

Simultaneous Inference under the Vacuous Orientation Assumption

Ruobin Gong

Department of Statistics, Rutgers University, New Jersey, USA

RUOBIN.GONG@RUTGERS.EDU

Abstract

I propose a novel approach to simultaneous inference that alleviates the need to specify a correlational structure among marginal errors. The *vacuous orientation* assumption retains what the normal i.i.d. assumption implies about the distribution of error configuration, but relaxes the implication that the error orientation is isotropic. When a large number of highly dependent hypotheses are tested simultaneously, the proposed model produces calibrated posterior inference by leveraging the logical relationship among them. This stands in contrast to the conservative performance of the Bonferroni correction, even if neither approaches makes assumptions about error dependence. The proposed model employs the Dempster-Shafer Extended Calculus of Probability, and delivers posterior inference in the form of stochastic three-valued logic.

Keywords: Dempster-Shafer theory; belief function; Bonferroni correction; familywise error rate; calibrated inference

1. Introduction

In scientific explorations, the analyst often needs to verify more than one hypothesis based on data from a single experiment. The multiplicity of hypotheses posits a threat to the trustworthiness of the overall conclusion. The inferential procedure employed, even if statistically valid for verifying single hypotheses, may no longer retain validity if not carefully compounded across hypotheses. The correct compounding of statistical procedures relies on adequate knowledge about the dependence relationship between the observed data as well as the hypotheses, neither of which is likely available to the analyst.

For concreteness, let $\mathbf{M} = (M_1, \dots, M_k)$ be a vector of unknown parameters, and $\mathbf{Y} = (Y_1, \dots, Y_k)$ a vector of observable data that aims to measure \mathbf{M} . Suppose for $i = 1, \dots, k$,

$$Y_i \sim N(M_i, S^2). \quad (1)$$

That is, each M_i is measured exactly once by Y_i , with a normally distributed measurement error of variance S^2 . Suppose for now that S^2 is known. The goal is to make inference about the uncertain values of \mathbf{M} without additional prior information.

It seems intuitive that for an individual M_i , a best guess at its value is Y_i , accompanied by a confidence statement as

an interval centered at Y_i with width proportional to S . This follows from the sampling model in (1). However, to make confidence statements about \mathbf{M} , (1) alone is not enough. Missing from the specification is a dependence structure among the observations given their unknown true means.

The scenario described here is an abstraction of a typical scientific experiment. A total of k unknown quantities are learned at the same time, and each Y_i is a summary statistic. As part of the inferential procedure, the marginal reference distributions of the Y_i 's are usually well-understood. However, information regarding their interdependence is much harder to come by. Often out of convenience, or perhaps a lack of better alternative, it is assumed that

$$Y_i \perp\!\!\!\perp Y_j \mid \mathbf{M} \quad \forall i \neq j, \quad (2)$$

hence \mathbf{Y} is multivariate normal with mean \mathbf{M} and covariance proportional to the identity matrix. We refer to (1) and (2) as the sampling model under the *normal i.i.d.* (or just i.i.d.) assumption. Figure 1 is an illustration of it for $k = 2$.

For problems of higher dimensions, computation can be vastly simplified if some form of independence can be assumed. Within the context of multiple hypothesis testing, a most widely adopted procedure to control the false discovery rate is the Benjamini-Hochberg procedure [1], which is valid under the assumption of independence or mild positive dependence [2]. However, independence (or for that matter, any known structure of dependence) among hypotheses is all but likely to hold. In particular, independence is easily violated when the parameters of interest and the summary statistics are devised sequentially according to previous observations and verification results. Such a trajectory is nevertheless a typical one in scientific explorations.

Things become more complicated when an even larger collection of hypotheses involving \mathbf{M} , highly collinear among themselves, are to be verified together. For example, one can test whether all pairwise contrasts are equal to zero: $M_i = M_j$, which makes a total of $k(k - 1)/2$ comparisons involving only k unknown quantities. This case is investigated in Example 3 in Section 3. Once the data is observed, one is always guaranteed to find *some* function g about the parameter, however obscure and scientifically insignificant it may be, such that the observed data produces strikingly strong support for the statement $g(\mathbf{M}) = 0$. This illegitimate maneuver is called “data-snooping”, and it invalidates the nominal significance level claimed by the test. Post-hoc

analyses aim to avoid it by controlling for the *familywise error rate*, that is, the probability of making at most one false rejection among a family of hypothesis tests when the nulls are all true.

A classic procedure to control for the familywise error rate is the Bonferroni correction [9]. If p hypotheses are tested together, the required significance level for each individual test is reduced to α/p , such that the familywise error rate is no more than α . Notably, the validity of the Bonferroni correction does not rely on an independence assumption about the hypotheses. However, it does not take into consideration the logical structure behind the hypotheses, contributing to its well-known conservative behavior. More discussions and comparisons involving the Bonferroni procedure can be found in Section 3.3.

From an estimation point of view, the model in (1) poses challenges to both the likelihood and Bayesian approaches. Since each M_i is measured exactly once, no large sample asymptotic approximation to the likelihood is appropriate, even when the variance parameter S^2 is known. In order to obtain a distributional description of \mathbf{M} , one must resort to Bayesian methods and assume a prior for the parameter vector \mathbf{M} . Again due to the small sample size, any prior will exude significant influence over the posterior for \mathbf{M} , which is undesirable unless a certain shrinkage effect is intentionally sought for. If S^2 is not known *a priori*, the maximum likelihood estimate for S^2 does not exist, whereas Bayesian methods need a prior for S^2 that cannot be updated based on the data. Challenges regarding unknown S^2 is addressed in Section 4. Lastly, if the dependence structure of \mathbf{Y} is not precisely available, both likelihood and Bayesian methods will fail to prescribe a variance quantification in their respective inference for \mathbf{M} .

Contribution of This Paper This paper discusses a novel approach to simultaneous inference that alleviates the need to specify a correlational structure among the marginal errors. In place of the i.i.d. assumption, the proposed approach invokes the assumption of vacuous orientation, a mathematically weaker assumption that delivers logically stronger conclusions. The vacuous orientation assumption retains what the i.i.d. assumption implies on the configuration of the measurement errors, i.e. their sum of squares follows a χ_k^2 distribution. On the other hand, it relaxes the implication that the error orientation is isotropic in \mathbb{R}^K .

