic drawn from the distribution g_θ of a stochastic object θ. In the stochastic predictor case, $\pi_{Q|\theta}$ is the conditional distribution. This section gives an overview of the measures of information for the non-stochastic case.

2.1. Uncertainty and divergence functions

The most unpredictable situation occurs when the absence of knowledge does not allow favoring an event over any other event in the support of π. By Laplace's Principle of Insufficient Reason all possible values of Q (intervals of equal width in the continuous case) are equally likely. This establishes the uniform probability distribution as the reference point for quantifying the uncertainty in terms of predictability.

DeGroot (1962) defined uncertainty by a nonnegative measurable function $\mathcal{U}(\pi) \geq 0$ on the space of distributions on finite support $\mathbb{Q} = \{q_1, \dots, q_J\}$. He stipulated that a "typical" uncertainty function will take $\mathcal{U}(\pi) = 0$ when $\pi(q_j) = 1$ for $j = j_0$ and $\pi(q_j) = 0$ for $j \neq j_0$, and $\mathcal{U}(\pi)$ attains its maximum at or near the uniform distribution. Ebrahimi et al. (2010b) identified two desirable properties for uncertainty functions, which we state more formally.

Definition 1. On a space of probability distributions Ω on a support $\mathbb{Q}$, the uncertainty function is defined by a measurable concave function $\mathcal{U} : \Omega \to \Re$ such that for all $\pi \in \Omega$, $\mathcal{U}(\pi) \leq \mathcal{U}(\pi_0^*)$ where π_0^* is the uniform distribution on $\mathbb{Q}$.

A distribution $\pi_{Q|\theta} \in \Omega$ is said to be more (or less) informative than $\pi_Q \in \Omega$ if

$$\Delta^{\mathcal{U}}(\theta) = \Delta^{\mathcal{U}}(\pi_{Q|\theta} : \pi_Q) = \mathcal{U}(\pi_Q) - \mathcal{U}(\pi_{Q|\theta}) \geq (\text{or} \leq) 0. \tag{2}$$

The sign of $\Delta^{\mathcal{U}}(\theta)$ depends on which of the two distributions is closer to the uniform distribution (more concentrated).

A concave $\mathcal{U}$ is the requirement for a reasonable uncertainty function to formalize the intuitive expectation that considering a variable for predicting another variable could increase, but would not decrease, the predictability. DeGroot (1962) showed that if θ is stochastic with a distribution g_θ and π_Q is the marginal distribution of Q, then

$$E_Q[\Delta^{\mathcal{U}}(\theta)] = \mathcal{U}(\pi_Q) - E_Q[\mathcal{U}(\pi_{Q|\theta})] \geq 0, \tag{3}$$

if and only if $\mathcal{U}$ is a concave.

The most well known and widely used uncertainty function in the IT literature is Shannon entropy (1); henceforth we will use $\mathcal{U} = H$ and $\Delta = \Delta^H$. The uncertainty relative to the uniform distribution can be quantified by the extent of discrepancy between probability distributions. A divergence function between a pair of probability distributions with PDFs (π_1, π_2) on a support $\mathbb{Q}$ is defined by $\mathcal{D}(\pi_1, \pi_2) \geq 0$ and attains the equality if and only if $\pi_1(q) = \pi_2(q)$ for all outcomes on $\mathbb{Q}$. See Borgonovo et al. (2021) for a similar definition of the separation function between probability distributions.

The prominent divergence function in IT is the Kullback–Leibler (KL) information divergence between two PDFs $(\pi_1, \pi_2) \in \Omega \times \Omega$, defined by

$$\mathcal{K}(\pi_1 : \pi_2) = \int_{\mathbb{Q}} \pi_1(q) \log \frac{\pi_1(q)}{\pi_2(q)} d\nu(q) \geq 0, \tag{4}$$

given that $\pi_1(q) = 0$ wherever the reference distribution $\pi_2(q) = 0$ (the absolute continuity condition). $\mathcal{K}(\pi_1 : \pi_2) = 0$ if and only if $\pi_1(q) = \pi_2(q)$ almost everywhere on $\mathbb{Q}$. $\mathcal{K}(\pi_1 : \pi_2)$ is invariant under one-to-one transformations of q; this property is inclusive of the desirable monotonic transformation invariance (Borgonovo et al., 2014).

The KL divergence provides the information for discriminating between two models for the distribution of Q in terms of the expected log-odds (Bayes Factor) in favor of π_1 against π_2 (Kullback, 1959, p. 38). In general, $\mathcal{K}(\pi_1 : \pi_2) \neq \Delta(\pi_1 : \pi_2)$. So, $\mathcal{K}(\pi_1 : \pi_2)$ does not indicate whether π_1 or π_2, is more informative (less concentrated).

The KL divergence is related to the extent of the reduction of uncertainty from the uniform distribution π_0^*. If $\mathcal{H}(\pi_0^*)$ is finite, then

$$\mathcal{K}(\pi : \pi_0^*) = \mathcal{H}(\pi_0^*) - \mathcal{H}(\pi) = \Delta(\pi : \pi_0^*). \tag{5}$$

On an unbounded support, π_0^* is not a proper distribution and $\mathcal{H}(\pi_0^*)$ is not well-defined. In that case π_0^* can be viewed as an idealization for comparing the predictive informativeness of distributions. It also can be viewed in terms of truncating the unbounded supports and extending the truncation bounds (a, b) such that $\mathcal{H}(\pi_j) \approx \mathcal{H}(\pi_j|(a, b)), j = 1, 2$ (Zellner, 1971); this idea is similar to locally uniform Bayesian priors.

2.2. ME model

A class of models where the KL divergence gives (5) in terms of (1) is as follows. Consider the following class of distributions on support $\mathbb{Q}$:

$$\Omega_\theta = \left\{\pi : E_\pi[T_k(q)] = \theta_k,\ k = 1, \dots, K\right\}, \tag{6}$$

where T_k's are integrable functions called *types of moments*, θ_k's are the *moment parameters*, and the moment equations in (6) are the *moment constraints*.

The ME model in Ω_θ is defined by the following optimization problem: $\pi^* = \arg\max_{\pi \in \Omega_\theta} \mathcal{H}(\pi)$. Let $\boldsymbol{T}(q) = (T_1(q), \dots, T_k(q))^T$ and $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)^T$, where supersrcipt T stands for the transpose. If

$$Z(\lambda) = \int_{\mathbb{Q}} \exp\left\{-\lambda^T \boldsymbol{T}(q)\right\} d\nu(q) < \infty, \tag{7}$$

then the ME model in Ω_θ is unique with the PDF in the following form:

$$\pi^*_{Q|\theta}(q) = \frac{1}{Z(\lambda)} \exp\left\{-\lambda^T \boldsymbol{T}(q)\right\}, \quad q \in \mathbb{Q}, \tag{8}$$

and $\lambda^T = (\lambda_1, \dots, \lambda_K)$ are Lagrange multipliers. $Z^{-1}(\lambda)$ is the normalizing factor for π^* and $Z(\lambda)$ is called the *partition function* in statistical physics (Jaynes, 1957, 1968). The existence of the ME model is determined by finiteness condition (7).

The entropy of the ME model is given by $\mathcal{H}(\pi^*_{Q|\theta}) = \log Z(\lambda) + \lambda^T \boldsymbol{\theta}$. Jaynes (1968) used maximum entropy reduction for assessing the value of an additional constraint by comparing the entropies of two nested ME models on a finite support. The extent of the reduction of uncertainty of moment constraints in (6) is given by the *information distinguishability* (ID) by the following equality between the KL and entropies of $\pi, \pi^*_{Q|\theta} \in \Omega_\theta$ Soofi et al. (1995):

$$\mathcal{K}(\pi : \pi^*_{Q|\theta}) = \mathcal{H}(\pi^*_{Q|\theta}) - \mathcal{H}(\pi) \geq 0,$$

provided that the entropies are finite.

The Fisher information matrix of the ME distribution (8) about λ is defined by

$$\mathcal{F}(\lambda) = E_{\pi^*}\left[\frac{\partial \log \pi^*}{\partial \lambda} \frac{\partial \log \pi^*}{\partial \lambda^T}\right].$$

In the traditional statistics the focal quantity is the model parameter λ. But in the IT approach the information moment vector $\boldsymbol{\theta}$ is the *focal parameter* and the model parameter λ is the implied Lagrange multiplier. By the uniqueness of the ME model $\lambda = \eta(\boldsymbol{\theta})$, where $\eta : \Re^K \to \Re^K$ is a continuous invertible function with partial derivatives on $\mathbb{D}_\lambda \subseteq \Re^K$. The following relationships are well-known (Jaynes, 1957; Kullback, 1959):

$$\boldsymbol{\theta} = -\frac{\partial}{\partial \lambda} \log Z(\lambda), \quad \frac{\partial \boldsymbol{\theta}}{\partial \lambda} = \Sigma^*, \quad \frac{\partial \lambda}{\partial \boldsymbol{\theta}} = \Sigma^{*^{-1}}, \tag{9}$$

where Σ^* is the covariance matrix of $\boldsymbol{T}(Q)$ with the ME distribution π^*. These relationships give the following result.

Lemma 1. *The Fisher information matrix of the ME model* (8) *about the moment parameters is given by*

$$\mathcal{F}(\boldsymbol{\theta}) = \mathcal{F}^{-1}(\lambda) = \Sigma^* = E_{\pi^*}[(\boldsymbol{T}(Q) - \boldsymbol{\theta})(\boldsymbol{T}(Q) - \boldsymbol{\theta})^T]. \tag{10}$$

For $T(q) = q$, $\theta = E(Q)$, and $\mathcal{F}(\theta)$ reduces to the variance of Q. By (9), this result relates the sensitivity of the ME model parameters to changes in the given moment parameters.

