Exploring the Accuracy of Joint-Distribution Approximations Given Partial Information

Luis V. Montiel\textsuperscript{a} and J. Eric Bickel\textsuperscript{b}

\textsuperscript{a}Department of Operations Research and Industrial Engineering, ITAM, Río Hondo 1, Mexico City, Mexico; \textsuperscript{b}Graduate Program in Operations Research & Industrial Engineering, The University of Texas at Austin, 1 University Station C2200, Austin, TX, USA

\textbf{ABSTRACT}
We test the accuracy of various methods for approximating under-specified joint probability distributions. In particular, we examine the maximum entropy and the analytic center approximations, and we introduce three methods for approximating a discrete joint probability distribution given partial probabilistic information. Our results suggest that recently proposed approximations and our new approximations more accurately represent the possible uncertainty models than do previous models such as maximum entropy.

\textbf{KEYWORDS}
Accuracy; joint probability distribution approximations; JDSIM; maximum entropy; analytic center

1. Introduction

In many engineering-economic applications involving uncertainty, the economist or analyst may be unable to fully specify the underlying probabilistic model. This under-specification can be due to a lack of resources, data, or the inability of experts to assess the required marginal and conditional probabilities. For example, consider a problem presented by Bickel and Smith (2006) (hereafter Bickel and Smith), which was inspired by an actual project with a major oil & gas company.

The company was developing its exploration strategy within a major oil field. They were considering six prospects (new drilling locations) that may produce oil (i.e., are “wet”) or may be dry holes (i.e., no oil is present). The geologists working for the company believed that these prospects were geologically dependent. By this, they meant that if they found oil at one location they would assign a higher probability to finding oil at another, nearby, location. The company wanted to use this information to develop the optimal drilling sequence. As Bickel and Smith show, such an optimal drilling sequence can be developed using dynamic programming if one is able to specify the joint probability distribution over all possible outcomes for the six wells.

If we assume that each well is either wet or dry, there are $2^6$ (or 64) potential outcomes. Thus, we must specify 64 joint probabilities. In practice, such an assessment is decomposed into a series of marginal and conditional assessments. For example, one might begin by assessing the probability that well 1 contains hydrocarbons, which...
requires the assessment of only one probability. Next, we would assess the probability that well 2 contains hydrocarbons, conditional on the outcome at well 1. This conditional assessment requires that we assess two probabilities: the chance that well 2 is wet given that well 1 is wet and the chance that well 2 is wet given that well 1 is dry. Continuing in this way, we would find that wells 3, 4, 5, and 6 require four, eight, sixteen, and thirty-two assessments, respectively. This is a daunting task. Even more challenging is the fact that many of these assessments would be heavily conditioned. In short, while this problem is theoretically straightforward, it is impossible to specify the required joint probability distribution in practice.

Bickel and Smith addressed this problem by assessing the marginal probability of finding hydrocarbons for each well and then each pairwise conditional probability, as shown in Table 1. For example, the marginal probability of finding hydrocarbons at well 1 was 0.35. The marginal probability of finding hydrocarbons at well 2 was 0.49. However, if hydrocarbons were discovered at well 1, the probability of finding them at well 2 increased from 0.49 to 0.59.

The assessments in Table 1 constrain, but do not fully specify the required joint distribution of all possible well outcomes. This is true because we have not specified any level of conditioning other than pairwise. For example, we have not specified the probability of finding hydrocarbons at well 3 given that well 1 was dry and well 2 was wet.

We may restate the information given in Table 1 as pairwise joint probabilities. These values are shown in Table 2. For example, the probability that wells 1 and 2 are both wet is 0.35*0.59 or 0.2065.

Formally, the assessments in Tables 1 and 2, along with the requirement that the joint probability distribution sums to 1, define a polytope $A\pi = b$ that contains all feasible joint distributions. We refer to this polytope as the Truth Set, because it contains the true distribution, and define it as $T = \{ \pi : A\pi = b, \pi \geq 0 \}$. Any $\pi$ that is a member of $T$ is a feasible joint distribution consistent with the assessed information.

Table 1. Bickel and Smith Pairwise Conditional Assessments of Well Success.

| i \ j | 1   | 2   | 3    | 4    | 5    | 6    | Marginal p_i |
|-------|-----|-----|------|------|------|------|--------------|
| 1     | 0.59| 0.63| 0.83 | 0.39 | 0.31 | 0.35 |              |
| 2     | 0.65| 0.83 | 0.55 | 0.24 | 0.49 |      |              |
| 3     | 0.83 | 0.42 | 0.31 | 0.53 |      |      |              |
| 4     | 0.33 | 0.18 | 0.83 |      |      |      |              |
| 5     | 0.26 | 0.33 |      |      |      |      |              |
| 6     | 0.18 |      |      |      |      |      |              |

Table 2. Bickel and Smith Pairwise Joint Assessments of Well Success.

| i \ j | 1   | 2   | 3    | 4    | 5    | 6    |
|-------|-----|-----|------|------|------|------|
| 1     | 0.2065 | 0.2205 | 0.2905 | 0.1365 | 0.1085 |
| 2     | 0.3185 | 0.4067 | 0.2695 | 0.1176 |
| 3     | 0.4399 | 0.2226 | 0.1643 |
| 4     | 0.2739 | 0.1494 |
| 5     |      | 0.0858 |
| 6     |      |      |      |      |      |      |

Table 1. Bickel and Smith Pairwise Conditional Assessments of Well Success.

Direct conditional assessments $p(w_j = 1|w_i = 1)$

Table 2. Bickel and Smith Pairwise Joint Assessments of Well Success.

$p(w_j = 1, w_i = 1)$
In our example, \( \mathbf{A} \) is a 22-by-64 matrix of zeros and ones (see Appendix A) that encodes the 22 constraints (the requirement that the probabilities sum to 1, 6 marginal constraints, and 15 pairwise constraints) and the 64 joint outcomes. The vector \( \mathbf{b} \) encodes the specific information of our 22 constraints described above, and \( \pi \) is the joint probability distribution that matches all our information requirements.

\[
\mathbf{b} = \begin{bmatrix}
1, 0.35, 0.49, 0.53, 0.83, 0.33, 0.18, 0.2065, 0.2205, 0.2905, 0.1365, 0.1085,
0.3185, 0.4067, 0.2695, 0.1176, 0.4399, 0.2226, 0.1643, 0.2739, 0.1494, 0.0858
\end{bmatrix}.
\]

Since \( \mathbb{T} \) contains an infinite number of feasible joint probability distributions, the problem now is how to select a single distribution under which to evaluate the decision? Bickel and Smith choose the distribution that has the highest entropy, which is known as the maximum-entropy (ME) distribution. The maximum entropy method, introduced by Jaynes (1957), has been used extensively in the decision analysis literature and elsewhere for nearly six decades. Speaking loosely, the ME distribution is the “flattest” distribution (i.e., closest to uniform), while still honoring any information we do have (e.g., marginal and pairwise probabilities.)

