Abstract

We consider the problem of generating a joint distribution for a pair of Bayesian networks that preserves the multivariate marginal distribution of each network and satisfies prescribed correlation between pairs of nodes taken from both networks. We derive the maximum entropy distribution for any pair of multivariate random vectors and prescribed correlations and demonstrate numerical results for an example integration of Bayesian networks.

Mathematics Subject Classification: 60E05; 62E10; 62B10; 62H10; 94A17

Keywords: Bayesian Networks, Maximum Entropy, model integration

1 Introduction

Bayesian networks provide a compact representation of relationships between variables in terms of conditional independence of events and associated conditional probabilities [3]. They were originally used to model causal relationships [9], and their use has extended to a wide range of applications. Graphically, a Bayesian network consists of a set of nodes, representing individual random variables, and directed arcs from “parent” nodes to “child” nodes indicating conditional dependence. Cycles are not allowed, so that the resulting structure
is an acyclic directed graph. The complete joint distribution of the set of nodes is defined by a conditional probability distribution for each node relative to its parent nodes. Bayes’ theorem can be used to make inferences about the underlying (unobserved) state of the modeled system from observations. This structure provides a means to decompose and rapidly compute conditional and marginal probability distributions, generating likelihoods of specific events or conditions.

Often only partial information is available on the dependence structure and conditional probability distributions of a Bayesian network. In this case some method for completing the model is needed. Schramm and Fronhöfer [11] for example consider the problem of completing a Bayesian network given an incomplete set of conditional probabilities, such that dependence structure is preserved. Additionally, distinct, complete Bayesian networks may be developed to model different aspects of a single process or system, and it may be desired to link the models together to create an integrated model. If the models are consistent (that is, they do not have conflicting assumptions in their structure or parameters), then they may be combined by determining dependence between pairs of nodes from each network (a sort of cross-network dependence) and estimating the conditional probabilities that are associated with that dependence. In general, model building in this manner can be time-consuming due to the challenges in obtaining the needed conditional probability parameters for the resulting combined model. Additionally—if the component models were developed by separate teams—there may not be a straightforward way to arrive at a combined model that is both a Bayesian network and is consistent with the component models.

We consider integration of Bayesian networks into a single, complete distributional model when only linear correlation values for pairs of cross-network nodes (or more generally, moment conditions) are provided, without requiring that the result be representable as a Bayesian network in itself. We impose the constraints that the resulting model match the (joint) marginal distributions given by the original networks as well as the prescribed correlations. Figure (1) illustrates this concept of integration for a simple example.

Figure 1: Bayes Net diagrams for example integration.
General problems of constructing joint distributions with known marginals and prescribed moments has been studied for some time (see for example [1], [5], and reviews in [7] and [10]). Maximum entropy and related methods provide natural approaches to such problems that have broad interest and applicability. Maximum entropy methods aim to derive the joint distribution that satisfies supposed a priori knowledge about the distribution while expressing the greatest possible ignorance or uncertainty about what is not known. Closely related, minimum cross-entropy methods aim to produce the distribution that is closest to a reference distribution while satisfying the imposed knowledge about the distribution. Most applications of these methods impose either marginal (almost always univariate) distributions or moments as constraints, but rarely combine both constraints as desired here. Csiszár [1] produced a set of foundational theorems for very general application of minimum cross-entropy for this purpose, with maximum entropy as a special case, for either marginal or moment constraints. He argues that the approach extends to multivariate marginal constraints. The author also generalized an iterative method for calculating the minimum cross-entropy solution. Miller and Liu [7] summarized existing theory and, following Csiszár, extended the applicability of minimum cross-entropy to include multivariate moment constraints, using the product of univariate marginals as the reference distribution.

Pasha and Mansoury [8] considered the maximum entropy approach using simultaneous constraints for univariate marginals and moments for multivariate distributions. They propose directly solving the set of nonlinear equations that obtain by imposing the constraints on the form of maximum entropy distribution, to calculate the distribution parameters. To solve the problem of combining distinct, complete Bayesian networks given some knowledge of the dependence between each network, we provide the extension of the maximum entropy method to constraints in the form of multivariate marginal distributions and moments. The resulting distribution for Bayesian networks with discrete conditional distributions is implicitly defined by a set of nonlinear equations. We demonstrate a method of numerical solution of those equations for a simple example to produce the joint distribution, and conclude with a discussion of the applicability and computational challenges of implementing this approach to model integration.