3. Noisy moment forecasts framework

If the moment parameter vector in (6) is a given forecast $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, then there will typically be some nonzero probability of the constraints in (6) being violated for any fixed Lagrangian multiplier. We consider the ME problem when the uncertainty about the moment parameter $\boldsymbol{\theta}$ is induced by the noise $\boldsymbol{\theta} = \boldsymbol{\zeta}(\boldsymbol{\theta}_0, \boldsymbol{\epsilon})$ where $\boldsymbol{\epsilon}$ is the vector of noise with a distribution g_ϵ such that $E(\boldsymbol{\theta}) = \boldsymbol{\theta}_0$. The uncertainty about the noisy moment $\boldsymbol{\theta}$ propagates through (6) into the ME model $\pi^*(q|\boldsymbol{\theta})$ in (8) and its parameter λ.

Fig. 1 depicts the propagation of the noise in the ME modeling outputs as a system. The ME model is conditional on the outcome of the noisy moments $\boldsymbol{\theta}$ and its Lagrange multipliers are conditional, as well. The conditional ME model and the distribution of the noisy moment provide the marginal distribution of Q:

$$\pi_Q(q) = \int \pi^*_{Q|\theta}(q|\theta) g_\theta(\theta) d\theta = \int \frac{\exp\{-\lambda^T \boldsymbol{T}(q)\}}{Z(\lambda)} g_\lambda(\lambda) d\lambda, \tag{11}$$

where g_λ is induced by g_θ via the one-to-one function $\lambda = \eta(\theta)$. The second integral represents the traditional mixture model $\pi_{mix}(q) = \pi_Q(q)$ based on a prior for λ.

The enumerated nodes in Fig. 1 are two well-known information measures as outputs of the noisy moments ME procedure. Each of these

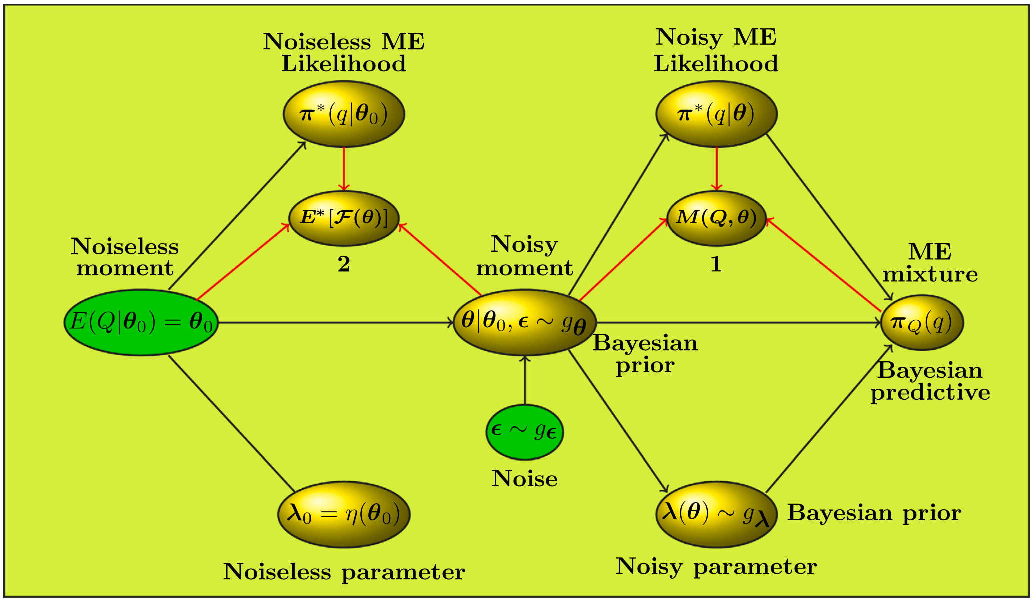

Fig. 1. Inputs (green) and outputs (golden) of noiseless and noisy ME model; links to distributions and parameters (black arrows); links for information measures (red arrows). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

measures is implied by three links shown in red. Node 1 is the *mutual information* between Q and θ which has the following representations:

$$M(Q,\theta) = \mathcal{H}(Q) - E_\theta[\mathcal{H}(Q|\theta)], \tag{12}$$

$$= E_\theta[\mathcal{K}(\pi_{Q|\theta} : \pi_Q)], \tag{13}$$

$$= \mathcal{K}(\pi_{\theta,Q} : \pi_Q g_\theta). \tag{14}$$

Representations (12) and (13) define $M(\theta,Q)$ as the *expected information* of the noisy moment constraint parameter θ by the uncertainty reduction (3) with $\mathcal{U} = H$ and the expected KL divergence. By (13), $M(\theta,Q)$ is a probabilistic sensitivity measure (Borgonovo et al., 2021) for assessing Q to the noisy inputs θ. The equality between (12) and (14) renders $M(\theta,Q)$ as the utility of dependence between Q and θ (Ebrahimi et al., 2014). Eqs. (12) and (13) are expressions for the Jensen–Shannon (JS) measure of the mixture model (11), i.e., $M(Q,\theta) = JS(\pi_{mix})$.

The noisy ME framework can be cast as a Bayesian model, where $\pi^*(q|\theta)$ provides the likelihood model with g_λ as the prior distribution for λ, and $\pi_Q(q)$ as the prior predictive distribution. Lindley's average amount of information about λ is defined by the expected entropy reduction $M(Q,\lambda) = \mathcal{H}(\lambda) - E_Q[\mathcal{H}(\lambda|Q)]$. Bernardo (1979) explicated Lindley's expected information in terms of the expected logarithmic utility of data for learning about the parameter. By the invariance of (14) under one-to-one transformation, $M(Q,\lambda) = M(Q,\theta)$, which implies that $M(Q,\theta)$ is the expected utility of constraints (6) about θ.

Node 2 in Fig. 1 is the expected Fisher information defined by

$$E_\theta[\mathcal{F}(\theta)] = \int \mathcal{F}(\theta) g_\theta(\theta) d\theta.$$

In general, $\mathcal{F}(\theta)$ is a function of the parameter θ and $E_\theta[\mathcal{F}(\theta)]$ accounts for the uncertainty about θ. As a prior Bayes estimate of the Fisher information, it is called the Bayes Fisher information (Asadi et al., 2019). This information measure is for assessing the sensitivity of model parameter to random variation of the moment parameter. By Lemma 1, $E_\theta[\mathcal{F}(\theta)]$ is a moment based sensitivity measure that generalizes the variance based sensitivity measure, see, e.g., Borgonovo et al. (2021).

3.1. Noisy mean and variance on continuous support

The PDFs of many ME distributions on continuous supports are in the exponential family (Ebrahimi et al., 2008). These ME likelihood models and their Bayesian conjugate priors in integrals (11) provide closed form predictive PDFs, which facilitate computing the mutual information by (12). We consider the most widely used ME/exponential family model.

The ME model with constraints $E(Y|\mu,\sigma^2) = \mu$ and $E[(Y-\mu)^2|\mu,\sigma^2] = \sigma^2$ on the real line is the normal distribution $\pi^*_{y|\mu,\sigma^2} = \mathcal{N}(\mu,\sigma^2),\ y \in \Re$. The Lagrange multipliers are $\lambda_1 = -\mu/\sigma^2$ and $\lambda_2 = 1/2\sigma^2$. On a finite interval, the ME model can be a truncated normal, upward or downward truncated exponential, uniform, or U-shaped depending on the relationship between θ_1 and $E(Y^2) = \sigma^2 + \mu^2$ (and the implied Lagrange multipliers); for details see Bajgiran et al. (2021) and references therein.

Consider the additive noise $\mu = \mu_0 + \epsilon$ with $g_\epsilon = N(0,\sigma^2)$. The implied distribution for the noisy mean is $g_\mu = \mathcal{N}(\mu_0,\sigma^2_\mu)$. The conditional predictive distribution is $\pi_{y|\sigma^2} = \mathcal{N}(\mu_0,\sigma^2_y)$ where $\sigma^2_y = \sigma^2 + \sigma^2_\mu$. The mutual information $M(\mu,Y|\sigma^2)$ can be written a follows:

$$M(\mu,Y|\sigma^2) = -\frac{1}{2}\log\frac{\sigma^2}{\sigma^2_y} = \frac{1}{2}\log\left(1+\frac{\sigma^2_\mu}{\sigma^2}\right) \tag{15}$$

$$= -\frac{1}{2}\log(1-\rho^2), \tag{16}$$

where $\rho = \rho(Y,\mu)$ is the correlation coefficient. Representation (16) is the well-known mutual information of the bivariate normal distribution.

The normal mutual information with noiseless variance is used in various problems including the regression model,

$$Y_i|\boldsymbol{x}_i,\boldsymbol{\beta},\sigma^2 = \boldsymbol{\beta}^T\boldsymbol{x}_i + \epsilon, \quad i = 1,\dots,N, \tag{17}$$

where $\boldsymbol{\beta}^T = (\beta_0,\beta_1,\dots,\beta_K)$, $\boldsymbol{x}_i = (1,x_{1i},\dots,x_{Ki})^T$ is a vector of predictors, and ϵ is the Gaussian noise with $g_\epsilon = N(0,\sigma^2)$. In the Bayesian approach, the prior for $\boldsymbol{\beta}$ is multivariate normal, and in the non-Bayesian stochastic regressors model, $\boldsymbol{X}_i$ is assumed to have a probability distribution, often a multivariate normal. Either of these approaches meets the requirements of (16) which are normal distributions for the mean $\mu_i|\boldsymbol{x}_i,\boldsymbol{\beta},\sigma^2 = \boldsymbol{\beta}^T\boldsymbol{x}_i$ and for the marginal distribution $\pi_{y|\sigma^2}$. Application of (16) gives the well-known regression mutual information in terms of the squared multiple correlation coefficient; see, e.g.,Theil and Chung (1988), Retzer et al. (2009), Ebrahimi et al. (2010a) and references therein.