Later, Montiel and Bickel (2013a) explored solutions for the same problem using different approximation methods. They found that, not surprisingly, differing methods resulted in different optimal drilling policies. In addition, however, they suggested that some approximations may be more robust or better representations of the set of all feasible distributions.

The previously cited papers examined the accuracy of discrete joint-distribution approximations within the context of a single example. In this paper, we undertake a systematic study of this question. Specifically, the question we address in this paper is how representative is a single, maximum-entropy, distribution of the entire set of feasible distributions? In addition, are there other distributions within \( \mathbb{T} \) that are, in a sense to be formally defined in the next section, better representations of the entire set of feasible distributions? Specifically, we systematically evaluate the accuracy of ME and four other approximation methods: the analytic center (AC), the Chebyshev center (CC), the maximum-volume inscribed-ellipsoid center (MV), and sample average approximation (SA).

This paper is organized as follows. §2 presents the five joint-distribution approximations considered in this paper. §3 provides the basis to measure the accuracy of an approximation. §4 presents and analyzes the results of four experiments that test the accuracy of our approximations. Finally, §5 concludes and discusses future research directions.

2. Joint-Distribution Approximations

We consider five approximation methods: maximum entropy (ME), analytic center (AC), Chebyshev center (CC), maximum-volume inscribed-ellipsoid center (MV), and sample average approximation (SA). These five methods have been extensively used in the literature in areas such as optimization, decision analysis, simulation, and statistics. However, except for ME, they have not been generally used to approximate joint-probability distributions under partial information. In this section, we present the original models used to calculate all five of them, and in a later section we study how representative is each method with respect to the Truth Set \( \mathbb{T} \). By “representative” we mean how
“close” an approximate distribution is to all other distributions within the Truth Set. We will measure closeness in several different ways, all of which are explained in §3.2.

2.1. Maximum Entropy

Jaynes (1957) proposed using the distribution in $\mathbb{T}$ that has the highest entropy, which we denote $\pi^{\text{ME}}$. Formally, $\pi^{\text{ME}}$ is defined as:

$$\pi^{\text{ME}} = \underset{\pi \in \{\pi_1, \ldots, \pi_n\}}{\text{arg max}} \sum_{i=1}^{n} \pi_i \ln \pi_i,$$

subject to $A\pi = b$, $\pi \geq 0$,

where $n$ is the total number of joint events in $\pi^{\text{ME}}$.

2.2. Analytic Center

The AC, denoted as $\pi^{\text{AC}}$, has been mainly used to initialize interior point algorithms. The formulation of AC is as follows:

$$\pi^{\text{AC}} = \underset{\pi \in \{\pi_1, \ldots, \pi_n\}}{\text{arg max}} \sum_{i=1}^{n} \ln \pi_i,$$

subject to $A\pi = b$, $\pi \geq 0$.

AC is not invariant with respect to the polytope representation. For example, the addition of a redundant constraint would result in a different AC (Ye, 1997). In this paper, we assume that the expert has not provided any redundant constraints.

2.3. Chebyshev Center

The CC is the center of the largest hypersphere in the positive orthant. We restrict this center to belong to the Truth Set, hence $\pi^{\text{CC}} \in \mathbb{T}$. Equation (3) presents an adaptation of the original problem defined by Boyd and Vandenberghe (2004) in full-dimensional polytopes (dimension equal to $n$), where $r$ represents the radius of the hypersphere and the symbol 1 is a vector of all ones.

$$\pi^{\text{CC}} = \underset{\pi \in \{\pi_1, \ldots, \pi_n\}}{\text{arg max}} r,$$

subject to $A\pi = b$, $\pi \geq r \cdot 1$.

Equation (3) determines the maximum $r$ and represents the minimum distance from $\pi^{\text{CC}}$ to the boundary of the positive orthant. Hence, $\pi^{\text{CC}}$ is the distribution in $\mathbb{T}$ for which the smallest probability $\pi_i$ is maximum, which can be solved efficiently using linear programming.

2.4. Maximum-Volume Inscribed-Ellipsoid Center

The maximum-volume inscribed-ellipsoid center (MV), first studied by John (1948), is the center of the largest ellipsoid in the interior of $\mathbb{T}$, denoted as $\pi^{\text{MV}}$. MV is part of
The "maxdet" family that can be solved by positive semi-definite programming formulations as described in Vandenberghe et al. (1998).

Consider the null space of matrix $A$, denoted as $\text{Null}(A) = \{z_i \in \mathbb{R}^n | Az_i = 0 \forall i\}$. Taking orthogonal vectors $z_i \in \text{Null}(A)$ where $z_i \neq 0$, we can define the matrix $Z = [z_1 z_2 \ldots z_k]$ as the orthogonal basis of $\text{Null}(A)$ of dimension $k$. Then, the transformation $\pi = Zy + \hat{\pi}$, where $\hat{\pi}$ is any predefined solution that satisfies $A\hat{\pi} = b$ and where $y \in \mathbb{R}^k$ is a free variable, provides an efficient implementation of the problem shown in Equation (4). The optimal solution in terms of the vector $y^* \in \mathbb{R}^k$ and the matrix $E^* \in \mathbb{R}^{k \times k}$ can be used to recover the $\pi^{MV}$ approximation.

$$y^* = \arg \max_{y \in \mathbb{R}^k} \log \det E,$$

$$s.t. -Zy + G(E) \leq \hat{\pi},$$

$$E \succeq 0,$$

$$G(E) = (||Z_1E||_2, \ldots, ||Z_nE||_2)^T,$$

where $Z_i \in \mathbb{R}^k$ is the $i^{th}$ row of $Z$. Finally, we recover $\pi^{MV}$ using the previous transformation

$$\pi^{MV} = Zy^* + \hat{\pi}.$$  

The objective function maximizes the determinant of $E$, i.e., it maximizes the volume of the ellipsoid inscribed in $\mathbb{T}$ with center at $\pi^{MV}$. A geometrical interpretation of $E$ is given by the eigenvectors and eigenvalues of the matrix, the first describe the position of the principal axes of the ellipsoid, and the second describe its length. For example, when $E$ is the identity matrix the ellipsoid reduces to the unit sphere. Because the number of elements in $E$ increases exponentially with the size of $\text{Null}(A)$, MV has longer running times than all other approximations. For example, cases with seven binary RVs with marginal and pairwise information could take up to ten minutes using CVX on Matlab. Please see Appendix B for an illustrative implementation of the MV. For further technical details, please refer to John (1948).