The proposed model employs the Dempster-Shafer Extended Calculus of Probability (DS-ECP) [7, 8], a logical framework for probabilistic reasoning. DS-ECP dictates the specification, combination and processing of marginal information, including observed data, model structure, as well as available (but optional) prior distributional information. The model delivers posterior inference in the form of a probability triple (p, q, r) that sums to one, representing evidence “for”, “against”, and “don’t know” towards a hypothesis one wishes to ascertain. As a generalization

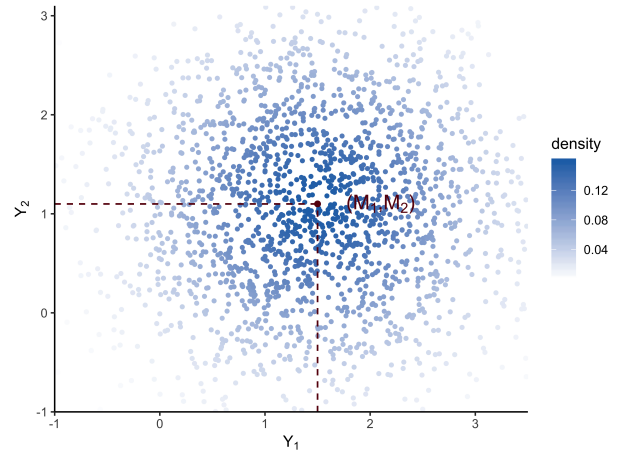


Figure 1: The normal i.i.d. sampling model of \mathbf{Y} given \mathbf{M} when $k = 2$. The i.i.d. assumption implies that 1) the sum of squares (configuration) of measurement errors is distributed χ_2^2 , and 2) their orientations are isotropic, i.e. uniform in all directions.

to ordinary probability reasoning, DS-ECP is a system of stochastic three-valued logic. The third value augmentation, i.e. a possibly nonzero r , provides a flexible vocabulary to express partial information, including the complete lack of information. An advantage of DS-ECP over the Bayesian approach is that does not require the modeler to possess *a priori* distributional knowledge about unknown parameters, while at the same time enjoying the logical coherence as does the Bayesian approach.

The DS-ECP analysis framework is a marriage between belief function, an imprecise probability construction, and a functional approach to statistical inference. The model specification in DS-ECP bears resemblance to structural [10] and functional models [4], as well as the modern approach of generalized fiducial inference [11]. Other statistical approaches that leverage non-additive probability measures include robust Bayes [3], inferential models (IM) [14], and outer probability measures [12]. Notably, the concept of validity emphasized by the IM approach in the sense of frequentist coverage is echoed in the analysis of calibration properties in Section 3.3 of this paper.

The remainder of this paper is organized as follows. Section 2 lays out the set of weakened assumptions underlying the proposed approach, and demonstrate the combination and projection operations following the inferential recipe of DS-ECP, including a small example to illustrate the recipe itself. Section 3 discusses posterior inference for important types of hypotheses, including linear and quadratic forms, their relationships to hypothesis testing and confidence regions, as well as frequentist coverage properties. In the

presence of an exploding number of dependent hypotheses, the posterior probabilities from the vacuous orientation model is uniformly distributed, just as a well-calibrated p -value under the null sampling model. In contrast, the Bonferroni approach, which boasts no assumption on the dependence structure of hypotheses, behaves unnecessarily conservatively. Section 4 discusses the case in which the error variance S^2 is not known but follows a prior distribution. Section 5 discusses potential generalizations to the proposed model.

2. Model

The proposed model is based on the following state space structure, consisting of¹

$$(\mathbf{Y}, \mathbf{M}, \mathbf{E}, S^2) \in \Omega, \quad (3)$$

where $\Omega = \Omega_{\mathbf{Y}} \times \Omega_{\mathbf{M}} \times \Omega_{\mathbf{E}} \times \Omega_{S^2} = \mathbb{R}^{3k+1}$. \mathbf{Y} is a k -vector of observable measurements, and \mathbf{M} the corresponding vector of unknown parameters whose values we wish to learn. \mathbf{E} is a vector of measurement errors and S^2 a variance parameter associated with them, both to be defined soon.

2.1. Marginal Evidence

A piece of marginal evidence is a mathematical statement regarding subsets of the state space. Observations, together with modeling assumptions, make up the marginal evidence available for analysis. In DS-ECP models, the word *margin* may carry a more general meaning than it does in ordinary statistical models, without being confined to individual dimensions of the state space. A margin can be subsets, as well as (single- or multi-valued) functions of subsets of the state space. This generalization gives the freedom to specify evidence on higher order structures.

The first piece of marginal evidence we specify is the assumption that the true means are observed subject to additive error. That is,

$$\mathbf{Y} - \mathbf{M} = \mathbf{E}. \quad (i)$$

The next piece of marginal evidence concerns the observed data. That is, the observable measurement \mathbf{Y} realizes to a particular value

$$\mathbf{Y} = \mathbf{y}. \quad (ii)$$

While this may look obvious, we spell it out explicitly in anticipation of potentially censored or truncated observations. The third assumption is on the distribution of *error configuration*:

$$\mathbf{E}'\mathbf{E} = S^2U, \quad \text{where } U \sim \chi_k^2, \quad (iii)$$

1. In general, we require the state space Ω be the Cartesian product of a collection of marginal spaces, and be endowed with the product topology.

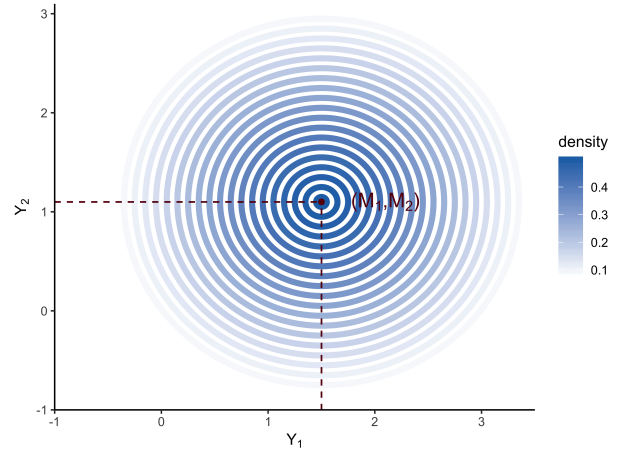


Figure 2: The error configuration assumption (iii) says that the sum of squares of the error terms follows a scaled χ_k^2 distribution. It posits a k -sphere centered at \mathbf{M} with a random radius. Assumption on the error orientation is vacuous, i.e., no knowledge of location on a given k -spherical shell is available, contrasting the isotropic orientation assumption implied by the normal i.i.d. model illustrated in Figure 1.

the Chi-squared distribution with k degrees of freedom. Notice that the sum of squares of the measurement errors is injected with stochastic evidence through the introduction of the variable U . It is the modeler's assertion that U be *auxiliary*, in the sense that (iii) is the only way through which the distribution of U injects knowledge into the state space. Besides through (iii), no further information can be learned from U about any aspect of the state space.