Considering both moments of $\mathcal{N}(\mu,\sigma^2)$ are uncertain is more realistic than the assumption of the variance being given. The usual conjugate normal-gamma Bayesian prior is defined in terms of the precision parameter $\tau = 1/\sigma^2$ by $g_{\mu,\tau} = g_{\mu|\tau} g_\tau$ as follows: Let $g_{\mu|\tau} = \mathcal{N}(\mu_0, \sigma_\mu^2)$ with $\tau_\mu/\tau = c$, i.e., $\sigma_\mu^2/\sigma^2 = 1/c$ and $g_\tau = \mathcal{G}(\nu/2, \nu/2\tau_0)$ be the gamma (scaled-chi-square) prior with $E(\tau) = \tau_0 = 1/\sigma_0^2$ and following PDF:

$$g_\tau(\tau) = \frac{(\frac{\nu}{2\tau_o})^{\nu/2}}{\Gamma(\nu/2)} \tau^{\nu/2-1} e^{-\frac{\nu}{2\tau_o}\tau}, \quad \tau > 0,\ \nu > 0,$$

Then the marginal distribution π_y is Student-t with ν degrees of freedom, location parameter μ_0, and scale parameter $(1 + 1/c)\sigma_0^2$, denoted by $\pi_y = \mathcal{T}_\nu(\mu_0, (1 + 1/c)\sigma_0^2)$.

The following proposition gives the mutual information for noisy mean and variance.

Proposition 1. *The mutual information for the normal model with normal-gamma prior for the noisy mean and variance with $\sigma_\mu^2 = \sigma^2/c$ is as follows:*

$$M_\nu[(\mu, \sigma^2), Y] = M_c(\mu, Y) + M_\nu(\sigma^2, Y), \tag{18}$$

where

$$M_c(\mu, Y) = M(\mu, Y|\sigma^2) = \frac{1}{2}\log\left(1 + \frac{1}{c}\right), \tag{19}$$

$$M_\nu(\sigma^2, Y) = \log\frac{\Gamma(\frac{\nu}{2})}{\Gamma(\frac{\nu+1}{2})} + \frac{\nu+1}{2}\psi\left(\frac{\nu+1}{2}\right) - \frac{\nu}{2}\psi\left(\frac{\nu}{2}\right) - \frac{1}{2}, \tag{20}$$

and ψ is the digamma function.

Proof is given in Appendix A.

In general, information decompositions hold for stochastically independent variables or as chain rules in terms of conditional information measures (Cover & Thomas, 2006, Theorems 2.5.1–2.5.3). But with the normal-gamma prior $\rho^2(\mu, \sigma^2) = c/(c + 1) > 0$. The mean–variance decomposition in (18) is due to the location-invariance of entropy. In (18), $M_\nu(\sigma^2, Y) \geq 0$ takes the variance uncertainty into account via the degrees of freedom, which controls the level of uncertainty about the variance; for $\nu \geq 1, \mathcal{H}(g_\tau)$ is decreasing. The difference between (16) and (18), mapped by $M_\nu(\sigma^2, Y) \geq 0$, is decreasing and approaches zero as $\nu \to \infty$, and the marginal Student-t distribution approaches the normal distribution. These limiting behaviors imply the following asymptotic for (18): $M_\nu[(\mu, \sigma^2), Y] \approx M(\mu, Y|\sigma_\mu^2 = \sigma^2/c)$.

In regression analysis the error variance σ^2 and the squared correlation ρ^2 are unknown and their sample values are plugged in (15) and (16). Decomposition of information in (18) provides insights about the consequence of ignoring the uncertainty about the error variance. With the normal-gamma prior for $(\boldsymbol{\beta}, \tau)$, the marginal distribution of Y_i is a Student-t distribution. Furthermore, the predictive distribution of n future outcomes $y_i = \boldsymbol{\beta}^T \boldsymbol{x}_i, i = N + 1, \dots, N + n$ for given x_i's, is the multivariate-t (Zellner, 1971, section 3.2.4). Using the expression for $\mathcal{H}(\pi_y)$ (Zografos & Nadarajah, 2005) extends (18) to $M_\nu[(\boldsymbol{\beta}, \tau), \boldsymbol{Y}]$ for application to the Bayesian regression forecasting problem.

Transformation to normality and reparameterization of the likelihood model are frequently used in statistical analysis. Let $h : \mathfrak{R} \to \mathfrak{R}$ and $\xi : \mathfrak{R}^2 \to \mathfrak{R}^2$ be one-to-one functions. Then, by the invariance property of the mutual information under one-to-one transformation of each of its arguments, we have

$$M[(\mu, \sigma^2), Y] = M[\xi(\mu, \sigma^2), h(Y)].$$

The following examples give some transformations considered in statistics.

(a) The distribution of $X = h(Y) = e^Y$ is lognormal $\mathcal{LN}(\mu, \sigma^2)$ with mean $\mu_x = \exp(\mu + \sigma^2/2)$ and variance $\sigma_x^2 = [\exp(\sigma^2) - 1]\exp(2\mu + \sigma^2)$. The mapping $(\mu_x, \sigma_x^2) = \xi(\sigma^2, \mu)$ is one-to-one, so $M[(\mu_x, \sigma_x^2), X] = M[(\mu, \sigma^2), Y]$. More generally, the Box–Cox (BC) transformation is defined by $Y = B(X) = (X^c - 1)/c,\ P(X > 0) = 1, c > 0, c \neq 0$ where $B(x)$ is continuous at $x = 0$ and $Y = B(X) = \log X, c = 0$. Define the Inverse BC (IBC) transform by $X = h(Y) = (1 + cY)^{1/c},\ c > 0, c \neq 1$ and $X = h(Y) = e^Y,\ c = 0$. Then, $(\mu_x, \sigma_x^2) = \xi(\sigma^2, \mu)$ is one-to-one, and $M[(\mu_x, \sigma_x^2), X] = M[(\mu, \sigma^2), Y]$.

(b) Consider the following the functions of the normal parameters,

$$M[(\mu, \tau), Y] = M[(\mu, \sigma), Y] = M[(\mu, \log\sigma), Y].$$

More generally, let $\omega = \xi(\tau) = \tau^r, r > 0$. The distribution of ω is the generalized gamma $\mathcal{GG}(\nu/2, r, \zeta), \zeta = 2\sigma_0^2/\nu = 2/\tau_0\nu$ with PDF

$$g(\omega) = \frac{r}{\zeta\Gamma(\nu/2)}\left(\frac{\omega}{\zeta}\right)^{\nu r/2-1} \exp\{-(\omega/\zeta)^r\}, \quad \omega > 0,\ \nu, r, \zeta > 0.$$

This flexible family includes the following distributions: gamma ($r = 1$), Weibull ($\nu = 2$), exponential $[(\nu, r) = (2, 1)]$, Half-normal $[(\nu, r) = (1, 2)]$, Maxwell–Boltzmann $[(\nu, r) = (3, 2)]$, and lognormal as a limiting distribution. Moreover, the IBC transform of μ implies that $(\mu, \tau, Y) \to [h(\mu), \tau^r, Y]$ is one-to-one, and $M[(h(\mu), \sigma^2), Y] = M[(\mu, \sigma^2), Y]$. Hence, the normal-gamma prior generalizes to the family of priors constructed by transformations of the mean and generalized gamma distribution.

The Fisher information matrix of $\mathcal{N}(\mu, \sigma^2)$ is diagonal with elements $\mathcal{F}(\mu) = 1/\sigma^2 = \tau$ and $\mathcal{F}(\sigma^2) = 1/2\sigma^4 = \tau^2/2$. Using the first two moments of the gamma distribution gives $E_\tau[\mathcal{F}(\mu)] = \tau_0$ and $E_\tau[\mathcal{F}(\sigma^2)] = (1/\nu + 1/2)\tau_0^2$.

3.2. Noisy ME model on finite support

In applications of the noisy moments framework to a finite support $\mathbb{Q} = \{q_1, \dots, q_J\}$ the ME model is the distribution of probabilities $\pi^*|\boldsymbol{\theta} = (\pi_1^*|\boldsymbol{\theta}, \dots, \pi_J^*|\boldsymbol{\theta})$ given by (8) where the integral in (7) is summation and (8) is the following logit model:

$$\begin{aligned}\pi_{ij}^*|\boldsymbol{\theta} &= \Pr(Q = q_j|\boldsymbol{\theta}) \\ &= \frac{\exp\left\{\boldsymbol{\beta}^T \boldsymbol{T}(q_j)\right\}}{\sum_{h=1}^J \exp\left\{\boldsymbol{\beta}^T \boldsymbol{T}(q_j)\right\}}, \quad j = 1, 2, \dots, J,\ i = 1, \dots, N,\end{aligned} \tag{21}$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_K)^T$ is the vector of logit coefficients given by the Lagrange multipliers $\boldsymbol{\beta} = -\boldsymbol{\lambda} = -\eta(\boldsymbol{\theta})$ as in (8).