### 2.5. Sample-Average Approximation

Finally, suppose we were able to uniformly sample $\mathbb{T}$. We could average these samples (element-wise) to create the sample average approximation (SA). In the limit as $N \to \infty$, $\pi^{SA}$ will converge to the center of mass of $\mathbb{T}$. Formally, if we sample $N$ points from $\mathbb{T}$, the SA is:

$$\pi^{SA} = \frac{1}{N} \sum_{i=1}^{N} \pi_i.$$  

### 3. Quantifying Approximation Accuracy

This section describes a general procedure to measure the accuracy of an approximate probability mass function (pmf) in terms of the geometry of $\mathbb{T}$. The procedure requires two main steps: (1) characterizing the geometry of the Truth Set by creating a discrete collection of pmfs that replicate the volume of $\mathbb{T}$, and (2) measuring the divergence from the sampled points within $\mathbb{T}$ to the approximate pmf. The distribution of this
divergence and its statistics, such as the mean, codify the accuracy of the approximation and is, therefore, a measure of how representative an approximation is with respect to $T$. Hence, we define accuracy as the sample average of the divergence measures. The following elaborates on both of these steps, formally defines how we measure accuracy, and presents a procedure to calculate it.

### 3.1. Characterizing and Sampling From $T$

As we shown in the Introduction, we start by gathering all the known partial information relevant to the structure of the joint distribution. In the decision analysis literature, it is common to assess marginal probabilities and perhaps pairwise probabilities. However, it is possible to consider higher order assessments, moments of the distribution, Spearman rank correlation information, and any other information that can be coded using linear equations. These constraints define the Truth Set $T = \{\pi : A\pi = b, \pi \succeq 0\}$.

Once the $T$ is specified, we create a collection of samples uniformly distributed with respect to the relative interior of $T$. We achieve this by means of the joint distribution simulation approach (JDSIM) proposed by Montiel (2012) and Montiel and Bickel (2013b). JDSIM is based on the Hit-and-Run sampler (HR) (Smith, 1984), which is the fastest known algorithm to sample the interior of an arbitrary polytope (Lovasz, 1998).

Our uniform sampling procedure implies that every distribution contained within $T$ is equally likely. Although this may not be the case in a particular setting, it is a reasonable starting point for exploring the accuracy of the various approximations considered in this paper. Moreover, uniform sampling provides a clear benchmark to which others could compare and build upon our work.

The collection of samples created with JDSIM can be used together with different accuracy measures to study the relationship between any particular approximation and points within $T$. For this study, the appropriate sample size was determined according to Figure 4 in Montiel and Bickel (2013b).

### 3.2. Accuracy Measures

After generating a collection of samples from $T$, we need a measure of divergence between the samples and the approximation of interest. We consider the three Bregman divergence measures (Bregman, 1967), also considered by Montiel and Bickel (2013a). The Bregman measures, or measures of accuracy, are defined in Equations (7a)-(7c), where $\pi$ is a sampled pmf, $\pi_i$ is the $i$-th element of $\pi$, $n$ is the number of elements in $\pi$, and $\pi^*$ is an approximate pmf, such as $\pi^{ME}$, $\pi^{AC}$, $\pi^{SA}$, $\pi^{CC}$, or $\pi^{MV}$.

\[
L^\infty_n(\pi, \pi^*) = \max\{|\pi_1 - \pi^*_1|, \ldots, |\pi_n - \pi^*_n|\}, \quad (7a)
\]
\[
L^2_n(\pi, \pi^*) = \left[\sum_{i=1}^{n} (\pi_i - \pi^*_i)^2\right]^{1/2}, \quad (7b)
\]
\[
KL_n(\pi || \pi^*) = \sum_{i=1}^{n} \pi_i \log_2 \left(\frac{\pi_i}{\pi^*_i}\right). \quad (7c)
\]

The maximum absolute difference ($L^\infty$-norm) between two pmfs is the maximum of all the differences between probabilities that describe a specific joint event. The
Euclidean distance (L²-norm) is the straight-line distance in n-dimensional space between the two pmfs. Finally, the KL divergence measures the relative entropy between a pmf and a reference distribution, which we take to be one of the approximate pmfs. In this paper, we use the base-2 logarithm.

3.3. The Accuracy of an Approximation

We quantify the accuracy of each approximation by measuring the divergence between the approximation and all other feasible distributions (i.e., all points within $\mathcal{T}$), and use it to determine how representative an approximation is with respect to $\mathcal{T}$. It would be infeasible to analytically obtain these measurements, given the numerous cases we consider here. However, they can be approximated using JDSIM. Specifically, we will measure the divergence between each sampled point and the approximation under consideration and will compute the sample mean and standard deviation of divergence for each approximation across the sampled points. We will focus on the mean and refer to it as “accuracy.” Formally, we define accuracy as shown in Definition 1.

**Definition 1.** Considering a collection of samples $\pi^i$ indexed by $i = 1, \ldots, N$; and a Bregman divergence measure $D(\pi^i, \pi^*)$, we define accuracy as:

$$\sum_{i=1}^{N} \frac{D(\pi^i, \pi^*)}{N}$$  

(8)

Our general approach can be summarized as follows:

- Define the Truth Set $\mathcal{T}$ as a system $A\pi = b$, $\pi \geq 0$.
- Determine the approximation $\pi^*$ in $\mathcal{T}$.
- Create a set of samples $\pi^i$ for $i = 1, 2, \ldots, N$ by sampling $\mathcal{T}$.
- Summarize the performance of approximation $\pi^*$ by computing the mean divergence between it and the $N$ sampled points.

4. Analysis

Quantifying the accuracy of each approximation method in general would require solving a high dimensional integral over a domain of integration defined by the vertices of $\mathcal{T}$. This approach is infeasible because, as noted by McMullen (1970) and Schmidt and Mattheiss (1977), the number of vertices in $\mathcal{T}$ can be enormous. Therefore, we consider a series of experiments that provide insight into the performance of the approximations, which we then extrapolate to more general settings.

Most of these experiments consider joint distributions comprised of three, four, or five binary/Bernoulli random variables (RVs). We design the experiments to test the accuracy under situations that we believe occur frequently. In particular, we analyze four general cases where marginal probability assessments and pairwise rank correlations are assessed. The first two cases consider identical RVs with equal pairwise rank correlations; the last two cases are defined using a random procedure to create information that is arbitrary, but probabilistically consistent. The complete list of experiments is summarized below.
1. Identical, but varying, marginal distributions.
2. Identical marginal distributions with identical, but varying, pairwise rank correlations.
3. Differing marginal distributions that are randomly selected.
4. Differing marginal distributions and pairwise correlations, both of which are randomly selected.

This paper presents a sample of the full experimental results. The online supplement contains the complete set of experiments.

We also design additional experiments to test the effect of the number of constraints, the accuracy of identical RVs with known first moments, and the accuracy of identical RVs with known first and second moments, all of which can be found in the online supplement. These experiments are summarized as follows:

1. Joint distributions with varying number of binary RVs and uniform constraints (see online supplement §1).
2. Identical, but varying, first moments (see online supplement §6).
3. Identical, but varying, first moments and covariance (see online supplement §7).