The error configuration assumption says that the sum of squares of the measurement errors follows a scaled χ^2 distribution with a scaling factor S^2 . As alluded to in Section 1, that the error configuration is distributed as Chi-squared is a necessary but insufficient consequence of the normal i.i.d. assumption. To see this insufficiency, consider the following probability specification for \mathbf{E} when the dimension $k \geq 2$: $E_1 \sim k^{-1/2} \cdot \text{sgn}(0.5) \cdot \chi_k$, where $\text{sgn}(0.5)$ is a random “+/-” sign, and χ_k is an independent Chi distribution with k degrees of freedom. Also, $E_i = E_1$ for all $i = 2, \dots, k$. The implied distribution of \mathbf{E} asserts that all E_i 's are fully correlated. This specification is compliant with the vacuous orientation model, but not with the normal i.i.d. model. Therefore, to replace the i.i.d. assumption with (iii) is a proper weakening of model assumptions. Conclusions derived from the vacuous orientation model are logically stronger.

The terminology *configuration* and *orientation* are due to [6], who used them in a regression context. For \mathbf{X} a linear regression design matrix, configuration refers to the matrix product $\mathbf{X}'\mathbf{X}$, whereas orientation refers to the information that remains in \mathbf{X} given $\mathbf{X}'\mathbf{X}$. The vacuous orientation assumption refers to the fact that only the marginal distribution of the error configuration is specified, while the error orientation is left unspecified. With this partial specification, the model admits to full ignorance on the correlational structure of the error terms \mathbf{E} . Figure 2 is an illustration of the error configuration assumption when $k = 2$.

We will need a last assumption concerning the variance parameter S^2 . Consider for now the *known variance* case:

$$S^2 = s^2, \quad (\text{iv})$$

for s a positive real number. In Sections 2 and 3, we discuss posterior inference for \mathbf{M} based on the known variance assumption. The unknown variance case is discussed in Section 4, where we introduce a variant of (iv) that allows S^2 to bear a prior distribution. Reasoning and computation contingent upon either variants are virtually the same.

2.2. Independence of Marginal Evidence

As a prerequisite for the next steps of DS-ECP analysis, it is important that the specified collection of marginal evidence be judged as an *independent* body of evidence. Independence justifies the use of Dempster's Rule to combine the marginal information in the joint state space.

To be concrete, let \mathbb{E} denote a collection of J pieces of marginal evidence. \mathbb{E} here stands for "Evidence", to be distinguished from \mathbf{E} the measurement error. For the model under contemplation, $\mathbb{E} = \{(i), (ii), (iii), (iv)\}$. Let U_j be the auxiliary variable associated with the j^{th} piece of evidence in \mathbb{E} , which follows a known distribution μ_j . For example, U in (iii) follows the χ_k^2 distribution. Marginal pieces of evidence that are deterministic, such as (i), (ii) and (iv) which are mere equality statements about margins of the state space, are regarded as associated with constant auxiliary variables. Evidence independence is defined as follows.

Definition 1 (Independent marginal evidence.) *A body of marginal evidence \mathbb{E} consisting of J pieces is said to be independent, if the auxiliary variables associated with each piece are all statistically independent. That is, for $U_j \sim \mu_j$ where $j = 1, \dots, J$, we have that*

$$(U_1, \dots, U_J) \sim \mu_1 \times \dots \times \mu_J. \quad (4)$$

Since constant variables are independent of other variables, all deterministic evidence are naturally independent. By Definition 1, the body of evidence for the vacuous orientation model $\mathbb{E} = \{(i), (ii), (iii), (iv)\}$ is independent, and is eligible for the next steps of DS-ECP analysis.

Note that the independence of evidence discussed here is distinct from the assumption of independence about the measurement errors. The latter is implied by the normal i.i.d. assumption, precisely what the proposed model attempts to rid by replacing with the mathematically weaker assumption of χ_k^2 error configuration. Contingent upon the configuration, the errors may or may not be independent. On the other hand, the independence of evidence assumption claims that the information that the error configuration behaves in a certain way is independent of other pieces of information, such as that the errors are additive, the variance is known, and so on. In other words, independence of evidence refers to statistical independence of the auxiliary variables, a separate notion from (in)dependence among margins of the state space. Independence of evidence is just as much a subjective judgment on the part of the modeler. In complex models, through careful specification using a join tree structure [13], one can successively construct independent bodies of evidence such that they're eligible for DS-ECP analysis using Dempster's Rule of Combination.

2.3. Evidence Projection and Combination

The processing of marginal evidence consists of three steps: 1) up-projection of marginal evidence to the joint state space, 2) combination of evidence in the joint state space, and 3) down-projection of combined evidence to margins of interest. The three steps are explicated in this section.

2.3.1. UP-PROJECTION OF MARGINAL EVIDENCE

To up-project a piece of marginally specified evidence is to extend its evidence statement involving only a subset of the state space variables, into one that concerns the entire state space. Algebraic relationships defined on margins of the state space are extended into cylinder sets, spanning the remainder dimensions that were not mentioned in the evidence statement. For example, (i) implies the following partitioning of Ω :

$$\{(\mathbf{Y}, \mathbf{M}, \mathbf{E}, S^2) \in \Omega : \mathbf{Y} - \mathbf{M} = \mathbf{E}\}, \quad (5)$$

and similarly for statement (iii),

$$\{(\mathbf{Y}, \mathbf{M}, \mathbf{E}, S^2) \in \Omega : \mathbf{E}'\mathbf{E} = S^2U\}. \quad (6)$$

In other words, the up-projection process levels the playground for all marginal evidence, such that they become comparable statements concerning the same joint space Ω , and ready to be combined there.