The noisy ME logit model (21) is the IT alternative for the random parameter logit model in the traditional literature, where $\boldsymbol{T}(q_j) = q_j$ and a normal distribution g_β is usually assumed for $\boldsymbol{\beta}$ (Harding & Hausman, 2007); some other distributions are also used when β_k's are nonnegative (Train, 2009). The second integral in (11) with (21) is intractable. Pakes and Pollard (1989) developed a general theory for the uniform convergence of simulation estimates. In the noisy moments framework a distribution for g_θ is assumed which induces g_β. The integrals in (11) with $\pi_{ij}^*|\boldsymbol{\theta}$ and g_θ (as well as the induced g_β) are intractable. For estimating the mutual information in Fig. 1, we utilize the simulation estimate of the marginal distribution π_Q. Consider data on the noisy moments $\boldsymbol{\theta}_i, i = 1, \dots, N$ as a random sample from G_θ. Such data can be produced by simulations when a model is assumed for g_θ or by a forecasting model. The noisy moments produce N copies of the ME logit model (21) with $\boldsymbol{\beta}_i = -\eta(\boldsymbol{\theta}_i)$.

The simulation estimate of π_Q is defined by the following empirical approximation of (11):

$$\begin{aligned}\tilde{\pi}_Q(q_j) &\approx \int h_1(q_j, \boldsymbol{\theta}) dG_n(\boldsymbol{\theta}) \\ &= N^{-1}\sum_{i=1}^N \frac{\exp\left\{\boldsymbol{\beta}_i^T \boldsymbol{T}(q_j)\right\}}{\sum_{h=1}^J \exp\left\{\boldsymbol{\beta}_i^T \boldsymbol{T}(q_j)\right\}}, \quad j = 1, 2, \dots, J.\end{aligned} \tag{22}$$

The expression in (22) with $\boldsymbol{T}(q_j) = q_j$ appears in Harding and Hausman (2007) where $\boldsymbol{\beta}_i$'s are generated from a Gaussian distribution.

In the noisy ME framework $\boldsymbol{\beta}_i = -\eta(\boldsymbol{\theta}_i), i = 1, \dots, N$ are random samples from G_β induced by G_θ and $\tilde{\pi}_Q$ provides an estimate of the marginal distribution. Using the notation of Pakes and Pollard (1989), we let $h_1(Q, \boldsymbol{\theta}) = \pi_{Q|\theta}^*$ and $h_2(Q, \boldsymbol{\theta}) = h_1(Q, \boldsymbol{\theta}) \log h_1(Q, \boldsymbol{\theta})$ to obtain the estimate of the expected information of the moment constraints in (6) given by the following proposition.

Proposition 2. *Let $\theta_i, i = 1,\dots,N$ be a sample of noisy moments, $h_1(Q,\theta) = \pi^*|\theta_i$ be the conditional distribution of Q given $\theta = \theta_i$, $h_2(Q,\theta) = h_1(Q,\theta)\log h_1(Q,\theta)$, and $\tilde{\pi}_Q$, be the empirical marginal distribution (22). Then the expected information of the constraints in (6) is the following simulation estimate:*

$$\widetilde{M}(Q,\theta) \approx \mathcal{H}(\tilde{\pi}_Q) + \int h_2(q,\theta)dG_n(\theta)$$

$$= \mathcal{H}(\tilde{\pi}_Q) - N^{-1}\sum_{i=1}^{N}\mathcal{H}(\pi^*|\theta_i) = N^{-1}\sum_{i=1}^{N}\mathcal{K}(\pi^*|\theta_i : \tilde{\pi}_Q) \geq 0. \quad (23)$$

The expected Fisher information does not involve (22). Lemma 1 with the sample average of the second moment of $\boldsymbol{T}(Q)$ gives

$$\widetilde{\mathcal{F}}(\theta) \approx N^{-1}\sum_{i=1}^{N} E_{\pi^*|\theta_i}\left(\boldsymbol{T}(Q)-\theta_i\right)\left(\boldsymbol{T}(Q)-\theta_i\right)^T. \quad (24)$$

A large number of data points provides reliable estimates of the information measures in Fig. 1. For the convergence properties of these estimates, see Pakes and Pollard (1989) on convergence properties of optimization estimators.

In the econometric literature, a parametric model with continuous PDF $g_{\beta|\alpha}$ is used for the distribution of the random parameter logit model. The hyper-parameter $\boldsymbol{\alpha}$ is estimated by the maximum likelihood (Cameron & Trivedi, 2005, p. 514) which requires data on the occurrences of outcomes, or a least squares approach (Harding & Hausman, 2007), which requires data on marketshares. The ME procedure does not rely on such occurrence or share data. However, when the share data is available, the procedure suggested by Harding and Hausman (2007) can be used for estimating the parameters of g_θ by the PDF transformation with $\theta = -\eta^{-1}(\beta)$. Pakes and Pollard (1989) theory does not require smoothness of g_β.

Remark 1. The family of ϕ-divergences, $\mathcal{D}_\phi(\pi_1 : \pi_2)$, includes the KL information, some of its generalizations, and several other well-known measures that can be used for the noisy moment ME framework. We have chosen the KL divergence as our focal measure because it provides the mutual information with attractive properties such as the uncertainty reduction representation (12), closed form expressions for many well-known continuous probability distributions, and coinciding with the mixture's Jensen–Shannon divergence. A member of the ϕ-divergence family with the attractive feature of symmetry and normalization is the *total variation distance (TV)*, defined by the normalized L_1-norm

$$TV(\pi_1,\pi_2) = \frac{1}{2}\int_{\mathbb{Q}}|\pi_1(q)-\pi_2(q)|d\nu(q). \quad (25)$$

Borgonovo (2007) defined the *moment independent sensitivity indicator* by

$$\xi^{TV} = E_\theta[TV(\theta)] = \frac{1}{2}\int\int_{\mathbb{Q}}|\pi_{\theta,Q}(\theta,q) - g_\theta(\theta)\pi_Q(q)|d\nu(q)d\mu(\theta). \quad (26)$$

Borgonovo et al. (2014) included (13) and (26) as examples of ϕ-divergence for the invariance in the sensitivity analysis. Borgonovo et al. (2021) called ξ^{TV} the *δ-importance* and showed that (13) and (26) are information value under a proper scoring rule. The integrals in (25) and (26) are intractable for the normal and many other continuous PDFs. We will use (26) for the ME models on finite support.

Various normalized mutual information indices are proposed in the literature. We use the normalized expected information index defined by the following expected entropy reduction due to the constraints (6) on the finite support:

$$I_M(\theta,Q) = \frac{M(\theta,Q)}{\mathcal{H}(\pi_Q)} = 1 - \frac{E_\theta[\mathcal{H}(Q|\theta)]}{\mathcal{H}(Q)}. \quad (27)$$

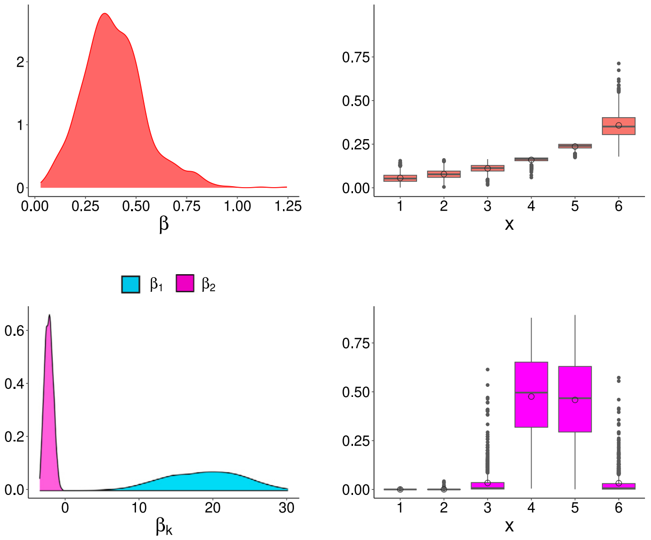

Fig. 2. Density plots of the ME model parameters and box plots of the ME probabilities of 1,000 simulations of the noisy mean ME model (top panels) and the noisy mean and variance ME model (bottom panels).

3.3. Noisy mean and variance on finite support

To illustrate the entropy concentration theorem, Jaynes (1982) introduced the following loaded dice problem (which has become a widely used problem in the entropy literature). He considered two scenarios. In the first scenario, there is no information beyond the possible outcomes X of the dice, so the ME model is the uniform distribution π_0^* on the support $\mathbb{X} = \{1,\dots,6\}$, which coincides with Laplace's Principle of Insufficient Reason. In the second scenario, he supposed that there is "evidence for some systematic influence causing the distribution to depart from the uniformity", so that the average number of spots is 4.5 (which exceeds 3.5 of the uniform distribution π_0^* for a fair dice).

We consider two noisy cases for the loaded dice problem.

(a) The noisy mean with $g_\mu = \mathcal{N}(\mu_0, \sigma_\mu^2)$.

(b) The noisy mean and variance with the normal-gamma prior $g_{\mu,\tau} = g_\mu g_\tau$, where $\tau = 1/\sigma^2$ is precision parameter of the distribution of X on the support $\mathbb{X} = \{1,\dots,6\}$, and $g_\tau = \mathcal{G}(a,b)$; a and b are the shape and scale parameters.

In both cases the ME models are logit (21), where for (a) $T(x) = x$ and for (b) $T_1(x) = x, T_2(x) = x^2$.

For implementing the noisy ME models, we simulated 1,000 noisy moments from $g_\theta = \mathcal{N}(4.5, 1/9)$ for (a) and $g_{\theta_1,\tau} = \mathcal{N}(4.5, 1/9)\mathcal{G}(8.8, 1.95)$ for (b). The prior variance σ_μ^2 chosen such that a substantially large number of draws reflect the upward bias; i.e., $E(\mu) > 3.5$ with a high probability. For (b), we set $E(X) = \mu_0 = 4.5$ and $E(\sigma^2) \approx 2.3$ as the variance of Jaynes' ME model. Among the simulation outcomes only two were slightly smaller than 3.5 which produced $\beta < 0$ reflecting downward bias.