For simplicity, all calculations have been performed for joint distributions with binary marginal variables. However, experimental results suggest that the observations hold for marginal variables with three, four, or more outcomes.

As discussed above, the experiments consider three measures of accuracy ($L^1$, $L^2$, and KL). However, only one set of full results is presented; the results for the first two measures are relegated to the online supplement. To compute our accuracy measures across the five approximations ($\pi_{ME}$, $\pi_{AC}$, $\pi_{CC}$, $\pi_{MV}$, and $\pi_{SA}$), we use a sample of one million joint distributions. $\pi_{SA}$ was calculated using a collection of one million samples different from the one used to calculate accuracy. This assures that SA accounts for the fact that the approximation is sample dependent.

### 4.1. Varying Marginal Assessments

In this section, we assume that each binary RV has the same marginal distribution $\rho$. Specifically, we assume that $b = [1, \rho, \rho, ..., \rho]^T$ for $\rho = 0.50, 0.51, ..., 0.99$ (50 cases in total). Figure 1 presents the accuracy results for the proposed approximations in the case of three and five RVs as a function of $\rho$. (A complete set of results is available in §2 in the online supplement.) Each point in Figure 1 is the mean of one million accuracy measures. Thus, each subfigure is based on 50 million randomly selected joint distributions.

The five approximations have the same accuracy (or error) when $b$ is equal to 0.5 or 1.0. This behavior is explained by the geometry of $T$ as follows. If $b = 0.5$, the constraints define a symmetric Truth Set with center of mass equal to the uniform distribution. In this case, AC, CC, MV, ME, and SA coincide with this uniform distribution. If $b = 1.0$, the Truth Set reduces to a single point where all approximations coincide due to feasibility.

Among the approximations considered, SA, MV, and AC exhibit the best performance, in that order. However, the differences are almost indistinguishable. CC exhibits
lower accuracy than SA, MV, or AC over almost the entire region. ME performs the worst. The accuracy results for all the approximations are consistent under all three accuracy measures. This consistency holds in the rest of the experiments presented in the paper. Therefore, we limit the remainder of the results to the KL divergence and include all other results in the online supplement.

4.2. Identical Marginal Assessments with Varying Rank Correlations

We now add a pairwise dependence structure to $T$ by specifying pairwise correlations, $\rho$. The linear equations to implement these correlation-based constraints, are rather
cumbersome and we do not present them for that reason. The interested reader can see Appendix C for an example of the procedure or refer to Montiel and Bickel (2013b) for a detailed explanation.

Figures 2 and 3 show the results for Truth Sets with three and five binary RVs. The complete set of results for three, four, and five binary/Bernoulli RVs can be found in Figures 4, 5, and 6 in the online supplement.

This experiment quantifies the effects of two parameters: $p$ and $q$. The parameter $p$ indicates the marginal probability of the binary variables (as in §4.1), whereas the parameter $q$ indicates the Spearman correlation among any two RVs. In the context of $T = \{\pi | A \pi = b\}$, the parameters $p$ and $q$ correspond to elements of $b = [p, \ldots, p, \rho, \ldots, \rho]$, where the matrix $A$ is defined by marginal and rank correlation constraints. As before, every point in each figure is the average of one million accuracy measures.

Figure 2 contains four subfigures, one for each marginal probability $p = 0.5, 0.6, 0.7, 0.8$. For $p = 0.5$, the domain of $\rho$ ranges from 0 to $\approx 0.75$ as determined by Mackenzie (1994), whereas for $p = 0.8$, the domain of $\rho$ ranges from 0 to $\approx 0.48$. The $n$ variables and $m$ constraints define the shape and dimension of $T$. In the case of three binary RVs, we have $n = 8$ ($2^3$ joint events) and $m = 7$ (three marginal distribution constraints, three rank correlation constraints, and one necessary constraint). Hence, the Truth Set is a line that intersects the positive orthant, with length determined by the vector $b$. For example, in Figure 2(b), the Truth Set length increases as $\rho$ goes from 0 to approximately 0.12, increasing the error of all the approximations, and decreases for
\( \rho \gtrsim 0.12 \), decreasing also the observed error. In fact, the error goes to zero for \( \rho \approx 0.71 \), where the Truth Set becomes a singleton.

Figure 3 is equivalent to Figure 2, but considers five binary RVs instead of three. Here, CC is more accurate than ME in most scenarios. In particular, if \( \rho \) is higher than \( \approx 0.14 \), CC is almost as good as AC, MV, or SA. Otherwise, its accuracy worsens, providing an unreliable model. Here again ME consistently performs poorly. This behavior is also explained by the geometric properties of \( T \), although this might not be intuitive in \( n - m = 16 \) dimensions. We noted earlier that CC maximizes the minimum joint event probability. As a result, when \( T \) contains a line that forms a small angle \( \theta \) with any unit vector \( e_i \), for some \( i \), the CC approximation skews toward one side of the Truth Set. Hence, for \( \rho > 0.5 \), as \( \rho \) increases, \( T \) changes position, thereby reducing the angle \( \theta \) and increasing the error measure. The position of \( T \) continues to change until the angle \( \theta \) starts increasing, which decreases the error measure. Finally, as \( \rho \) approaches 0.14 (approximately), the volume of the relative interior of \( T \) starts decreasing, thereby improving the accuracy of CC as well as some of the other approximations.

These results show that SA, MV, and AC outperform ME and CC. SA, MV, and AC perform very similarly, although AC’s performance degrades close to \( \rho \approx 0.0 \). In this experiment, we found that MV and SA have a better accuracy than all other approximations, but require longer running times. These results and those presented in §3 of the online supplement suggest that AC is a very accurate and practical representation of \( T \).
provided that the marginal distributions are identical and the dependence interactions are the same for all pairs of RVs. However, if additional precision is required, SA and MV can provide better results at the expense of larger running times.

4.3. Randomly Selected Marginal Distributions

In the previous two sections, we tested Truth Sets where all RVs had the same marginals and pairwise correlations. In this and the next section, we relax this requirement by allowing each RV to have a randomly chosen marginal distribution. Specifically, we now use a two-step simulation procedure:

**Step 1:** We generate 1,000 random constraint vectors $b$, each of which specifies a Truth Set for the marginal probabilities of the RVs $V_i \forall i = 1, ..., M$, where $M = 3, 4,$ or

![Figure 4. Accuracy of ME, CC, AC, and MV compared to SA in sets with three RVs having arbitrary marginals. The x-axes index the $T$s from 1 to 1,000 by the sort values of SA (black-dot), and the y-axes show the error measure using KL divergence. In each subfigure, a second approximation is shown (gray-dot), where a line between approximations indicates the difference in accuracy.](image)
5 depending on how many RVs are being considered. Each constraint vector was generated by selecting a uniform random number in the range (0,1) for each element of $b$.