2 Maximum entropy solutions

Consider random vectors $X = (X_1, X_2, \ldots, X_m)$ and $Y = (Y_1, Y_2, \ldots, Y_n)$ with probability density (or mass) functions $g(x)$ and $h(y)$, respectively. We will ultimately apply the maximum entropy solution to the problem of integrating Bayesian networks, where elements of $X$ and $Y$ correspond to nodes in each network. First, a more general case is studied. We wish to find a joint prob-
ability density (mass) function $f(z)$ for $Z = (X_1, X_2, \ldots, X_m, Y_1, Y_2, \ldots, Y_n)$ defined on $\Omega \subseteq \mathbb{R}^{m+n}$ subject to the marginal constraints

$$f_X(x) = g(x), \quad f_Y(y) = h(y),$$

and a vector of constraints on moments

$$E_f \gamma(z) = \bar{\gamma},$$

where each element of the vector $\gamma$ is real-valued and has finite expectation.

Let $\mathcal{F}$ denote the class of all $(m+n)$-variate probability density (mass) functions $f(z)$ taking on positive values on $\Omega$ and satisfying the above constraints:

$$\mathcal{F}(g,h,\gamma) = \left\{ f(z) \left| \int_{\Omega_x} f(z) \, dy = g(x), \int_{\Omega_y} f(z) \, dx = h(y), E_f \gamma(z) = \bar{\gamma} \right. \right\},$$

where $\Omega_x = \{ y \in \mathbb{R}^n | z \in \Omega \}$ and $\Omega_y = \{ x \in \mathbb{R}^m | z \in \Omega \}$, and $dx$ and $dy$ are shorthand for $dx_1 dx_2 \cdots dx_m$ and $dy_1 dy_2 \cdots dy_n$, respectively.

The entropy of the distribution $f(z)$ is defined by

$$H(f) = -E[\log f(Z)] = -\int_{\Omega} [f(z) \log f(z)] \, dz.$$

A maximum entropy solution is a member of $\mathcal{F}$ that maximizes $H(f)$. The form of the maximum entropy distribution is easily derived, or assumed based on similar maximum entropy problems. Proof can follow directly from arguments in Csiszár [1] or Ebrahimi et al. [2]; a proof is given here for completeness. Note that the maximum entropy solution may not exist, conditions for which are discussed by Csiszár and others.

**Theorem 2.1** If $\mathcal{F} \neq \emptyset$, then the maximum entropy solution is uniquely given by

$$f^*(z) = f^*_X(x) f^*_Y(y) \exp [\nu' \gamma(z)]$$

for some functions $f^*_X(x)$ and $f^*_Y(y)$ and parameter vector $\nu$.

**Proof:** The solution is found by calculus of variations. The Lagrangian is

$$L(f) = -\int_{\Omega} f \log f \, dz + c \left( \int_{\Omega} f \, dz - 1 \right)$$

$$+ \int_{\Omega_x} \lambda(x) \left[ \int_{\Omega_x} f(x,y) \, dy - g(x) \right] \, dx$$

$$+ \int_{\Omega_y} \mu(y) \left[ \int_{\Omega_y} f(x,y) \, dx - h(y) \right] \, dy$$

$$+ \nu' \int_{\Omega} [\gamma(z) - \bar{\gamma}] \, f \, dz.$$
where $\lambda(x)$ and $\mu(y)$ are Lagrange multipliers corresponding to the marginal constraints, and $\nu$ is a vector of Lagrange multipliers corresponding to the moment constraints. The extremal function $f^*$ of $L$ is a unique maximum due to the concavity of $L$ in $f$, a consequence of the concavity of entropy in $f$. Setting a variational derivative with respect to $f$ equal to zero at $f^*$ gives

$$-\log f^* + c - 1 + \lambda(x) + \mu(y) + \nu'\gamma(z) = 0$$

(equivalently, the Euler-Lagrange equations apply). Solving for $f^*$ gives

$$f^*(z) = \exp \left[ \lambda_0 + \lambda(x) + \mu(y) + \nu'\gamma(z) \right] = f^*_X(x)f^*_Y(y)\exp[\nu'\gamma(z)],$$

for some functions $f^*_X(x)$ and $f^*_Y(y)$, and parameter vector $\nu$, where $\lambda_0 = c - 1$.