Fig. 2 displays the density plots of model parameters $\beta = -\lambda$ for (a) and $\beta_k = -\lambda_k, k = 1,2$ for (b) and the box plots of the distributions of the noisy ME probabilities for the two cases. The mean of the model parameter for (a) is $\bar{\beta} = 0.388$ and for (b) $\bar{\beta}_1 = 18.70$ and $\bar{\beta}_2 = -2.08$. The box shows the middle 50% of the distribution and the line in the box shows the median probability. The circle inside the box shows the mean and the dots indicate outliers defined by 1.5 times the inter-quartile range above or below the quartiles. For case (a), the distributions of the ME probabilities show exponential upward shifts and for (b) the distributions of ME probabilities shift according to a truncated normal pattern.

Table 1
Noiseless and mixture of the noisy ME distributions of the loaded dice outcomes.

	Input[1]		x								$\mathcal{F}^{-1}(\theta)$		
	θ or g_θ	β	1	2	3	4	5	6	H	$I_{\Delta^*}(\theta)$	$\mathcal{F}^{-1}_{11}$	$\mathcal{F}^{-1}_{12}$	$\mathcal{F}^{-1}_{22}$
Mean													
Noiseless	4.5	.371	.054	.079	.114	.165	.240	.347	1.614	.10	.44		
Mixture	g_θ		.058	.080	.113	.161	.235	.352	1.617				
Mean and Precision													
Noiseless	(5.5,2.3)	(.366,.001)	.055	.079	.114	.165	.240	.348	1.614	.10	12.60	−1.56	.20
Mixture	g_{θ_1,θ_2}		.0000	.0004	.0333	.4757	.4585	.0321	.938				

[1] $g_\theta = \mathcal{N}(4.5, 1/9), g_{\theta_1,\theta_2} = \mathcal{N}(4.5, 1/9)\mathcal{G}(8.8, 1.95)$.

Table 2
Information measure of two noisy ME models for the loaded dice.

	Input					$\tilde{\mathcal{F}}^{-1}(\theta)$		
	g_θ	$\tilde{H}$	$\widetilde{M}(\theta, Q)$	$I_{\widetilde{M}}$	$\widetilde{TV}$	$\tilde{\mathcal{F}}^{-1}_{11}$	$\tilde{\mathcal{F}}^{-1}_{12}$	$\tilde{\mathcal{F}}^{-1}_{22}$
Mean	$\mathcal{N}(4.5, 1/9)$	1.593	.023	.015	.073	.462		
Mean and precision	$\mathcal{N}(4.5, 1/9)\mathcal{G}(8.8, 1.95)$	.753	.185	.197	.563	1281.44	−142.06	15.85

[1] $g_\theta = \mathcal{N}(4.5, 1/9), g_{\theta_1,\theta_2} = \mathcal{N}(4.5, 1/9)\mathcal{G}(8.8, 1.95)$.

Table 1 shows the noiseless ME models and the mixture distributions of the simulated noisy moment ME models. For the Jaynes' noiseless model, $\beta = 0.371 > 0$, which produces an exponentially increasing $\pi^*(x)$ for the upward bias $E(X) = 4.5 > 3.5$. The mixtures of the noisy mean ME probabilities (second row) are very close to the noiseless ME probabilities, in spite of the variations shown in the upper right panel of Fig. 2. For the noiseless ME model with mean and variance constraints, $\beta_1 = 0.366 > 0$ and $\beta_2 = 0.001$, which indicate that the addition of the second moment has a negligible effect on the ME distribution, and that is a discrete analog of an upward truncated exponential on $\mathbb{X}$. But the mixtures of the noisy mean and variance ME probabilities (fourth row) is very different from the noiseless ME probabilities. This mixture model distributes more than 93% of the distribution almost equally to $x = 4, 5$. For noiseless ME models only the entropy, the normalized (5) of the constraints defined by $\Delta^*(\theta)/\mathcal{H}(\pi^*_0)$, and the inverse of Fisher information Σ^* given by Lemma 1 are available. The entropy of the noiseless mean constraint model is close to the uniform distribution and the ID index gives the relative cost of 10% for the minimum adjustment of the uniform distribution to satisfy the given information about the mean. The entropy of the mixture of the noisy mean ME probabilities is about the same as that for the noiseless ME model. But the entropy of the mixture of the noisy mean ME probabilities is substantially less than that for the noiseless ME model and the inverse of Fisher information indicates substantial sensitivity of β_1 to the small changes of the moment parameters.

Table 2 compares the two noisy models according to the empirical mutual information $\widetilde{M}$, information index $I_{\widetilde{M}}$, the total variation index $\widetilde{TV}$, and the expected inverse of the Fisher information $\widetilde{\mathcal{F}}$. For the noisy mean ME model, these measures indicate little probabilistic and parameter sensitivity. For the noisy mean and variance ME model, these measures indicate probabilistic sensitivity and substantially higher parameter sensitivity of β_1 to the random variation of $\boldsymbol{\theta}$.

Assigning probability to outcomes of the loaded dice exemplifies the more general problem of assigning probability to alternatives (e.g. different products, different projects) based on some attributes (e.g. price, cost) across the alternatives.

4. Choice probability forecasts

The probabilistic choice models estimate the probabilities of different choices for a random sample of N DMs from a set of alternatives $\mathbb{Q} = \{A_1, \dots, A_J\}$, given data on alternative-varying attribute variables and/or DM's attributes. We consider the problem with alternative-varying attribute variables represented by the $K \times J$ matrix $X_i = [x_{ijk}], i = 1, \dots, N, j = 1, \dots, J, k = 1, \dots, K$, where the rows are data on K attributes which vary across J columns for the alternatives. We denote the choice probabilities by $\pi_{ij} = P_i(Q = A_j), i = 1, \dots, N, j = 1, \dots, J$, and the distribution for each DM on the support $\mathbb{Q}$ by the vector $\boldsymbol{\pi}_i = (\pi_{i1}, \dots, \pi_{iJ})^T$. An analogy to the loaded dice problem can be made by considering N dies with J faces that are loaded differently and $T(q_{ij}) = x_{ijk}$.

In the traditional statistics and econometrics a probability model is formulated as $\pi_{ij} = F(\boldsymbol{\beta}^T \boldsymbol{x}_{ij})$, where F is a cumulative distribution function, $\boldsymbol{x}_{ij} = (x_{ij1}, \dots, x_{ijK})^T$ is the jth column vector of individual ith choice attribute matrix X_i, and $\boldsymbol{\beta}^T = (\beta_1, \dots, \beta_K)$ is a vector of unknown parameters. The data requirement for estimating $\boldsymbol{\beta}$, in addition to the choice attributes, includes choices already made by the sample of N DMs, defined by $y_{ij} = 1$ if DM i has chosen alternative j and $y_{ij} = 0$ otherwise, for all $i = 1, \dots, N, j = 1, \dots, J$.

In probability forecasting problems, future attribute scores, denoted as $X_{i(t+1)}, i = 1, \dots, N_{t+1}$, are given, but data on choice indicators $y_{ij(t+1)}, i = 1, \dots, N_{t+1}$ are not available because future choices are not known yet; stated preference data gathered in surveys, if available, is subject to controversy (Cameron & Trivedi, 2005, p. 499).

Traditional forecasting approaches first estimate the model parameter $\boldsymbol{\beta}$ using data available at time t, denoted here as $(x_{ijkt}, y_{ijt}), i = 1, \dots, N_t$. The likelihood function for $\boldsymbol{\beta}$ is defined by $L(\boldsymbol{\beta}) = \prod_{i=1}^{N_t} \prod_{j=1}^{J} [F(\boldsymbol{\beta}^T \boldsymbol{x}_{ijt})]^{y_{ijt}}$. The MLE is defined by $\hat{\boldsymbol{\beta}}_t = \arg\max L(\boldsymbol{\beta})$. Then probability forecasts are computed by $\hat{\pi}_{ij(t+1)} = F(\hat{\boldsymbol{\beta}}_t^T \boldsymbol{x}_{ij(t+1)})$.

4.1. Individual level ME choice probability forecasts

The data required for deriving the ME choice probability forecasts for individual DMs consists of attribute matrix $X_{i(t+1)}$ and the attribute forecasts defined by

$$\theta^f_{ik(t+1)} = \sum_{j=1}^{J} \pi_{ij} T_k(x_{ijk(t+1)}) = \boldsymbol{\pi}_i^T \boldsymbol{T}(\boldsymbol{x}_{ik(t+1)}), \quad k = 1, \dots, K, \tag{28}$$

where $\boldsymbol{x}_{ik(t+1)}$ is the kth row of X_i. The attribute forecast data can be estimated from the MLE probability forecasts $\hat{\theta}^f_{ik(t+1)} = \hat{\boldsymbol{\pi}}^T_{i(t+1)} \boldsymbol{x}_{ik(t+1)}$ or provided by the experts in the application problem. For example, at time t a manager asks her data scientist to compute the probability forecasts that the firm would be selected by a DM who is considering J producers of a product which will be offered with new prices and sizes for time $t + 1$. The data scientist's probability model based on the new attributes will be different from the current probability model. The weights can be found in various ways, such as expert opinion, market shares, and applying probabilities given by a time t model to the time $t + 1$ specifications.

The system of Eqs. (28) displays two notable features: (a) Each attribute varies across the alternatives in the choice set $\mathbb{Q}$ with probabilities π_{ij}. (b) Each attribute forecast θ_{ik} is a linear function of π_{ij}, and therefore is a moment of the probability distribution $\boldsymbol{\pi}$.