2.3.2. EVIDENCE COMBINATION VIA DEMPSTER'S RULE

Since $\mathbb{E} = \{(i), (ii), (iii), (iv)\}$ is judged to be independent, we apply Dempster's rule to combine its component pieces,

that is, by taking the intersections of all up-projected evidence whenever they are nonempty. These nonempty intersections become the new focal sets, representing the combined evidence from \mathbb{E} . In this model, the combination of \mathbb{E} implies a class of subsets of Ω of the following structure

$$\begin{aligned} R_{\mathbb{E}} &\stackrel{\text{def}}{=} \{(\mathbf{Y}, \mathbf{M}, \mathbf{E}, S^2) \in \Omega : \\ &\mathbf{Y} = \mathbf{y}, \mathbf{Y} - \mathbf{M} = \mathbf{E}, \mathbf{E}'\mathbf{E} = S^2U, S^2 = s^2\}, \end{aligned} \quad (7)$$

where $U \sim \chi_k^2$. Notice that $R_{\mathbb{E}}$ is a multi-valued map [5] from U to subsets of Ω . Since U bears a known distribution, $R_{\mathbb{E}}$ can be regarded as a random subset of Ω whose distribution is inherited from that of U . The probability density function of U dictates the mass function of $R_{\mathbb{E}}$.

Upon combining arbitrary body of evidence, the auxiliary variable distribution may need to be revised. To be precise, the domain on which the auxiliary variable is *a priori* defined may be reduced, to exclude those values that result in algebraic incompatibility among pieces of marginal evidence. These incompatible values are those that correspond to marginal focal sets that result in empty intersections with other marginal focal sets. Empty intersections are eliminated from the combination process, while the weights of the remainder non-empty intersections renormalize to one. The revision of the auxiliary distribution proceeds as follows.

Denote μ the prior probability of the auxiliary variable \mathbf{U} , associated with a body of evidence \mathbb{E} and measurable with respect to $\sigma(\mathfrak{E})$. Upon combining \mathbb{E} , μ is revised to $\mu_{\mathbb{E}}$, the posterior probability measurable with respect to $\sigma(\mathfrak{E}_{\mathbb{E}}) \subset \sigma(\mathfrak{E})$, where $\mathfrak{E}_{\mathbb{E}} = \{u \in \mathfrak{E} : R_{\mathbb{E}}(u) \neq \emptyset\}$, and

$$\mu_{\mathbb{E}} = (\mu \times \mathbf{1}_{\mathfrak{E}_{\mathbb{E}}}) / \mu(\mathfrak{E}_{\mathbb{E}}), \quad (8)$$

where the indicator function $\mathbf{1}_A(S) = 1$ if $S \subseteq A$ and 0 otherwise. In case the denominator $\mu(\mathfrak{E}_{\mathbb{E}})$ is 0, (8) may be alternatively defined via regular conditional probability or limiting arguments. For the case at hand, it just so happens that none of the four component evidence of $\mathbb{E} = \{(i), (ii), (iii), (iv)\}$ raises algebraic conflict for any given value of the auxiliary variable U . Thus, the revision of auxiliary distribution is trivial, namely $\mu_{\mathbb{E}} = \mu$ which is still the χ_k^2 distribution.

2.3.3. DOWN-PROJECTION TO MARGINS OF INTEREST

Rarely is the case that we wish to draw inference about the entire state space. Often, we are only interested in a particular margin of the state space, such as \mathbf{M} the parameter of interest. To reduce computational burden, the random subset $R_{\mathbb{E}}$ defined on Ω is projected onto the margin of interest $\Omega_{\mathbf{M}}$. This process is called down-projection. The projection of $R_{\mathbb{E}}$ onto $\Omega_{\mathbf{M}}$ is

$$R_{\mathbf{M}|\mathbb{E}} \stackrel{\text{def}}{=} \{\mathbf{M} \in \Omega_{\mathbf{M}} : (\mathbf{M} - \mathbf{y})'(\mathbf{M} - \mathbf{y}) = s^2U\} \quad (9)$$

where $U \sim \mu_{\mathbb{E}}$, the χ_k^2 distribution. $R_{\mathbf{M}|\mathbb{E}}$ is again a random subset of $\Omega_{\mathbf{M}}$ whose distribution is dictated by U . For every realization $U = u$, $R_{\mathbf{M}|\mathbb{E}}(u)$ is a k -sphere centered at \mathbf{y} with radius $s\sqrt{u}$. Repeated draws of U following the χ_k^2 distribution result in a collection of concentric k -spheres of varying radii. The random k -sphere $R_{\mathbf{M}|\mathbb{E}}$ embodies posterior inference for \mathbf{M} . Section 3 discusses posterior inference based on $R_{\mathbf{M}|\mathbb{E}}$, expanding on its properties in greater detail.

2.3.4. PROJECTION AND COMBINATION: AN EXAMPLE

We given an example to illustrate the projection and combination operations discussed above. Let $\mathbf{X} = (X_1, X_2)$ be two independent tosses of a same coin, whose chance of landing as head is θ . Suppose the following Bernoulli experiment for $i = 1, 2$:

$$X_i = \mathbf{1}(U_i \leq \theta), \quad (10)$$

where the auxiliary variables $U_i \stackrel{\text{iid}}{\sim} \mu_i = \text{Unif}(0, 1)$. A head followed by a tail was observed: $X_1 = 1$ and $X_2 = 0$. Gathering the above evidence as \mathbb{E}' , we have that

$$R_{\mathbb{E}'} = \{(\mathbf{X}, \theta, \mathbf{U}) : \mathbf{X} = (1, 0), X_i = \mathbf{1}(U_i \leq \theta)\}, \quad (11)$$

where the prior auxiliary variable distribution $\mu = \mu_1 \times \mu_2$ is uniform on the unit square. Notice that the data implies $U_1 \leq \theta$ and $U_2 > \theta$, hence for any given θ , U_2 must be greater than U_1 . This restricts the domain of the auxiliary variable to $\mathfrak{E}_{\mathbb{E}'} = \{\mathbf{u} \in [0, 1]^2 : u_1 < u_2\}$, the northwest triangle of the unit square. It follows from (8) that the revised auxiliary variable distribution $\mu_{\mathbb{E}'} \propto \mathbf{1}_{\mathfrak{E}_{\mathbb{E}'}}$, uniform over the northwest triangle of the unit square, which is equivalent to the probability distribution induced by a pair of Uniform order statistics. Down-projecting $R_{\mathbb{E}'}$ to the θ margin, we obtain the random subset that embodies posterior inference for θ :

$$R_{\theta|\mathbb{E}'} = \{\theta \in [0, 1] : U_1 < \theta \leq U_2\}, \quad (12)$$

where $\mathbf{U} \sim \mu_{\mathbb{E}'}$. $R_{\theta|\mathbb{E}'}$ is a half-closed, half-open random interval on $[0, 1]$. Its left and right end points are marginally distributed as $Beta(1, 2)$ and $Beta(2, 1)$ respectively.