**Step 2:** As before, for each Truth Set, we generate a collection of one million joint distributions, using JDSIM. We compute the average divergence from each sample point to each of our five approximations, using our accuracy measures.

Figures 4 and 5 display the accuracy results for three and five RVs, respectively, using KL divergence. The complete set of results for joint distributions with three, four, or five RVs and all measures of accuracy appear in §4 of the online supplement.

Each subfigure uses the accuracy of SA, shown as the increasing solid black line, as a point of reference. The results are by ascending SA error. Each subfigure includes a cloud of gray dots corresponding to the approximations. The light gray vertical lines highlight the difference in the errors of the two featured approximations. The dots above (below)
SA show that on average for the simulated Truth Set, the listed approximation (e.g., ME) was farther from (closer to) the one million sampled points than was SA. For example, in the vast majority of randomly generated Truth Sets, ME was on average further from all other points in the set than was SA. In fact, SA outperformed ME, by this measure, in about 940 of the 1,000 randomly generated Truth Sets (Figure 4(a)). After considering all the results (see online supplement), we concluded that SA outperformed ME about 91% to 94% of the time, depending on the measure used and the dimension of $T$ (noting that SA was calculated independently from the measuring sample).

Comparing Figure 4(b) to 4(c), and Figure 5(b) to 5(c), indicates that AC outperformed CC by a small margin. Furthermore, both AC and CC exhibited a considerably

---

**Figure 6.** Accuracy of ME, CC, AC, and MV compared to SA in sets with three RVs having arbitrary marginals and rank correlation. The $x$-axes index the $T$s from 1 to 1,000 by the sort values of SA (black dot), and the $y$-axes show the error measure using KL divergence. In each subfigure, a second approximation is shown (gray dot), where a line between approximations indicates the difference in accuracy.
lower error than did ME. Finally, Figures 4(d) and 5(d) show that MV and SA were usually very close, except for few special cases where the differences became exceptionally large. These cases correspond to sets $T_i$ with extreme marginal constraints ($P(X=1)\approx0.98$). These Truth Sets tend to have elongated corners, similar to a needle, that increases the mixing time of the simulation procedure.

Table 3 presents a summary of these results, including those presented in the online supplement. Here we show the sample mean and standard deviation (SD) of the average error (using one million samples), where the mean and the SD are calculated using 1,000 Truth Sets for each measure of accuracy and each approximation. In this analysis, we considered three, four, and five RVs.

As shown in the AC and MV columns of Table 3, these two approximations have consistently smaller sample mean errors. Hence, we consider them to have the highest accuracy among our approximations. For example, Table 3 for three RVs and $L_2$-norm indicates that the sample mean errors of AC and MV are 0.1226 and 0.1218, respectively, which clearly outperform the sample mean error of ME (0.1383) and improve slightly over the sample mean error of CC (0.1228). As we increase the number of RVs, AC performs better than MV (0.1256 to 0.1277, respectively, for five RVs and $L_2$ norm). We surmise that this behavior is related to the complexity of the polytopes in higher dimensions and to the decrease in accuracy of the MV approximation as the problem size is increased. These results are statistically significant for a sample of size $N=1,000$, a sample standard deviation $\sigma=0.0016$, and a confidence interval error of $\alpha=0.999$.

In general, the sample mean errors for CC and SA are higher than for AC and MV. However, in the case of SA, we need to consider that as the number of variables increases, the sampling error also increases. Hence, for a constant sample size of one million $n^T \in \mathbb{T}$, the SA results are less precise for Truth Sets with five RVs, which is consistent with the aggregate information presented on the last column of Tables 3 and 4. Additionally, the constant sample size can also have an important effect in some random polytopes, as shown in Figures 4(d) and 5(d), where the mixing time require larger

| Table 3. Results summary for Truth Sets with randomly selected marginal values. |
|-----------------|---|---|---|---|---|
|                 | ME | CC | AC | MV | SA |
| **Three Variables** |    |    |    |    |    |
| Sample Mean  | $L_\infty$ | 0.0821 | 0.0728 | 0.0724 | 0.0720 | 0.0722 |
|               | $L_2$   | 0.1383 | 0.1228 | 0.1226 | 0.1218 | 0.1222 |
|               | $KL$    | 0.1754 | 0.1402 | 0.1385 | 0.1369 | 0.1377 |
| Sample SD   | $L_\infty$ | 0.0009 | 0.0010 | 0.0010 | 0.0010 | 0.0010 |
|               | $L_2$   | 0.0015 | 0.0016 | 0.0016 | 0.0016 | 0.0016 |
|               | $KL$    | 0.0021 | 0.0023 | 0.0023 | 0.0023 | 0.0023 |
| **Four Variables** |    |    |    |    |    |
| Sample Mean  | $L_\infty$ | 0.0953 | 0.0782 | 0.0772 | 0.0778 | 0.0782 |
|               | $L_2$   | 0.1534 | 0.1307 | 0.1297 | 0.1305 | 0.1312 |
|               | $KL$    | 0.2645 | 0.2037 | 0.1992 | 0.2004 | 0.2045 |
| Sample SD   | $L_\infty$ | 0.0007 | 0.0008 | 0.0008 | 0.0008 | 0.0009 |
|               | $L_2$   | 0.0012 | 0.0014 | 0.0014 | 0.0014 | 0.0015 |
|               | $KL$    | 0.0025 | 0.0028 | 0.0028 | 0.0028 | 0.0030 |
| **Five Variables** |    |    |    |    |    |
| Sample Mean  | $L_\infty$ | 0.0986 | 0.0749 | 0.0736 | 0.0751 | 0.0774 |
|               | $L_2$   | 0.1542 | 0.1278 | 0.1256 | 0.1277 | 0.1317 |
|               | $KL$    | 0.3429 | 0.2556 | 0.2478 | 0.2516 | 0.2681 |
| Sample SD   | $L_\infty$ | 0.0008 | 0.0007 | 0.0007 | 0.0007 | 0.0009 |
|               | $L_2$   | 0.0010 | 0.0012 | 0.0012 | 0.0012 | 0.0014 |
|               | $KL$    | 0.0028 | 0.0030 | 0.0030 | 0.0030 | 0.0035 |
This effect needs to be considered when deriving conclusions from Tables 3 and 4 on relation to the accuracy of SA.

The main two results of this section concern the accuracy of ME and AC. The first result shows that ME is less accurate than the other approximations. The second result suggests that as the number of RVs increases, AC outperforms all other approximations, including MV.