In terms of (6), consider the following class of distributions:

$$\Omega_{\theta_i} = \left\{\boldsymbol{\pi}_i : \boldsymbol{\pi}_i^T \boldsymbol{T}(\boldsymbol{x}_{ik(t+1)}) = \theta_{ik(t+1)}, \quad k = 1,\dots,K\right\}. \tag{29}$$

If Ω_{θ_i} is not empty, for $K < J-1$, (28) is not determined and there are uncountably many solutions for π_i.

The following proposition gives the ensemble of individual level ME models in Ω_{θ_i} for N DMs.

Proposition 3. *Let* $\boldsymbol{\theta}_{i(t+1)} = (\theta_{i1(t+1)},\dots,\theta_{iK(t+1)})^T$. *Then* (21) *gives the ensemble of individual level ME models* $\pi_i^{*f}|X_{i(t+1)}, \boldsymbol{\theta}_{i(t+1)}$ *with elements given by the following probabilities:*

$$\pi^{*f}_{ij(t+1)}|X_{i(t+1)}, \boldsymbol{\theta}^f_{i(t+1)} = \frac{\exp\{\boldsymbol{\beta}^T_{i(t+1)}\boldsymbol{T}(\boldsymbol{x}_{ij(t+1)})\}}{Z_{i(t+1)}}, \quad i = 1,\dots,N_{t+1}, j = 1,\dots,J. \tag{30}$$

where $\boldsymbol{\beta}_{i(t+1)} = -\boldsymbol{\lambda}_{i(t+1)}$, $\boldsymbol{x}_{ij(t+1)}$ *is the* j*th column of* $X_{i(t+1)}$ *and* $Z_{i(t+1)} = Z(\boldsymbol{\lambda}_{i(t+1)})$.

To simplify notations conditioning on $X_{i(t+1)}$ and/or $\theta_{i(t+1)}$ will be used in (30) only when emphasis is needed.

Proposition 3 introduces a general family of individual level choice models. In (30), $T_k(x_{ijk})$ represents various functions of x_{ijk}. Two important special cases are as follows.

(a) When in (30) $T_k(x_{ijk(t+1)}) = x_{ijk(t+1)}$, Proposition 3 gives the following ensemble of individual level generalization of the conditional logit model:

$$\pi^{*f}_{ij(t+1)} = \frac{\exp\{\boldsymbol{\beta}^T_{i(t+1)}\boldsymbol{x}_{ij(t+1)}\}}{\sum_{h=1}^J \exp\{\boldsymbol{\beta}^T_{i(t+1)}\boldsymbol{x}_{ih(t+1)}\}}, \quad i = 1,\dots,N_{t+1}, j = 1,\dots,J. \tag{31}$$

(b) When in (30) attributes are positive and $T_k(x_{ijk(t+1)}) = \log(x_{ijk(t+1)})$, Proposition 3 gives the following ensemble of individual level generalization of the multiplicative competitive interaction model:

$$\pi^{*f}_{ij(t+1)} = \frac{\prod_{k=1}^K x_{ijk(t+1)}^{\beta_{ik(t+1)}}}{\sum_{h=1}^J \prod_{k=1}^K x_{ihk(t+1)}^{\beta_{i(t+1)}}}, \quad i = 1,\dots,N_{t+1}, j = 1,\dots,J. \tag{32}$$

The individual moments $\theta_{ik(t+1)}, i = 1,\dots,N, k = 1,\dots,K$ provide data for application of Proposition 2 to (31). The marginal PDFs of (30)–(32) are given by the ME mixture model (22) with $\pi^*_{ij}|X_i, \boldsymbol{\theta}_{i(t+1)}$. The entropies of the individual level ME models and their marginal distributions provide the empirical mutual information (23). By (Bernardo, 1979) interpretation of (13), (23) measures the expected utility of constraints (29) for the ME model (30).

The future attribute scores are often subject to uncertainty. Then $X_{i(t+1)}, i = 1,\dots,N$ are generated from the probability distribution of the random matrix X. The empirical mutual information (23) for the ensemble of ME forecast models written as $\widetilde{M}(X,Q)$ provides a measure of dependence between Q and the random attribute score matrix.

Remark 2. The ensemble of conditional logit model (31) is the IT counterpart of the traditional random coefficients logit choice model specified by Train (2009, p. 157), where the model parameters $\boldsymbol{\beta}_i, i = 1,\dots,N$ vary across the DMs $i = 1,\dots,N$ to represent each DM's "taste". However the two approaches are fundamentally different. In the random coefficients logit model, $\boldsymbol{\beta}_i, i = 1,\dots,N$ are unknown parameters assumed to have a continuous PDF $g(\boldsymbol{\beta}|\boldsymbol{\alpha})$. The hyper-parameter $\boldsymbol{\alpha}$ is estimated using data from a sample of individuals who have already chosen a set of alternatives with different attributes, $(X_i, y_{ij}), i = 1,\dots,N, j = 1,\dots,J$. Then $\boldsymbol{\beta}_i$'s are estimated by $E(\boldsymbol{\beta}_i|X_i, y_{ij})$ via simulations (Train, 2009).

In contrast, in the ME derivation of the ensemble of logit models (31), $\boldsymbol{\beta}_i, i = 1,\dots,N$ are given by the Lagrangian multipliers for the individual constraints in (28) that vary across the DMs. For a sample of N DMs from a population, the set of attribute moments $\boldsymbol{\theta}_{i(t+1)}, i = 1,\dots,N$ constitutes a sample of the population attribute moments. The distribution of population attribute moments is unknown and induces uncertainty about the population ME model parameters. The empirical distributions of $\boldsymbol{\theta}_{i(t+1)}$'s and $\boldsymbol{\beta}_i$'s represent the distribution of the respective population parameters. Thus, unlike the traditional random coefficients logit model, the individual level ME model (31) does not require specification of probability distributions for the moment and model parameters. However, using a parametric distribution for the noisy moments is an option, as shown in Section 3.3.

4.2. ME/MLE attribute forecasts

The existing ME logit choice models are derived by pooling the attribute moments $\boldsymbol{\theta}_{ik(t+1)}$ for N DMs in terms of $\sum_{i=1}^N \boldsymbol{\theta}_{ik(t+1)}$ (Soofi, 1992, 1994). A slight reformulation of constraints for deriving these logit models in terms of (6) is to let

$$\Omega_{\bar{\theta}} = \left\{\pi : N^{-1}\sum_{i=1}^N \pi_i^T \boldsymbol{x}_{ik} = \bar{\theta}_k, \quad k = 1,\dots,K\right\}. \tag{33}$$

For $K < N(J-1)$, constraints (33) provide the following *pooled moment ME model*:

$$\pi^*_{ij}|X, \bar{\boldsymbol{\theta}} = \frac{1}{Z}\exp\{\boldsymbol{\beta}^T\boldsymbol{x}_{ij}\}, \quad i = 1,\dots,N, \tag{34}$$

where $\bar{\boldsymbol{\theta}} = (\bar{\theta}_1,\dots,\bar{\theta}_K)^T$ and $X = [X_1 : \cdots : X_N]$ is the $K \times NJ$ matrix of the attribute data for the entire N DMs.

The distinction between the individual level logit model (31) and the pooled ME logit model (34) is notable. Unlike the individual level logit model (31), where $\boldsymbol{\beta}_i$ varies across the N DMs, $\boldsymbol{\beta}$ in (34) is the same for all DMs. This distinction is due to the fact that (33) includes K information constraint for all N DMs and (28) includes K constraints for each DM, which accounts to NK constraints for N DMs.

A common Lagrange multiplier still produces different attribute forecasts $\theta_{ik(t+1)}, i = 1\dots,N, k = 1,\dots,K$ due to the variation of $X_{i(t+1)}, i = 1,\dots,N$ across the sample for application of Proposition 2. A special case of the empirical mutual information (23) arises when the marginal distribution of the pooled moment ME logit model (34) is uniform, $N^{-1}\sum_{i=1}^N \pi^*_{ij}|X_i = 1/J, j = 1,\dots,J$. For this case, (27) gives the expected information in terms of (5) for (31), called by Soofi (1994) the information value of the set of constraints. This relationship provides additional insight about the existing information indices of the logit models in terms of representations of the mutual information and their interpretations.

The pooled moment model (34) includes the MLE of the logit model as a special case (Soofi, 1992). The MLE logit model is in the form of (34) where $\boldsymbol{\beta}$ is replaced with $\hat{\boldsymbol{\beta}}$ defined by the vector of solutions to $\partial \log L(\boldsymbol{\beta})/\partial\boldsymbol{\beta} = \boldsymbol{0}$, which can be represented as follows:

$$N^{-1}\sum_{i=1}^N\sum_{j=1}^J \pi_{ij}x_{ijk} = N^{-1}\sum_{i=1}^N\sum_{j=1}^J y_{ij}x_{ijk} = \hat{\theta}_k, \quad k = 1,\dots,K. \tag{35}$$

With $\bar{\theta}_k = \hat{\theta}_k$ in (33) gives the pooled ME model logit with $\hat{\boldsymbol{\beta}} = -\eta(\hat{\boldsymbol{\theta}})$. We call this special case the *full information ME/MLE*.

This model facilitates application of (31) for probability forecasting as stated next.

Corollary 1. *Denote the pooled ME/MLE attribute forecasts by*

$$\hat{\boldsymbol{\pi}}^{fT}_{i(t+1)}\boldsymbol{x}_{ik(t+1)} = \hat{\theta}^f_{ik(t+1)}, \quad i = 1,\dots,N_{t+1}, k = 1,\dots,K, \tag{36}$$

and the matrix of future attribute scores by $X_{i(t+1)} = [x_{ijk(t+1)}]$. *Then the individual level ME probability forecast distribution consistent with the attribute forecasts* (36) *is given by* (34) *where* $\boldsymbol{\beta}_{i(t+1)} = -\eta(\hat{\theta}^f_{i(t+1)}), i = 1,\dots,N_{t+1}$.