For any function $f$ in $\mathcal{F}$ and $f^*$ as in Equation (3), the relative entropy (or Kullback-Leibler divergence) of $f$ with respect to $f^*$ is given by $D(f \parallel f^*) = E_f\log(f/f^*)$. The relative entropy is non-negative, and is zero if and only if $f = f^*$, by Gibbs’ inequality. The entropy of $f$ may be written as

$$H(f) = -D(f \parallel f^*) - E_f\log f^*$$

so that

$$H(f) \leq -E_f\log f^*(Z) = -E_f\log \left( \exp \left[ \lambda_0 + \lambda(X) + \mu(Y) + \nu'\gamma(z) \right] \right) = -\lambda_0 - E_f\lambda(X) - E_f\mu(Y) - \nu'\bar{\gamma} = -\lambda_0 - E_f\lambda(X) - E_f\mu(Y) - \nu'\bar{\gamma} = H(f^*).$$

Equality holds if and only if $f = f^*$ almost everywhere on $\Omega$.

The maximum entropy solution can be explicitly obtained in a direct manner by solving the nonlinear system of equations that is obtained by substituting Equation (3) into the constraint equations (1) and (2). While we have focused on the combination of two multivariate marginal densities along with moment constraints, the theory easily extends to any number of multivariate marginal densities. In particular, letting $X_i \in \mathbb{R}^{d_i}$ be random vectors with probability density (or mass) functions $g_i(x_i)$ for $i = 1, 2, \ldots, n$ where $d_i \in \mathbb{Z}^+$ and letting $Z = (X_1, X_2, \ldots, X_n)$, the maximum entropy solution $f^*(z)$ that satisfies marginal constraints equivalent to Equation (1) and moment constraints equivalent to Equation (2) is given by

$$f^*(z) = \prod_{i=1}^n f^*_X(x_i)\exp[\nu'\gamma(z)]$$

for some functions $f^*_X(x_i), i = 1, 2, \ldots, n$ and parameter vector $\nu$. 
3 Application to Bayesian networks

We use the maximum entropy solution to demonstrate the calculation of a complete distribution for a pair of simple Bayesian networks with discrete state space, given their multivariate marginal distributions and a subset of the \((x, y)\) correlations:

\[
EX_iY_j = \rho_{ij}\sigma_i\tau_j + \mu_i\nu_j \quad \text{for } (i, j) \in I,
\]

for prescribed values of \(\rho_{ij}\), where \(\sigma_i^2 = \text{Var } X_i\), \(\tau_j^2 = \text{Var } Y_j\), \(\mu_i = EX_i\), \(\nu_j = EY_j\), and \(I\) is the set of paired indices \((i, j)\) for which correlation is imposed. The nonlinear system that obtains by applying the constraint equations (1) and (2) to the formal maximum entropy distribution in Equation (3) is

\[
f^*_X(x) \int_{\Omega_y} f^*_Y(y) \exp \left[ \sum_{(i,j) \in I} \lambda_{ij} x_i y_j \right] dy = g(x),
\]

\[
f^*_Y(y) \int_{\Omega_x} f^*_X(x) \exp \left[ \sum_{(i,j) \in I} \lambda_{ij} x_i y_j \right] dx = h(y),
\]

\[
\int_{\Omega} x_k y_l f^*_X(x) f^*_Y(y) \exp \left[ \sum_{(i,j) \in I} \lambda_{ij} x_i y_j \right] dx dy = \rho_{kl} \sigma_k \tau_l + \mu_k \nu_l,
\]

for \((k, l) \in I\).

Consider the example shown earlier in Figure 1, with conditional probabilities in Table 1. Let the probability of success of \(X_1\) be 0.6, and likewise for \(Y_1\). We impose the single correlation constraint \(\rho(X_3, Y_2) = \rho_{32} = 0.036\). The system of equations was implemented in Matlab [6], and a complete numerical solution is given in Table 2. The required positive correlation between \(X_3\) and \(Y_2\) is reflected in the table by comparing the results to the joint probabilities in Table 3 obtained when \(X\) and \(Y\) are assumed to be independent.