### 4.4. Randomly Selected Marginal Distributions and Rank Correlations

Section 4.3 considered random marginal distributions with a free dependence structure. In this section, we add a randomly generated dependence structure. This procedure starts in the same way that the procedure described in §4.3. Once the marginal distributions are chosen, the intersection of the respective constraints and the Truth Set generates a proper subset of $T$, where any line that intersects this proper subset will have joint distributions with the same marginal probabilities but with different correlation structures. To select a random correlation structure, we randomly select a line among all the lines that intersect the proper subset that include the joint distribution that assumes all RVs are independent ($\rho_{ij} = 0 \forall i \neq j$). A random point in this line will provide a joint distribution that goes from a zero Spearman rank-correlation matrix $\rho_{ij} = 0$ to the highest Spearman rank-correlation matrix structure that is consistent with previous information. For an illustrative example of the implementation of marginal and correlation constraints see Appendix §C. The algorithm is as follows:

**Step 1:** Generate 1,000 random constraint vectors $b$, each of which specifies the marginal probabilities of the RVs $V_i \forall i = 1, ..., M$, where $M = 3, 4, or 5$ depending on how many RVs are being considered. Each constraint vector was generated by selecting a uniform random number in the range $(0,1)$ for each element of $b$.

**Step 2:** For each marginal probability vector $b$ created in Step 1, recover the independent distribution ($\pi^I$) and randomly select a non-trivial direction orthogonal to the coefficients of the marginal constraints $Ad = 0$. Hence, for direction $d$ and scalar $\alpha$.

#### Table 4. Results summary for Truth Sets with random selected marginal and rank correlation values.

|                  | ME    | CC    | AC    | MV    | SA    |
|------------------|-------|-------|-------|-------|-------|
| **Three Variables** |       |       |       |       |       |
| Sample Mean      | $L^\infty$ | 0.0071 | 0.0065 | 0.0066 | 0.0065 | 0.0065 |
|                  | $L^2$  | 0.0200 | 0.0183 | 0.0186 | 0.0183 | 0.0183 |
|                  | $KL$   | 0.0180 | 0.0138 | 0.0144 | 0.0138 | 0.0138 |
| Sample SD        | $L^\infty$ | 0.0003 | 0.0002 | 0.0002 | 0.0002 | 0.0002 |
|                  | $L^2$  | 0.0007 | 0.0007 | 0.0007 | 0.0007 | 0.0007 |
|                  | $KL$   | 0.0008 | 0.0007 | 0.0007 | 0.0007 | 0.0007 |
| **Four Variables** |       |       |       |       |       |
| Sample Mean      | $L^\infty$ | 0.0161 | 0.0141 | 0.0141 | 0.0139 | 0.0140 |
|                  | $L^2$  | 0.0361 | 0.0317 | 0.0319 | 0.0314 | 0.0315 |
|                  | $KL$   | 0.0487 | 0.0361 | 0.0361 | 0.0352 | 0.0355 |
| Sample SD        | $L^\infty$ | 0.0004 | 0.0004 | 0.0004 | 0.0004 | 0.0004 |
|                  | $L^2$  | 0.0008 | 0.0008 | 0.0008 | 0.0008 | 0.0008 |
|                  | $KL$   | 0.0013 | 0.0011 | 0.0011 | 0.0011 | 0.0011 |
| **Five Variables** |       |       |       |       |       |
| Sample Mean      | $L^\infty$ | 0.0224 | 0.0187 | 0.0185 | 0.0183 | 0.0191 |
|                  | $L^2$  | 0.0479 | 0.0406 | 0.0403 | 0.0400 | 0.0419 |
|                  | $KL$   | 0.0904 | 0.0664 | 0.0651 | 0.0643 | 0.0690 |
| Sample SD        | $L^\infty$ | 0.0004 | 0.0004 | 0.0004 | 0.0004 | 0.0004 |
|                  | $L^2$  | 0.0009 | 0.0009 | 0.0009 | 0.0009 | 0.0009 |
|                  | $KL$   | 0.0019 | 0.0018 | 0.0017 | 0.0017 | 0.0018 |
we have $A(\pi^l + xd) = b$. The line $L = \{\pi^l + xd | \forall x\}$ that intersects $T$ is a source of consistent Spearman rank-correlation matrices and can be randomized by choosing $x^* \in [x_{\min}, x_{\max}]$, where the minimum and maximum are defined according to $L \cap T$. To recover the Spearman rank-correlation matrix, calculate $\rho$ for the distribution $\pi^l + x^*d$ and include the corresponding constraints into the structure of the Truth Set.

**Step 3:** For each Truth Set, generate a collection of one million joint distributions, using JDSIM. Compute the average divergence from each sample point to each of our five approximations, using our accuracy measures.

The procedure described above does not attempt to create a uniform sample of $T_i$'s over the space of all possible Truth Sets. Instead, we select a sample of polytopes with

---

**Figure 7.** Accuracy of ME, CC, AC, and MV compared to SA in sets with five RVs having arbitrary marginals and rank correlation. The x-axes index the $T_i$'s from 1 to 1,000 by the sort values of SA (black dot), and the y-axes show the error measure using KL divergence. In each subfigure, a second approximation is shown (gray dot), where a line between approximations indicates the difference in accuracy.
correlations that vary uniformly between independent and fully dependent variables. This strategy increases the diversity of the polytopes sampled and provides flexibility over sets for which marginals have previously been defined.

The new results are shown in Figures 6 and 7, in the case of three and five RVs, using KL divergence. A complete set of results can be found in the online supplement’s Figures 10, 11, and 12.

The addition of information regarding the correlation alters the accuracy of SA, but the results are consistent with the prior section. ME again has the lowest accuracy (highest error), and both AC and MV provide more accurate results than the other approximations. However, in this case, MV has better overall accuracy.

Table 4 provides a summary of this section’s results, including those given in the online supplement. These results and conclusions are similar to Table 3 for marginal and correlation information: ME underperforms the other approximations, across all of our error measures. AC performs very well and, considering the relative ease with which it can be computed, looks to be a very attractive approximation in cases of partial information. Finally, MV and SA provide the most accurate approximations in general at the expense of longer computing times.

5. Discussion and Conclusion

We have performed a series of experiments that suggest maximum entropy (ME) is not the most representative distribution in the case of partial probabilistic information. Given the widespread use and theoretical underpinnings of ME, this result deserves elaboration, with regard to both the rationale for ME methods and our computational procedure.

ME will select the distribution that is the most uncertain (i.e., highest entropy) while being consistent with the information that is known. For example, if one states that they have “no idea” regarding the probabilities for a RV with n outcomes, ME will assign $1/n$ to each. Thus, ME imbues the omission of a probability assessment with the life of a formal mathematical statement. This is not the problem addressed in this paper. Rather, as stressed at the outset, we are concerned with the situation in which the decision maker or the expert has not been asked to assess particular probabilities. This is not the same as asking them, their thinking carefully, and their responding that they have no idea. In fact, a decision on the part of the analyst to treat the omission of probabilistic information as a statement of ignorance is mixing their probabilistic knowledge with that of the decision maker or the expert. The fact that the analyst has no idea what probability to put into their model is not a justification for using ME; the maximum entropy principle applies to a single individual’s probabilistic reasoning, not to the analyst’s reasoning about the expert’s thoughts. Instead, we believe that the analyst should use a distribution that is representative of the set of feasible distributions. This representative distribution is unlikely to be one in which an assumption of ignorance, and thereby independence, is enforced upon all unassessed information.