4.3. Illustrative data analysis: Fishing data

The fishing data (Cameron & Trivedi, 2005) is a sample of choices made by 1182 individuals in the United States from four fishing modes (beach, pier, boat, and charter), two choice attributes (price and the catch rate), and one individual-specific attribute (income). We use the choice attributes to compare the individual level ME forecast model (31) with the pooled full information ME/MLE model (34) and three versions of the traditional random parameters logit models.

We split the data into two randomly selected subsamples of equal sizes ($N = 591$), denoted as (X_t, y_t) and (X_{t+1}, y_{t+1}), to represent time t and time $t+1$ data, respectively. The subsamples are used as follows.

(a) The time t complete data (X_t, y_t) is used for computing coefficients $\hat{\boldsymbol{\beta}}_t$ of pooled ME/MLE logit and the traditional random parameters models.

(b) We apply the estimated coefficients to the time $t+1$ predictors X_{t+1} to compute the forecast probabilities of these models and attribute forecasts $\theta^f_{ik(t+1)}, k = 1, 2$ in (36).

(c) The time $t+1$ predictors X_{t+1} and attribute forecasts $\hat{\theta}^f_{ik(t+1)}$ are used to compute the individual level ME forecast model (31).

(d) We use time $t+1$ data (X_{t+1}, y_{t+1}) to compute the time $t+1$ MLE (full information ME/MLE) model to serve as the benchmark, π^{BM}, for comparing accuracies of probability forecasts obtained from the models under consideration.

Table 3 summarizes the results for moment forecasts, model parameters, and the following two measures of forecast accuracy: the averages of the KL divergences, and the mean absolute errors (MAE) between each forecast model π^f and π^{BM}. The moments parameters $\hat{\theta}_{k(t+1)}, k = 1, 2$ for the benchmark model are computed using (35), which are treated here as actual $\theta_{k(t+1)} = \hat{\theta}_{k(t+1)}$ for the purpose of comparison with other models. So its error measures are zero. Economic consideration suggests negative price coefficient and positive catch rate coefficient. Two versions of the individual level ME model are considered: one uses the attribute forecasts for all 591 individuals with average model parameters, and one restricted to 585 cases where the Lagrange multipliers in (31) produced $\beta_{i1(t+1)} < 0$ and $\beta_{2i(t+1)} > 0$. The individual level ME model parameters reported in the table are averages of logit coefficients obtained from individual Lagrange multipliers $\bar{\beta}_k = \sum_{i=1}^{N_{t+1}} \beta_{ik}/N_{t+1}$. The individual level logit coefficients varied to great extents, however, their averages reported in the table are very close to the coefficients of the pooled ME/MLE logit models used for computing the moment forecasts. The results for the two versions of the individual level ME model are approximately identical (for the number of decimal places shown in the table). The results for the three versions of the traditional random parameters models are substantially different, particularly substantially larger catch rate coefficient and higher forecast errors. Also notable in the table are that all forecast models underestimate the moment forecast for the price $\hat{\theta}_{1(t+1)}$ which is more serious for the traditional random parameters models with the uniform and triangular distributions. The individual level ME models slightly underestimate the moment forecast for the catch rate $\hat{\theta}_{2(t+1)}$ while the traditional random parameters model overestimate it, again substantially.

Table 4 shows the information measures for assessing the probabilistic sensitivity (divergence measures) and parameter sensitivity (expected Fisher information) of the benchmark model (a pooled moments ME/MLE) and the individual level ME forecast model. The forecast model is slightly more probabilistically sensitive than the benchmark model. The table includes the elements of the expected Fisher information matrix. Accordingly, the price coefficients of the two models are not sensitive but their catch rate coefficients are highly sensitive. The cross-sensitivity between the price and catch rate is negative.

We also examined probabilistic sensitivity and forecast error of the individual level ME forecast model (31) by adding Gaussian noises to the ME/MLE moment forecasts (36) with $g_{\theta_{ik}} = \mathcal{N}(\theta_{0ik}, \sigma^2_{ik})$; the upper limits for coefficients of variation $cv_{ik} = \sigma_{ik}/\theta_{0ik}$ are chosen such the moments remain positive with very high probabilities.

Table 3
Comparison of the individual level ME forecast model (time t model parameter and time $t+1$ attribute data) with the traditional random parameters models.

	Price		Catch rate		Forecast error[1]	
	$\bar{\theta}_1$	β_1	$\bar{\theta}_2$	β_2	$\bar{\mathcal{K}}$	MAE
Benchmark π^{BM}	53.882	−.018	.396	.929	0	0
Forecast model π^f						
Individual level ME[2]						
β_1, β_2 unrestricted	50.899	−.023	.389	.955	.007	.019
$\beta_1 < 0, \beta_2 > 0, n = 585$	50.800	−.023	.391	.972	.007	.019
Traditional random parameters logit[3] (mixed logit)						
Normal $\beta_k \sim N(\mu_k, \sigma^2_k)$	50.131	−.028	.427	1.478	.023	.036
Triangular $\beta_k \sim Tr(0, a_k)$	49.480	−.027	.396	1.140	.018	.031
Uniform $\beta_k \sim U(0, a_k)$	46.904	−.039	.407	1.545	.059	.060

[1] Error measures are reference to the benchmark model; [2] Model parameters are obtained by the averages of individual Lagrange multipliers; [3] Model parameters are estimated expected values of β_{kt}.

Fig. 3 displays plots of the empirical sensitivity measures (left) and the two forecast error measures (right) for samples of noisy moments θ_{ik} as functions of $cv \in (0, .2]$. These plots reflect the fact that when the noisy moments variances increase, variations of the forecast probabilities increase and they become farther from the marginal distribution; i.e., the marginal distribution (average probabilities) becomes less representative. The right panel indicates that when variations of the forecast probabilities increase they become farther from π^{BM}_i's given by the benchmark model shown in Table 3.

5. Concluding remarks

Probabilistic forecasting seeks to provide distributions for future outcomes of a variable of interest based on some partial information. The maximum entropy approach is a formal method to derive optimal forecast distributions consistent with the given moment forecasts. The moments forecasts are uncertain and estimated using currently available data or expert judgments. We proposed an ME framework where the moment forecasts are treated as noisy. The mutual information and expected Fisher information measures are included for assessing probabilistic and parameter sensitivity, respectively. This framework is equivalent to the Bayesian model where the prior distribution is assigned to the moments instead of the likelihood model parameters.

The noisy ME framework led to some new information measures and ME models. First, the mutual information of the normal model with normally distributed mean and given variance is used in various applications. We derived a new mean–variance decomposition of the mutual information using the usual normal-gamma prior for the more realistic problem when both moments are uncertain. Extensions of this result to various generalizations of the normal distribution and other priors provide interesting research topics for future.

Second, computing the mutual information requires the marginal distribution. The ME models on finite supports are in logit form and the integral for computing the marginal distribution is intractable. We proposed an estimate for the mutual information.

Third, this paper developed an ensemble of individual level forecast models with parameters varying across the sample – recognizing that individual attribute forecasts vary across the sample. In the context of the probabilistic choice modeling, the existing pooled ME logit models provide information about the average behavior, but the new individual level ME logit model provides information about the individual's behavior. Traditionally, probability forecasts are estimated by plugging in the future attribute scores into the maximum likelihood estimate of the logit computed with the complete sample currently available. We utilize the traditional attribute forecasts as the moment parameters.

Table 4
Probabilistic and parameter sensitivity of the benchmark and forecast models.

	Divergence measures			Expected Fisher information		
	$\widetilde{M}$	$I_{\widetilde{M}}$	$\widetilde{TV}$	$\bar{E}[\mathcal{F}]_{11}$	$\bar{E}[\mathcal{F}]_{12}$	$\bar{E}[\mathcal{F}]_{22}$
Benchmark π^{BM}	.225	.165	.262	0.056	–4.719	5206.821
Individual level ME forecast model π^{f}	.275	.203	.290	0.052	–5.403	7814.991

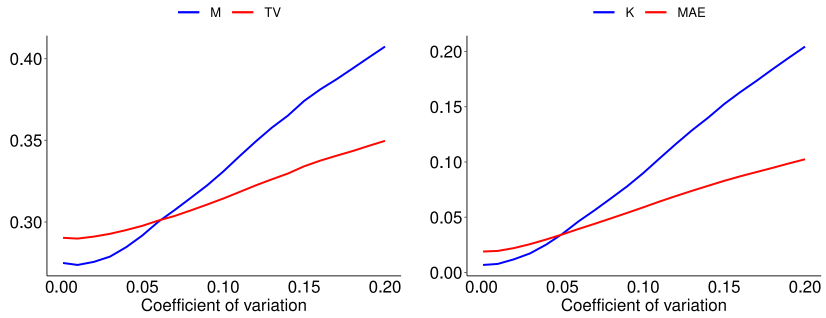

Fig. 3. Plots probabilistic sensitivity measures (left) and forecast errors (right) of Gaussian noise added to individual level ME model.

This new ME model advances the existing ME logit models for forecasting application problems. An illustrative application example indicated that the individual level ME logit model produces substantially more accurate forecasts than the traditional random parameters logit models.

Some remaining issues provide interesting problems for future research. Implications of the proposed noisy moments ME framework for Bayesian analysis are currently under study. In this paper, the noisy ME models are conditional on the given attribute forecasts. We noted that the literature provides various approximations of the marginal distributions of the noisy ME models on the finite support. Evaluating the implications of these approximations for estimating the mutual information can be insightful. We also briefly pointed out an interpretation of the empirical mutual information when the attribute scores are random draws from a probability distribution. Considering parametric models for distributions of the attribute scores provides a challenging research problem for future research.