3. Posterior inference

Posterior inference about unknown quantities in the state space are expressed through a probability triple (p, q, r) , representing weights of evidence “for”, “against”, and “don’t know” about an assertion concerning subsets of the state space. When $R_{\mathbf{M}|\mathbb{E}}$ is the down-projected random subset that embodies posterior inference for \mathbf{M} , it inherits randomness from the revised auxiliary variable $\mathbf{U} \sim \mu_{\mathbb{E}}$. Define a trio of set functions $p, q, r : \Omega_{\mathbf{M}} \rightarrow [0, 1]$ such that for all $H \in \sigma(\Omega_{\mathbf{M}})$,

$$p(H) = \int_{\{u \in \mathfrak{E}_{\mathbb{E}} : R_{\mathbf{M}|\mathbb{E}}(u) \subseteq H\}} d\mu_{\mathbb{E}}, \quad (13)$$

$$q(H) = \int_{\{u \in \Xi_{\mathbb{E}}: R_{\mathbb{M}|\mathbb{E}}(u) \subseteq H^c\}} d\mu_{\mathbb{E}} = p(H^c), \quad (14)$$

$$r(H) = 1 - p(H) - q(H), \quad (15)$$

with $p + q + r = 1$. Note that (p, q, r) are implicit functions of \mathbb{E} , a dependence we suppress for notational simplicity. The (p, q, r) representation of posterior inference is an alternative to using a pair of belief and plausibility functions [15]. In particular, p is a belief function on $\Omega_{\mathbb{M}}$. $1 - q$, or equivalently $p + r$, is its conjugate plausibility function. The three-valued representation has the advantage that it explicitly acknowledges a possibly non-zero r , the “don’t know” probability which reflects the extent of structural uncertainty within the model.

3.1. Linear Forms and Hypothesis Tests

Let \mathbf{C} be a real-valued p by k matrix, where p can be smaller than, the same as, or larger than k . A consistent system of equations

$$\mathbf{CM} = \mathbf{a} \quad (16)$$

is a linear margin of the parameter space $\Omega_{\mathbb{M}}$ of dimension $k - \text{rank}(\mathbf{C})$. Inference about linear margins of the parameter space is the most common type of posterior inference, encompassing a variety types of hypotheses.

We are interested in drawing inference about a linear margin in the form of (16). Define the summary statistic

$$t_{\mathbf{y}} = (\mathbf{a} - \mathbf{C}\mathbf{y})' (\mathbf{C}\mathbf{C}')^{-1} (\mathbf{a} - \mathbf{C}\mathbf{y}), \quad (17)$$

where in case $p > \text{rank}(\mathbf{C})$, the inverse is defined as the Moore-Penrose pseudoinverse. Results below concern posterior inference for one- and two-sided linear forms of \mathbf{M} , expressed in terms of $t_{\mathbf{y}}$.

Theorem 2 (Two-sided linear form) *For a two-sided linear hypothesis $H : \mathbf{CM} = \mathbf{a}$, The DS posterior probabilities concerning H are*

$$\{p(H), q(H), r(H)\} = \{0, F(t_{\mathbf{y}}), 1 - F(t_{\mathbf{y}})\} \quad (18)$$

where F is the CDF of the scaled χ_k^2 distribution with scaling factor s^2 .

Proof of Theorem 2 recognizes that $t_{\mathbf{y}}$ is the minimum square radius of k -spheres of the form $R_{\mathbb{M}|\mathbb{E}}$ to intersect with the linear subspace $\mathbf{CM} = \mathbf{a}$. The probability that $R_{\mathbb{M}|\mathbb{E}}$ does not intersect with the linear subspace contributes to $q(H)$, and the probability that it does contributes to $r(H)$.

Theorem 3 (One-sided linear form) *For a one-sided linear hypothesis $H : \mathbf{CM} \leq \mathbf{a}$, The DS posterior probabilities concerning H are*

$$\{p(H), q(H), r(H)\} = \{F(t_{\mathbf{y}}), 0, 1 - F(t_{\mathbf{y}})\} \quad (19)$$

if $\mathbf{C}\mathbf{y} \leq \mathbf{a}$, and

$$\{p(H), q(H), r(H)\} = \{0, F(t_{\mathbf{y}}), 1 - F(t_{\mathbf{y}})\} \quad (20)$$

otherwise. F is defined as in Theorem 2.

The hypothesis in Theorem 3 is a halfspace formed by the hyperplane which Theorem 2 posits as hypothesis. Depending on \mathbf{y} , the location of the center of $R_{\mathbb{M}|\mathbb{E}}$, the hypothesized halfspace may or may not contain $R_{\mathbb{M}|\mathbb{E}}$ with positive probability. If H is not supported at face value by empirical evidence, that is, either H is degenerate relative to $\Omega_{\mathbb{M}}$ or it asserts to the contrary of what the observation appears to be (e.g. $H : \mathbf{CM} \leq \mathbf{a}$ but $\mathbf{C}\mathbf{y} \not\leq \mathbf{a}$), then $p(H)$, the probability “for” H , is zero. On the other hand, if H is supported by empirical evidence, then $q(H)$, or the probability “against” H , is zero. The probability of “don’t know”, $r(H)$, is the same as long as H concerns the same linear subspace in $\Omega_{\mathbb{M}}$. This again demonstrates the intuitive appeal of r as a posterior summary statistic: its value is reflective of the inherent structural uncertainty regarding the hypothesis with respect to the collection of evidence. That is, it reflects the extent to which \mathbb{E} is able to discern anything about H at all, while agnostic to the direction of support for it.

Examples below showcase a variety of hypothesis types to which the above theorems are applicable. Posterior probabilities are examined in comparison to the frequentist answers under the normal i.i.d. model.