An obvious shortcoming of our work is that we have sampled the Truth Set uniformly. As discussed early on, this approach was taken simply because we do not know
how to assign a probability distribution over the set of all feasible probability distributions. We do believe, however, that the approach taken here provides a great deal of insight into the behavior of the differing approximations and begins to cast doubt upon the ME approach. We hope that others will be able to extend our work and consider other distributional assumptions. These assumptions, if valid, could overturn our conclusions.

These caveats notwithstanding, our current results argue for use of the analytic center (AC) in case of partial probabilistic information. AC performed very well (the best or very close to the best) for nearly all of the cases considered. Additionally, like ME, which has enjoyed widespread use, AC is easy to compute exactly, using well-known optimization methods. If very accurate approximations are needed, then the maximum-volume inscribed ellipsoid and the sample average should be considered. However, these approximations require significantly more computational resources than does AC.

A final shortcoming of our work is that we have analyzed a series of experiments that we believe test the accuracy of differing approximations in a variety of settings. However, we have not tested every possible situation or developed analytic results that prove that one approximation is always best under a particular metric. Nonetheless, we believe that the experiments considered here are a representative sample of the problems found in practice. If true, our results argue strongly for the use of AC in research and practice.

Acknowledgments

This work was supported by the National Council of Science and Technology of Mexico (CONACyT), and the National System of Researchers (SNI) under Luis V. Montiel Grant: No 259968.

Funding

Consejo Nacional de Ciencia y Tecnología - Sistema Nacional de Investigadores; National Science Foundation.

References

Bickel, J. Eric, James E. Smith. 2006. Optimal sequential exploration: A binary learning model. *Decision Analysis* 3(1) 16–32.
Boyd, Stephen P., Lieven Vandenberghe. 2004. *Convex Optimization*. Cambridge University Press, Cambridge, UK.
Bregman, Lev M. 1967. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics* 7(3) 200–217.
Jaynes, Edwin T. 1957. Information theory and statistical mechanics. *Part 1. Physical Review* 106(4) 620–630.
Jimenez, Luis, David Landgrebe. 1998. Supervised classification in high dimensional space: Geometrical, statistical and asymptotical properties of multivariate data. *IEEE Trans. on Sys., Man, and Cybernetics* 28-C(1) 39–54.
John, F. 1948. Extreme problems with inequalities as subsidiary conditions. *Interscience, Studies and Essays*, presented to R. Courant on his 60th Birthday 2 187–204.

Lovasz, Laszlo. 1998. Hit-and-run mixes fast. *Mathematical Programming* 86(3) 443–461.

Mackenzie, George R. 1994. Approximately maximum entropy multivariate distributions with specified marginals and pairwise correlations. Dissertation, University of Oregon.

McMullen, P. 1970. The maximum number of faces of a convex polytope. *Mathematika* 17(2) 179–184.

Montiel, Luis V. 2012. Approximations, simulation, and accuracy of multivariate discrete probability distributions in decision analysis. Ph.D. thesis, The University of Texas at Austin, Austin, TX.

Montiel, Luis V., J. Eric Bickel. 2012. A simulation-based approach to decision making with partial information. *Decision Analysis* 9(4) 329–347.

Montiel, Luis V., J. Eric Bickel. 2013a. Approximating joint probability distributions given partial information. *Decision Analysis* 10(1) 26–41.

Montiel, Luis V., J. Eric Bickel. 2013b. Generating a random collection of discrete joint probability distributions subject to partial information. *Methodology and Computing in Applied Probability* 15(4) 951–967.

Montiel, Luis V., J. Eric Bickel. 2014. A generalized sampling approach for multilinear utility functions given partial preference information. *Decision Analysis* 11(3) 147–170.

Schmidt, Brian K., T. H. Mattheiss. 1977. The probability that a random polytope is bounded. *Mathematics of Operations Research* 2(3) 292–296.

Smith, Robert L. 1984. Efficient Monte Carlo procedures for generating points uniformly distributed over bounded regions. *Operations Research* 32(6) 1296–1308.

Vandenberghe, Lieven, Stephen P. Boyd, S. Wu. 1998. Determinant maximization with linear matrix inequality constraints. *SIAM Journal on Matrix Analysis and Applications* 19 499–533.

Ye, Yinyu. 1997. *Interior Point Algorithms, Theory and Analysis*. Wiley Interscience, New York, NY.

### Appendix A. Structure of the Truth Set

We define the Truth Set as $T \equiv \{ \pi | A \pi = b, \pi \geq 0 \}$. The first row of $A$ constrains the probabilities to sum to one. Rows 2 to 7 select the joint elements of $\pi$ necessary to generate the marginal event probabilities for each of the wells’ being wet. Rows 8 to 22 select the joint elements of $\pi$ necessary to generate the pairwise event probabilities of wells $i$ and $j$, $i \neq j$, to be wet. $A$ is then defined as follows:

$$A = \begin{bmatrix}
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\

L. V. MONTIEL AND J. E. BICKEL
Appendix B. Maximum-Volume Inscribed-Ellipsoid Center

Here we illustrate the implementation of the maximum-volume inscribed-ellipsoid center using an instance with two binary RVs \((X, Y)\). Suppose that the marginal probability of \(X = 1\) is known to be 0.9, while \(P(Y)\) is unknown. Hence, the matrix \(A\) and vector \(b\) are as follows:

\[
A = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 0.9 \end{bmatrix}.
\]

The set of all non-negative solutions to \(A\pi = b\) provides all consistent joint distribution \(\pi = [\pi_1, \pi_2, \pi_3, \pi_4]\), where \(\pi_1 = P(X = 1, Y = 1), \pi_2 = P(X = 1, Y = 0), \pi_3 = P(X = 0, Y = 1),\) and \(\pi_4 = P(X = 0, Y = 0)\). An arbitrary solution for \(A\pi = b\) is \(\pi = [0.6, 0.3, 0.05, 0.05]\), and a possible orthogonal basis for \(Z \equiv \text{Null}(A)\) is:

\[
Z = \frac{1}{2} \begin{bmatrix} -1 & -1 \\ 1 & 1 \\ 1 & -1 \\ -1 & 1 \end{bmatrix}
\]

After applying the appropriate transformations we have the following semidefinite program:

\[
y^* = \arg \max_{\pi^*} \log \det \mathbf{E},
\]

\[
\begin{align*}
\frac{1}{2}[-1, -1] - y + \frac{1}{2}[-1, -1]\|E\|_2 & \leq 0.6, \\
\frac{1}{2}[-1, -1] - y + \frac{1}{2}[-1, -1]\|E\|_2 & \leq 0.3, \\
\frac{1}{2}[-1, -1] - y + \frac{1}{2}[-1, -1]\|E\|_2 & \leq 0.05,
\end{align*}
\]

\(\pi^* = [0.35, 0.28],\) where the first four constraints are tight. Hence, the center of the ellipsoid is given by \(\pi^{MV} = Zy^* + \tilde{\pi} = [0.45, 0.45, 0.05, 0.05]\).