4 Discussion

The maximum entropy solution presented here for deriving a joint probability distribution subject to both multivariate marginal distribution constraints and moment constraints may have general applicability beyond the present goal of integrating Bayesian network models. The primary challenge to its application is in the computational complexity that arises from solving a nonlinear
### Table 1: Conditional probabilities for example integration

<p>| X₁ | X₂ | P(X₂|X₁) | Y₁ | Y₂ | P(Y₂|Y₁) | Y₃ | P(Y₃|Y₂) |
|-----|-----|---------|-----|-----|----------|-----|----------|
| 0   | 0   | 0.4     | 0   | 0   | 0.4      | 0   | 0.4      |
| 0   | 1   | 0.6     | 0   | 1   | 0.6      | 0   | 0.6      |
| 1   | 0   | 0.3     | 1   | 0   | 0.3      | 1   | 0.3      |
| 1   | 1   | 0.7     | 1   | 1   | 0.7      | 1   | 0.7      |</p>
<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>Z</th>
<th>X * Y</th>
<th>X * Z</th>
<th>Z * Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.0054</td>
<td>0.0075</td>
<td>0.0077</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.0081</td>
<td>0.0091</td>
<td>0.0106</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.0045</td>
<td>0.0051</td>
<td>0.0068</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.0106</td>
<td>0.0119</td>
<td>0.0158</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.0051</td>
<td>0.0058</td>
<td>0.0077</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.0068</td>
<td>0.0076</td>
<td>0.0098</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.0068</td>
<td>0.0076</td>
<td>0.0098</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.0158</td>
<td>0.0177</td>
<td>0.0229</td>
</tr>
</tbody>
</table>

Table 2: Joint Probability Table

Correlation between \( X_z \) and \( X \)
<table>
<thead>
<tr>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$Y_1$ 0</th>
<th>1</th>
<th>0</th>
<th>1</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.0041</td>
<td>0.0046</td>
<td>0.0046</td>
<td>0.0081</td>
<td>0.0061</td>
<td>0.0069</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.0035</td>
<td>0.0039</td>
<td>0.0039</td>
<td>0.0068</td>
<td>0.0052</td>
<td>0.0058</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.0061</td>
<td>0.0069</td>
<td>0.0069</td>
<td>0.0121</td>
<td>0.0092</td>
<td>0.0104</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.0081</td>
<td>0.0091</td>
<td>0.0091</td>
<td>0.0159</td>
<td>0.0121</td>
<td>0.0136</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.0061</td>
<td>0.0069</td>
<td>0.0069</td>
<td>0.0121</td>
<td>0.0092</td>
<td>0.0104</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.0081</td>
<td>0.0091</td>
<td>0.0091</td>
<td>0.0159</td>
<td>0.0121</td>
<td>0.0136</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.0092</td>
<td>0.0104</td>
<td>0.0104</td>
<td>0.0181</td>
<td>0.0138</td>
<td>0.0156</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.0188</td>
<td>0.0212</td>
<td>0.0212</td>
<td>0.0370</td>
<td>0.0282</td>
<td>0.0318</td>
</tr>
</tbody>
</table>
system with a number of parameters increasing rapidly with increasing event space dimension. Cost reduction should be studied through approximations that reduce the dimensions of the parameter space as well as optimization techniques that take maximum advantage of the structure of this particular problem, including iterative, indirect methods [1], [4].

We considered a special case in which the marginals and moments are required to be exactly matched by the combined distribution, thus assuming that these constraints are all consistent. However, Bayesian networks, and more generally any set of distributions to be integrated, can have conflicting model assumptions such that not all the constraints can be exactly met. Miller and Liu [7] point to a solution when the marginals may be inconsistent with correlation, by formulating the problem in terms of minimizing the cross-entropy where the reference distribution is the product of marginals. While they considered only univariate marginals, the concept is easily extended to incorporate multivariate marginals. Yet this still requires that moments be exactly met, and a generalization that allows for approximate matching of the moments as well as marginals is needed to allow balance between the competing objectives.

ACKNOWLEDGEMENTS

This project was supported by Laboratory Directed Research and Development at Pacific Northwest National Laboratory (PNNL). PNNL is operated by Battelle Memorial Institute for the U.S. DOE under Contract DE-AC06-76RLO 1830. The work described in this paper was developed within the context of the Technosocial Predictive Analytics Initiative (http://predictiveanalytics.pnl.gov) at the Pacific Northwest National Laboratory.

References


Integrating Bayes nets using maximum entropy


Received: September, 2010