Example 1 (test for all means) *Suppose the null hypothesis is $H : \mathbf{M} = \mathbf{0}$, and the alternative $H^c : \mathbf{M} \neq \mathbf{0}$. \mathbf{C} is the identity matrix, and $\mathbf{a} = \mathbf{0}$. The test statistic simplifies to $t_{\mathbf{y}} = \sum_{i=1}^k y_i^2$. The posterior probability “for” H is 0, and posterior probability of “don’t know” about H is*

$$r(H) = \Gamma(k/2, \sum y_i^2 / 2s^2) / \Gamma(k/2), \quad (21)$$

which is the survival probability (i.e. one minus the cumulative probability) of a χ_k^2 distribution evaluated at $\sum y_i^2 / s^2$.

Notice that (21) is identical to the p -value one would obtain under the i.i.d. sampling model, for which the likelihood ratio test is based on the same test statistic. It has an exact reference distribution of χ_k^2 under the null. Example 2 to follow seems like a slight modification from Example 1, but the solution it supplies can be rather distinct.

Example 2 (test for one mean) *Suppose the null hypothesis is $H : M_1 = 0$, and the alternative $H^c : M_1 \neq 0$. Here, \mathbf{C} is a column vector of 0’s at all except the first entry, which takes the value of 1. The test statistic is $t_{\mathbf{y}} = y_1^2$. The posterior probability “for” H is 0, and posterior probability of “don’t know” about H is*

$$r(H) = \Gamma(n/2, y_1^2 / 2s^2) / \Gamma(n/2), \quad (22)$$

which is the survival probability of a χ_k^2 distribution evaluated at y_1^2 / s^2 .

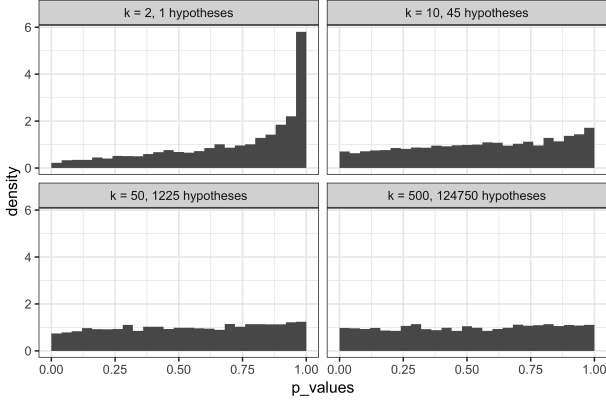


Figure 3: Distribution of $r(H)$ for all pairwise contrasts under the null sampling model for various k . As the dimension k increases, the distribution of $r(H)$ approaches uniform, which is the distribution of a correctly calibrated p -value.

When $k = 1$, Example 2 reduces to Example 1, and (22) agrees with the p -value obtained under the i.i.d. sampling model using the likelihood ratio test. However when $k > 1$, (22) is larger than the p -value obtained under the i.i.d. sampling model, which utilizes χ_1^2 as the reference distribution for $t_{\mathbf{y}}$ regardless of k . Under the vacuous orientation assumption, the reference distribution χ_k^2 grows with k , and is more conservative than the i.i.d. model.

Example 3 (test for all pairwise contrasts) Suppose we conduct a simultaneous test for the null hypothesis that all pairwise means are identical. That is,

$$H = \bigcap_{1 \leq i < j \leq k} H_{i,j}, \quad H_{i,j} : M_i = M_j$$

The alternative hypothesis H^c is that at least one of the equalities doesn't hold. \mathbf{C} is a $k(k-1)/2$ by k matrix of all pairwise contrasts with $\text{rank}(\mathbf{C}) = k-1$, and $\mathbf{a} = \mathbf{0}$. The test statistic simplifies to

$$t_{\mathbf{y}} = \sum_i (\bar{y} - y_i)^2 - \frac{1}{k} \sum_{i,j} (\bar{y} - y_i)(\bar{y} - y_j). \quad (23)$$

The posterior probability “for” H is zero, and that of “don't know” about H can be found via (20) with the reference distribution again a s^2 -scaled χ_k^2 distribution.

Figure 3 displays values of $r(H)$ for simulations under the null i.i.d. sampling model for various k . The number of pairwise contrasts tested is on quadratic order of k , but the compound hypothesis H always spans a 1-dimensional subspace of $\Omega_{\mathbf{M}}$. As k increases, the distribution of $r(H)$ using simulated null data approaches the uniform distribution, which is the distribution of a correctly calibrated p -value

under the null model. This stands in sharp contrast to the Bonferroni procedure. Observations obtained from a similar simulation experiment shows that, while the Bonferroni correction controls for the overall test size α , for larger k it becomes increasingly conservative, in that the fraction of null samples leading to a rejection of H is far less than α for all α in range. The culprit to Bonferroni's conservatism is that it adjusts the individual test size by mindlessly dividing α with the number of hypotheses tested. It neither respects nor utilizes the logical connection among the large number of hypotheses. The proposed model captures such feature, and delivers logically coherent posterior inference reflective of the geometry of the hypothesis space.

3.2. Quadratic Forms and Calibrated Credible Regions

The sampling distribution under the i.i.d. assumption is the spherical multivariate normal distribution. The associated confidence region for its mean vector is thus spherical. The proposed model \mathbb{E} induces random subsets in the parameter space in the form of concentric spherical shells. In this section, we consider credible regions for \mathbf{M} that are of quadratic forms. These credible regions are special in that they deliver *sharp* posterior inference that are also *calibrated* with respect to the normal i.i.d. model.

Definition 4 (sharp inference) Given a body of marginal evidence \mathbb{E} , we say that the posterior inference for $A \in \sigma(\Omega_{\mathbf{M}})$ is sharp if $r(A) = 0$.

Definition 5 (calibrated inference) Given \mathbb{E} , we say that posterior inference is lower-calibrated for $A \in \sigma(\Omega_{\mathbf{M}})$ with respect to a sampling model $\mathbf{Y} \mid \mathbf{M}^* \sim P^*$, if the posterior probability “for” A is equal to the frequentist coverage probability of A under the sampling model, when A is viewed as a function of \mathbf{Y} . That is,

$$p(A) = P^*(\mathbf{M}^* \in A). \quad (24)$$

Similarly, posterior inference for A is upper-calibrated if

$$q(A) = P^*(\mathbf{M}^* \in A^c). \quad (25)$$

Posterior inference for A is calibrated if it is both lower- (or upper-) calibrated and sharp.