Appendix C. Marginal and Correlation Information

Suppose we have three Bernoulli random variables \(V_1, V_2,\) and \(V_3\). The joint distribution will be given by \(p_{v_1,v_2,v_3} = P(V_1 = v_1, V_2 = v_2, V_3 = v_3)\) where \(v_i\) can take values \(\{0,1\}\) for \(i = 1, 2, 3\). As a joint distribution, the sum of all \(p_{v_1,v_2,v_3}\) has to be equal to one, and each element needs to be non-negative. The joint distribution is unknown, but if we had information of the marginal distributions, for example \(P(V_1 = 1) = 0.5, P(V_2 = 1) = 0.6,\) and \(P(V_3 = 1) = 0.7,\) we could use this information to restrict the joint distribution using the following equations:

\[
\begin{align*}
p_{111} + p_{110} + p_{101} + p_{100} &= 0.5, \\
p_{111} + p_{110} + p_{011} + p_{010} &= 0.6, \\
p_{111} + p_{101} + p_{011} + p_{001} &= 0.7.
\end{align*}
\]

Now assume we know the rank correlation between \(V_1\) and \(V_2\) to be \(\rho_{12} = 0.6\). Following Montiel and Bickel (2013b), we obtain the discrete cumulative marginal distributions of \(V_1\) and \(V_2\) as \(\{0,0.5,1\}\) and \(\{0,0.4,1\}\) and construct \(B\) using the cumulative marginals as in Figure A.C1.
Figure A.C1. Set of vertices $B_i$ and sections $B_i = [x_1, x_2] \times [y_1, y_2]$ where $x_1, x_2, y_1,$ and $y_2,$ are the corresponding vertices for each section.

We calculate the $H$-volume $V_H[B_i] = H(x_2, y_2) - H(x_2, y_1) - H(x_1, y_2) + H(x_1, y_1)$ using the vertices of $B_i$ for each $i$, where $H = x^2 \cdot y^2$. For example:

$V_H(B_1) = 0.5^2 \cdot 0.4^2 = 0.04,$

$V_H(B_2) = 0.25 - 0.04 = 0.21,$

$V_H(B_3) = 0.16 - 0.04 = 0.12,$

$V_H(B_4) = 1 - 0.16 - 0.25 + 0.04 = 0.63.$

Each $B_i$ corresponds to a joint event of the distribution of $P(V_1 = v_1, V_2 = v_2) = p_{v_1v_2} + p_{v_1v_0}$. Then $B_1$ corresponds to $P(V_1 = 0, V_2 = 0), B_2$ corresponds to $P(V_1 = 1, V_2 = 0)$, and so on. For each joint event define the constants $C_i$ as:

$C_1 = \frac{V_H(B_1)}{P(V_1 = 0) \cdot P(V_2 = 0)} = 0.2,$

$C_2 = \frac{V_H(B_2)}{P(V_1 = 1) \cdot P(V_2 = 0)} = 0.6,$

$C_3 = \frac{V_H(B_3)}{P(V_1 = 0) \cdot P(V_2 = 1)} = 0.7,$

$C_4 = \frac{V_H(B_4)}{P(V_1 = 1) \cdot P(V_2 = 1)} = 2.1.$

Using constants $C_1$ to $C_4$ we can restrict the correlations of variables $V_1$, and $V_2$ as:

$C_4(p_{111} + p_{110}) + C_2(p_{101} + p_{100}) + C_3(p_{011} + p_{010}) + C_4(p_{001} + p_{000}) = \frac{0.6 + 3}{3}.$

Following a similar procedure for all pairs of Bernoulli random variables we have the following set of restrictions:

$2.10(p_{111} + p_{110}) + 0.60(p_{101} + p_{100}) + 0.70(p_{011} + p_{010}) + 0.20(p_{001} + p_{000}) = \frac{\rho_{12} + 3}{3},$

$1.95(p_{111} + p_{101}) + 0.45(p_{110} + p_{100}) + 0.65(p_{011} + p_{010}) + 0.15(p_{001} + p_{000}) = \frac{\rho_{13} + 3}{3},$

$1.82(p_{111} + p_{011}) + 0.42(p_{110} + p_{010}) + 0.52(p_{101} + p_{001}) + 0.12(p_{001} + p_{000}) = \frac{\rho_{23} + 3}{3}.$

Given these equations, the independent joint distribution requires $\rho_{12} = \rho_{13} = \rho_{23} = 0$. Hence, the Truth Set will be radically different if we include these constraints or not. The maximum entropy approximation implicitly encode independence, unless it has been explicitly specified, whereas other alternatives tend to be more conservative by not assuming a specific correlation when assessments have not been made.

Appendix D. Authors Brief

Luis V. Montiel

Luis V. Montiel is an Assistant Professor at Instituto Tecnologico Autonomo de Mexico (ITAM) at Mexico City. His main research interest is in mathematical modeling for optimization under uncertainty, with a special interest in decision analysis and simulation learning for optimization. His current research provides a theoretical and practical framework for approximating
joint distributions under partial information. Most recently Dr. Montiel has used this framework to develop algorithms for applications in areas such as negotiations with multiple actors and electoral models for presidential election predictions and optimization. Dr. Montiel has a Post-Doc and a Ph.D. in Operations Research from The University of Texas at Austin, a M.S. in financial engineering from Columbia University and a M.S. in management science and engineering from Stanford University.

Address: Industrial Engineer and Operations. Instituto Tecnologico Autonomo de Mexico. Rio Hondo 1, Altavista, 01080 Ciudad de Mexico, CDMX.
Phone: +52-5518-676167, Email: lvmontiel@utexas.edu.

J. Eric Bickel

J. Eric Bickel is Associate Professor and Director of the Graduate Program in Operations Research and Industrial Engineering at The University of Texas at Austin. In addition, he is graduate advisor for the Engineering Management program and Academic Director of the McCombs School of Business Strategic Decision and Risk Management program. Prior to returning to academia, Professor Bickel was a Senior Engagement Manager for Strategic Decisions Group, where he remains on the board of directors. He has consulted around the world in a range of industries, including oil and gas, electricity generation/transmission/delivery, energy trading and marketing, commodity and specialty chemicals, life sciences, financial services, and metals and mining. Professor Bickel is the President of the Decision Analysis Society. He holds an MS and PhD from the Department of Engineering-Economic Systems at Stanford University.

Address: Graduate Program in Operations Research and Industrial, 1 University Station, C2200, The University of Texas at Austin, Austin, Texas, 78712-0292, ETC 5.128C.
Phone: 512.232.8316, Email: ebickel@utexas.edu.