CRediT authorship contribution statement

Omid M. Ardakani: Writing – review & editing, Software, Investigation, Data curation. **Robert F. Bordley:** Writing – review & editing, Methodology. **Ehsan S. Soofi:** Writing – original draft, Methodology, Conceptualization.

Acknowledgments

The authors are thankful to three reviewers for their comments that led to substantial improvement of the exposition of this paper.

Appendix A

Proof of Proposition 1

Letting $\sigma_\mu^2/\sigma^2 = 1/c$ in (15) gives (19). The mutual information (20) is

$$M(\sigma^2, Y|\mu) = \mathcal{H}(\pi_{y|\mu}) - E_{\sigma^2}\left[\mathcal{H}\left(\pi^*_{y|\mu,\sigma^2}\right)\right], \tag{A.1}$$

where the distribution of Y conditional on μ is Student-t, $\pi_{y|\mu} = \mathcal{T}_\nu[\mu, (1+1/c)\sigma_0^2]$, and

$$\mathcal{H}[\mathcal{T}_\nu(\mu, \sigma_0^2)]$$
$$= \log \frac{\sqrt{\nu\pi}\,\Gamma(\nu/2)}{\Gamma(\nu/2+1/2)} + \frac{\nu+1}{2}[\psi(\nu/2+1/2) - \psi(\nu/2)] + .5\log[(1+1/c)\sigma_0^2]. \tag{A.2}$$

The expression for the normal entropy with $\sigma^2 = 1/\tau$ and the expression for $E_\tau(\log\tau)$ give

$$E_\tau[\mathcal{H}(\pi^*_{y|\mu,\sigma^2})]$$
$$= .5 + .5\log(2\pi) - E_\mu(\log\tau) = .5 + .5\log(2\pi) - .5\psi(\nu/2) + .5\log(\nu/2\tau_0). \tag{A.3}$$

By the invariance of mutual information $M[(\mu,\sigma^2), Y] = M[(\mu,\tau), Y]$. Substituting (A.2) with $\tau_0 = 1/\sigma_0^2$ and (A.3) in (A.1) we obtain $M(\sigma^2, Y|\mu)$. The location-invariance of entropy (as seen in (A.2) and (A.3)) gives (20).

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ejor.2024.12.024.

References

Asadi, M., Ebrahimi, N., Kharazmi, O., & Soofi, E. S. (2019). Mixture models, Bayes Fisher information, and divergence measures. *Institute of Electrical and Electronics Engineers. Transactions on Information Theory*, *65*, 2316–2321.

Bajgiran, A. H., Mardikoraem, M., & Soofi, E. S. (2021). Maximum entropy distributions with quantile information. *European Journal of Operational Research*, *290*, 196–209.

Bernardo, J. M. (1979). Expected information as expected utility. *The Annals of Statistics*, *7*, 686–690.

Bordley, R. F. (2011). An anti-ideal point representation of economic discrete choice models. *Economics Letters*, *110*, 60–63.

Borgonovo, E. (2007). A new uncertainty importance measure. *Reliability Engineering System Safety*, *92*, 771–784.

Borgonovo, E., Hazen, G. B., Jose, V. R. R., & Plischke, E. (2021). Probabilistic sensitivity measures as information value. *European Journal of Operational Research*, *289*, 595–610.

Borgonovo, E., Tarantola, S., Plischke, E., & Morris, M. D. (2014). Transformations and invariance in the sensitivity analysis of computer experiments. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *92*, 5–947.

Brockett, P. L., Charnes, A., Cooper, W. W., Learner, D., & Phillips, F. Y. (1995). Information theory as a unifying statistical approach for use in marketing research. *European Journal of Operational Research*, *84*, 310–329.

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: methods and applications*. Cambridge University Press (Chapter 15).

Cooper, L. G. (1993). Market-share models. In *Handbooks in operations research and management science*: *vol. 5*, (pp. 259–314).

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory* (2nd ed.). John Wiley (Chapter 2).

DeGroot, M. H. (1962). Uncertainty, information, and sequential experiments. *The Annals of Mathematical Statistics*, *33*, 404–419.

Ebrahimi, N., Jalali, N. Y., & Soofi, E. S. (2014). Comparison, utility, and partition of dependence under absolutely continuous and singular distributions. *Journal of Multivariate Analysis*, *131*, 32–50.

Ebrahimi, N., Soofi, E. S., & Soyer, R. (2008). Multivariate maximum entropy identification, transformation, and dependence. *Journal of Multivariate Analysis*, *99*(6), 1217–1231.

Ebrahimi, N., Soofi, E. S., & Soyer, R. (2010a). Information measures in perspective. *International Statistical Review*, *78*, 383–412.

Ebrahimi, N., Soofi, E. S., & Soyer, R. (2010b). On the sample information about parameter and prediction. *Statistical Science*, *25*, 348–367.

Golan, A., Judge, G., & Perloff, J. M. (1996). A maximum entropy approach to recovering information from multinomial response data. *Journal of the American Statistical Association*, *91*, 841–853.

Granger, C., & Lin, J.-L. (1994). Using the mutual information coefficient to identify lags in nonlinear models. *Journal of Time Series Analysis*, *15*, 371–384.

Harding, M. C., & Hausman, J. (2007). Using a Laplace approximation to estimate the random coefficients logit model by nonlinear least squares. *International Economic Review*, *48*, 1311–1328.

Hong, T., Pinson, P., Fan, S., Zareipour, H., Troccoli, A., & Hyndman, R. J. (2016). Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond. *International Journal of Forecasting*, *32*, 896–913.

Huff, D. L. (1962). A probabilistic analysis of consumer spatial behavior. In W. S. Decker (Ed.), *Emerging concepts in marketing*. Chicago, IL: American Marketing Association.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical Review*, *106*, 620–630.

Jaynes, E. T. (1968). Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics*, *SEC-4*, 227–241.

Jaynes, E. T. (1982). On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, *70*, 939–952.

Kullback, S. (1959). *Information theory and statistics*. John Wiley, Reprinted in 1968 by Dover (Chapter 3).

Lahiri, K., Teigland, C., & Zaporowski, M. (1988). Interest rate and subjective distribution of inflation forecasts. *Journal of Money, Credit and Banking*, *20*, 233–248.

Lahiri, K., & Wang, W. (2020). Estimating macroeconomic uncertainty and discord using infometrics. In M. Chen, & et al. (Eds.), *Innovations in info-metrics: a cross-disciplinary perspective on information and information processing*. Oxford University Press.

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, *27*(4), 986–1005.

Ludwig, N., Arora, S., & Taylor, J. W. (2023). Probabilistic load forecasting using post-processed weather ensemble predictions. *Journal of the Operational Research Society*, *74*, 1008–1020.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in econometrics* (pp. 105–142). Academic Press.

Pakes, A., & Pollard, D. (1989). Simulation and the asymptotics of optimization estimators. *Econometrica*, *57*, 1027–1557.

Pourahmadi, M., & Soofi, E. S. (2000). Predictive variance and information worth of observations in time series. *Journal of Time Series Analysis*, *21*, 413–434.

Retzer, J. J., Soofi, E. S., & Soyer, R. (2009). Information importance of predictors: Concept, measures, bayesian inference, and applications. *Computational Statistics and Data Analysis*, *53*, 2363–2377.

Rich, R., & Tracy, J. (2010). The relationships among expected inflation, disagreement, and uncertainty: Evidence from matched point and density forecasts. *The Review of Economics and Statistics*, *92*, 200–207.

Shoja, M., & Soofi, E. S. (2017). Uncertainty, information, and disagreement of economic forecasters. *Econometric Reviews*, *36*, 796–817.

Simon, H. (1992). What is an explanation of behaviour? *Psychological Science*, *3*, 150–161.

Snyder, L. V., & Shen, Z.-J. M. (2011). *Fundamentals of supply chain theory*. John Wiley.

Soofi, E. S. (1992). A generalizable formulation of conditional logit with diagnostics. *Journal of the American Statistical Association*, *87*, 812–816.

Soofi, E. S. (1994). Capturing the intangible concept of information. *Journal of the American Statistical Association*, *89*, 1243–1254.

Soofi, E. S., Ebrahimi, N., & Habibullah, M. (1995). Information distinguishability with application to analysis of failure data. *Journal of the American Statistical Association*, *90*, 657–668.

Stone, M. (1959). Application of a measure of information to the design and comparison of regression experiments. *The Annals of Mathematical Statistics*, *29*, 55–70.

Theil, H., & Chung, C.-F. (1988). Information-theoretic measures of fit for univariate and multivariate linear regressions. *The American Statistician*, *42*, 249–252.

Tinsley, P. A., Spindt, P. A., & Friar, M. E. (1980). Indicator and filter attributes of monetary aggregates. *Journal of Econometrics*, *14*, 61–90.

Train, K. (2009). *Discrete choice models with simulation* (2nd ed.). Cambridge University Press.

Wallis, K. F. (2005). Combining density and interval forecasts: A modest proposal. *Oxford Bulletin of Economics and Statistics*, *67*, 983–994.

Wilson, T., & Bell, M. (2007). Probabilistic regional population forecasts: The example of Queensland, Australia. *Geographical Analysis*, *39*, 1–25.

Zellner, A. (1971). *An introduction to Bayesian inference in econometrics*. John Wiley, Reprinted in 1996 by Wiley (Chapter 2).

Zografos, K., & Nadarajah, S. (2005). Expressions for Rényi and Shannon entropies for multivariate distributions. *Statistics & Probability Letters*, *71*, 74–81.