For $\alpha \in [0, 1]$, define the $(1 - \alpha)$ posterior credible region

$$A_\alpha = \{ \mathbf{M} \in \Omega_{\mathbf{M}} : (\mathbf{M} - \mathbf{y})'(\mathbf{M} - \mathbf{y}) \leq F_{1-\alpha}^{-1} \}, \quad (26)$$

where F_α^{-1} is the α^{th} -quantile of the posterior auxiliary distribution $\mu_{\mathbb{E}}$. By (13) and (14), we have that

$$p(A_\alpha) = 1 - \alpha, \quad q(A_\alpha) = \alpha. \quad (27)$$

The posterior inference coincides with the i.i.d. model inference, when hypotheses of quadratic forms (such as A_α) are contemplated. This should come as no surprise, since when A_α is viewed as a quadratic function of \mathbf{y} , its probabilistic property is precisely the aspect of the i.i.d. assumption that is preserved under the vacuous orientation assumption.

Theorem 6 (sharp credible region.) A_α is a sharp posterior credible region for \mathbf{M} . That is, $r(A_\alpha) = 0$.

Theorems 6 is a direct consequence of (27). We also have the following.

Theorem 7 (calibrated credible region.) A_α is a calibrated posterior credible region with respect to the i.i.d. sampling model, for all \mathbf{M}^* and all α .

To prove Theorem 7, one just need to show that $\alpha = P^* \left((\mathbf{Y} - \mathbf{M}^*)' (\mathbf{Y} - \mathbf{M}^*) > F_{1-\alpha}^{-1} \right)$ for all \mathbf{M}^* and all α . That indeed is the case, when P^* is the normal i.i.d. sampling distribution specified in (1) and (2).

3.3. Rectangular Parallelepipedal Regions

Under the sampling model in (1), for $i = 1, \dots, k$, a size- α test for hypothesis $H_i : M_i = 0$ is dual to a confidence interval for M_i of the form $(y_i \pm \Phi(1 - \alpha/2) \cdot s)$. If a compound null hypothesis $H : \bigcap_{i=1}^k H_i$ is contemplated, and if the confidence intervals for each component hypothesis are calculated as above, the familywise error rate (i.e. the probability of making at least one false rejection) exceeds α . As alluded to in Section 1, the Bonferroni procedure accounts for the fact that k hypotheses are tested simultaneously. To maintain the familywise error rate at no more than α , the test size for each component hypothesis is reduced to α/k . Hence, a Bonferroni-corrected test for hypothesis H is dual to the rectangular confidence region

$$\bigotimes_{i=1}^k (y_i \pm b_\alpha \cdot s), \quad (28)$$

where $b_\alpha = \Phi(1 - \alpha/2k)$ is the Bonferroni-corrected and standardized half width of the univariate interval.

To parallel the Bonferroni confidence region, consider rectangular parallelepipedal regions of the form

$$C_\alpha = \left\{ \mathbf{M} \in \Omega_{\mathbf{M}} : \mathbf{M} \in \bigotimes_{i=1}^k (y_i \pm c_\alpha \cdot s) \right\}. \quad (29)$$

Posterior probabilities associated with C_α can be regarded as a function of the standardized half width c_α . We have the following results.

Theorem 8 The posterior probabilities of C_α are

$$\{p(C_\alpha), q(C_\alpha), r(C_\alpha)\} = \{F(c_\alpha^2), F(kc_\alpha^2) - F(c_\alpha^2), 1 - F(kc_\alpha^2)\} \quad (30)$$

where F is the CDF of the χ_k^2 distribution.

To prove Theorem 8, notice that $p(C_\alpha)$ is the probability that $\mathbf{R}_{\mathbf{M}|\mathbb{E}}$ is fully contained in C_α , i.e. its radius less than or equal to $c_\alpha \cdot s$. On the other hand, $q(C_\alpha)$ is the probability that $\mathbf{R}_{\mathbf{M}|\mathbb{E}}$ fully contains C_α , i.e. its radius greater than $\sqrt{k}c_\alpha \cdot s$. A lemma immediately follows.

Lemma 9 Let \underline{c}_α and \bar{c}_α be the standardized half widths of C_α such that it is lower-calibrated (i.e. $p(C_\alpha) = 1 - \alpha$) and upper-calibrated (i.e. $q(C_\alpha) = \alpha$), respectively. We have

$$\underline{c}_\alpha = (F_{1-\alpha}^{-1})^{1/2}, \quad \bar{c}_\alpha = (F_{1-\alpha}^{-1}/k)^{1/2} \quad (31)$$

where F_α^{-1} is the α^{th} -quantile of the χ_k^2 distribution.

Just like the Bonferroni half width b_α , \underline{c}_α is an increasing function of the dimension k . On the other hand, \bar{c}_α is a decreasing function of k . The three quantities are identical when $k = 1$. Table 1 displays a comparison among the three quantities for various k at $\alpha = 0.05$.

Table 1: Half widths of C_α for various k , $\alpha = 0.05$. $\bar{c}_\alpha = b_\alpha = \underline{c}_\alpha$ when $k = 1$. $\bar{c}_\alpha < b_\alpha < \underline{c}_\alpha$ when $k > 1$.

k	$\bar{c}_{0.05}$	$b_{0.05}$	$\underline{c}_{0.05}$
1	1.96	1.96	1.96
2	1.73	2.24	2.45
5	1.49	2.58	3.33
10	1.35	2.81	4.28
100	1.12	3.48	11.15

We also consider the posterior probabilities for rectangles to which the Bonferroni procedure assigns $1 - \alpha$ confidence. A comparison for various k and α is displayed in Table 2. As k increases, both the posterior “for” and “against” probabilities approach 0, and the posterior “don’t know” probability approaches 1. Unlike spherical regions, posterior inference for rectangular regions is not sharp under the vacuous orientation assumption. This reveals that, to make sharp probabilistic statements on rectangular regions require substantial model input concerning the dependence structure of observation errors. The growing extent of “don’t know” posterior probability quantifies the extent to which a sharp probabilistic assignment on a rectangular region need to depend on such assumptions, increasingly so as the dimension of the parameter space grows.

4. Case With Unknown Variance

As alluded to in Section 1, under the sampling model in (1) every unknown parameter M_i is measured only once. There is no hope to extract knowledge about the variance parameter S^2 from the data. If S^2 is not known precisely,

Table 2: Posterior probabilities associated with Bonferroni $(1 - \alpha)$ rectangles (28), for $\alpha = 0.05$ (left) and 0.2 (right).

k	$\alpha = 0.05$			$\alpha = 0.2$		
	p	q	r	p	q	r
1	0.95	0.05	0.00	0.80	0.20	0.00
2	0.92	0.01	0.07	0.74	0.07	0.19
5	0.75	0.00	0.25	0.48	0.00	0.52
10	0.36	0.00	0.64	0.14	0.00	0.86
100	0.00	0.00	1.00	0.00	0.00	1.00

it needs to be assumed to follow a prior distribution. This section describes a modification to the proposed model, to accommodate the case that S^2 is not known but rather follow a prior distribution. The *unknown variance* assumption posits that

$$S^2 = U_s, \quad (\text{iv.2})$$

where U_s is an auxiliary variable bearing some known prior distribution for S^2 , and is independent of the auxiliary variable U from (iii). Write $\mathbf{U} = (U, U_s)$, and denote by $\mathbb{E}.2 = \{(i), (ii), (iii), (iv.2)\}$ the body of evidence using the alternative unknown variance assumption. By assuming independence between U and U_s , $\mathbb{E}.2$ is judged independent according to Definition 1, and is eligible for combination via Dempster’s Rule. The projection and combination operations follow the same fashion as described in Section 3. $\mathbb{E}.2$ implies the following class of subsets of Ω :

$$\begin{aligned} R_{\mathbb{E}.2} = \{(\mathbf{Y}, \mathbf{M}, \mathbf{E}, S^2) \in \Omega : \\ \mathbf{Y} = \mathbf{y}, \mathbf{Y} - \mathbf{M} = \mathbf{E}, \mathbf{E}'\mathbf{E} = S^2\mathbf{U}, S^2 = U_s\}, \end{aligned} \quad (32)$$

and the down projection of $R_{\mathbb{E}.2}$ onto $\Omega_{\mathbf{M}}$ is

$$R_{\mathbf{M}|\mathbb{E}.2} = \{\mathbf{M} \in \Omega_{\mathbf{M}} : (\mathbf{M} - \mathbf{y})'(\mathbf{M} - \mathbf{y}) = U/U_s\}. \quad (33)$$

One can use any preferred distribution as the prior distribution for S^2 . One computationally convenient choice is the Inv- χ^2 distribution with ν degrees of freedom. By independence of U and U_s , the ratio $U/U_s \sim \mu_{\mathbb{E}.2} = (k/\nu) F_{(k,\nu)}$, the scaled F-distribution with scaling factor k/ν and degrees of freedom k and ν . All calculations in Section 3 remain the same after replacing $\mu_{\mathbb{E}}$ with $\mu_{\mathbb{E}.2}$. Posterior inference based on other prior specifications follows the same logic.

5. Discussion

Several useful generalizations to the proposed model are to be explored in future studies.

Elliptical distributions. The sampling distribution employed by the i.i.d. model is a multivariate normal distribution with identity covariance matrix, which is a special kind of elliptical distribution. The vacuous orientation assumption can be generalized to other families of elliptical distributions with covariance Σ , for which the probability distribution is a function of the configuration $(\mathbf{M} - \mathbf{Y})' \Sigma^{-1} (\mathbf{M} - \mathbf{Y})$. Of particular interest are the multivariate t and Laplace distributions, suitable for measurement errors are expected to be both heavy-tailed and dependent.

Multivariate regression. The current model setting can be generalized to accommodate covariate information. Specifically, let $\mathbf{M} = \mathbf{X}\beta$, where \mathbf{X} is a k by p design matrix and β a p -vector of coefficients. A vacuous orientation assumption on the observation errors now induces posterior inference about β in a similarly vacuous manner. Note that the current model produces meaningful simultaneous posterior inference when the number of hypotheses far exceeds the number of unknown parameters. In the same way, the regression model with the vacuous orientation assumption can deliver “large p , small k ” inference, that is, when the regression model itself is underdetermined.

Finer variance decomposition. Lastly, the vacuous orientation assumption itself is also subject to extension. Notice that the configuration of k i.i.d. normal error terms can be decomposed into a collection of up to k variance components, such that they’re all independent among each other and with degrees of freedom that sum to k . In fact, both the i.i.d. model and the vacuous orientation model are special cases of such variance decomposition. The former is the trivial decomposition into k components, each of degrees of freedom 1. The latter is the canonical decomposition into two components, configuration and orientation, of degrees of freedom 1 and $k - 1$ respectively. One can imagine generalizing the vacuous assumption onto other variance decompositions to express degrees of ignorance over the dependence structure among errors.

Acknowledgments

The author thanks Professor Arthur P. Dempster for inspiring this work and for valuable discussions, as well as three anonymous reviewers for helpful suggestions.

References

- [1] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

- [2] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The annals of statistics*, 29(4):1165–1188, 2001.
- [3] James O Berger. An overview of robust Bayesian analysis. *Test*, 3(1):5–124, 1994.
- [4] Philip A Dawid and Mervyn Stone. The functional-model basis of fiducial inference. *The Annals of Statistics*, 10(4):1054–1067, 1982.
- [5] Arthur P Dempster. Upper and lower probabilities induced by a multivalued mapping. *The Annals of Mathematical Statistics*, 38(2):325–339, 1967.
- [6] Arthur P Dempster. *Elements of continuous multivariate analysis*. Addison-Wesley, 1969.
- [7] Arthur P Dempster. The Dempster-Shafer calculus for statisticians. *International Journal of Approximate Reasoning*, 48(2):365–377, 2008.
- [8] Arthur P Dempster. Statistical inference from a dempster–shafer perspective. *Past, Present, and Future of Statistical Science*, pages 275–288, 2014.
- [9] Olive Jean Dunn. Multiple comparisons among means. *Journal of the American statistical association*, 56(293):52–64, 1961.
- [10] Donald AS Fraser. *Structural inference*. John Wiley & Sons, New York, NY, 1968.
- [11] Jan Hannig, Hari Iyer, Randy CS Lai, and Thomas CM Lee. Generalized fiducial inference: A review and new results. *Journal of the American Statistical Association*, 111(515):1346–1361, 2016.
- [12] Jeremie Houssineau. Parameter estimation with a class of outer probability measures. *arXiv preprint arXiv:1801.00569*, 2018.
- [13] Chung Tung Augustine Kong. *Multivariate Belief Functions and Graphical Models*. PhD thesis, Harvard University, 1987.
- [14] Ryan Martin and Chuanhai Liu. *Inferential Models: reasoning with uncertainty*. CRC Press, Boca Raton, FL, 2015.
- [15] Glenn Shafer. *A mathematical theory of evidence*. Princeton University Press, Princeton, NJ, 1976.