Bayesian Networks with Degenerate Gaussian Distributions*

Christopher Raphael†

May 31, 2001

Abstract

Bayesian networks compute marginal distributions through the "message passing" algorithm — a series of local calculations involving neighboring cliques of variables in a clique tree. In this work, these message passing computations are extended to the case of degenerate Gaussian potentials which are multivariate Gaussian densities that can have two different kinds of degeneracies corresponding to projections with zero variance and projections with infinite variance. The basic operations of the message passing algorithm, such as representing conditional distributions, extending potentials, and conditioning on observations, create or operate on potentials with various kinds of degeneracies thereby demonstrating the need for such methodology. Computer implementation of this scheme follows easily from the details within and some computational aspects are discussed. We also demonstrate an application of our methodology to automatic musical accompaniment.

Keywords: Bayesian Belief Networks, Graphical Models, Degenerate Gaussian, Conditional Gaussian Potential, Message Passing Algorithm

1 Introduction

Bayesian networks represent the dependency structure of a collection of random variables through a graph: The vertices of the graph are identified with variables while the collection of arcs convey conditional independence relations, [1]. Well-known algorithms facilitate the computation of marginal posterior distributions, identification of MAP configurations, and training of model parameters through a series of local operations known as "message passing" or "flow" calculations, as in the work of Spiegelhalter, Lauritzen, Cowell, Dawid, Jensen, and others, [1], [2], [3], [4], [5]. The computational aspects of such models are well-understood for variables with finite state space, however the methodology for evidence propagation in networks of continuous variables has received less attention.

We focus here on extending the well-known evidence propagation methodology for discrete networks to networks of Gaussian variables. The main difficulty in doing so is that the Gaussian distributions that arise in performing the message passing operation have two kinds of degeneracies. For instance, if we begin with a multivariate Gaussian density on a collection of variables

*This work is supported by NSF grant IIS-9987898.

†Department of Mathematics and Statistics, University of Massachusetts at Amherst, Amherst, MA 01003-4515, raphael@math.umass.edu

and wish to view the density as a function of a larger collection of variables, then we introduce variables the density does not depend on. One could think of representing this "extended" density as a multivariate Gaussian potential with a concentration matrix (inverse covariance matrix) having as null space the space spanned by the new variables. This is precisely the effect of the extension operation of the message passing algorithm. Representing conditional densities also leads to Gaussian potentials whose concentration matrix has nontrivial null space. On the other hand, the incorporation of evidence necessary when variables of the network are observed leads to Gaussian potentials with fixed variables. These can be thought of as potentials whose covariance matrix has a nontrivial null space. Similarly, allowing linear constraints among variables also leads to potentials whose covariance has nontrivial null space. We introduce a methodology in which these degeneracies can be handled in a natural manner throughout the evidence propagation calculations.

We know of two other research directions that address evidence propagation with degenerate Gaussian distributions. The first of these is the work of Shachter and Kenley [6] which is expressed within the "node-elimination" framework, initially formalized by Howard and Matheson [7], [8]. While the node-elimination framework provides a method for computing posterior marginal distributions on individual variables by iteratively reducing the graph through "arc reversals" and "barren node reductions," it does not enjoy certain properties of Bayesian networks. For example, Bayesian networks allow for the simultaneous computation of *all* marginal distributions through an equilibrium representation. This latter representation forms a starting point from which evidence can be entered and disseminated quickly and which facilitates multiple parameter estimation. Additionally, Bayesian networks allow for the efficient computation of some *joint* marginal distributions.

More closely related to our work is the research of Lauritzen and Jensen [9], [10], which deals with evidence propagation in the Bayesian network context in the case of conditional Gaussian distributions — collections of discrete and continuous variables in which the conditional distribution on the continuous variables, given the discrete variables, is multivariate Gaussian. The earlier work [9] deals with one particular kind of degeneracy among the continuous variables, nontrivial null space of the concentration matrix, but does not allow distributions with linear constraints among the variables. The lack of a means of representing this latter kind of degeneracy makes the incorporation of new evidence somewhat awkward, since the evidence must be entered separately into each clique containing the observed variables. This problem, as well as an issue with numerical stability, is partly rectified in [10], however the corresponding methodology is considerably more complicated involving recursive combination of potentials.

Our goal here is to present a straightforward methodology that deals with these two kinds of degeneracies in a systematic way. In particular, we will track two special subspaces corresponding to zero variance projections and "infinite variance" projections through the message passing calculations. We believe that the resulting methodology is simpler than the method of Lauritzen and Jensen, and more flexible that the method of Shachter.

In what follows we provide a precise description of the message passing calculations necessary to use Bayesian networks with degenerate Gaussian distributions. Our results are presented with an eye toward numerical implementation so we provide ample detail to facilitate a computer implementation of our ideas. Section 2 gives a brief account of some general aspects of Bayesian networks we use; Section 3 describes our representation of a possibly degenerate Gaussian potential function and defines the operations necessary to describe the message passing calculations; Section 4 demonstrates the correctness of our algorithm, i.e. that our extension of the message passing algorithm leads to a representation of the joint distribution in term of marginal distributions;

Section 5 discusses some issues regarding numerical implementation; Section 6 demonstrates our approach on a network containing hundreds of variables arising from an application to automatic musical accompaniment. Finally, Section 7 gives the proofs of our results.

2 Background

Suppose we have a random vector, X, whose components are indexed by a finite set V. If $C \subseteq V$, we use the notation X_C to denote variables of X whose indices are in C, thus $X_V = X$. We treat here joint probabilities for X whose densities, $\phi(x)$, can be factored by

$$\phi(x) = \prod_{C \in \mathcal{C}} \phi_C(x_C) \tag{1}$$

such that each ϕ_C depends only on the variables x_C and $\mathcal{C}$ is a collection of subsets organized into a "junction tree." That is, the elements of $\mathcal{C}$ are arranged into a tree, (an undirected graph with no cycles), with the property that for any $C_1, C_2, C_3 \in \mathcal{C}$, with C_3 on the unique path between C_1 and C_2, $C_1 \cap C_2 \subseteq C_3$.

Distributions with such factorizations are attractive since algorithms exist that facilitate the computation of marginal and conditional marginal distributions, as will be discussed. Most often these factorizations arise from a *recursive factorization* of a probability density. That is, assume $\phi(x)$ factors as a product of conditional densities

$$\phi(x) = \prod_{v \in V} f_v(x_v | x_{\mathrm{pa}(v)})$$

where $\mathrm{pa}(v)$ are the parents of v in the directed acyclic graph G. Recursive factorizations are natural when one can order the variables in such a way that the conditional distribution of a variable given the entire "past" only depends on a small subset of the past variables. Suppose that G is first "moralized," by dropping the direction the edges and connecting any pair of nodes that share a child in G, and then triangulated, by adding edges until each cycle of length greater than three is cut by a chord. If we let $\mathcal{C}$ be the cliques of the triagulated graph, then it is easy to see that for any $v \in V$, the family of v, $\mathrm{fa}(v) = v \bigcup \mathrm{pa}(v)$, can be associated with a clique $C(v) \in \mathcal{C}$ such that $\mathrm{fa}(v) \subseteq C(v)$. Thus

$$\phi(x) = \prod_{C \in \mathcal{C}} \prod_{v : C(v) = C} f_v(x_v | x_{\mathrm{pa}(v)})$$

and hence the representation of Eqn. 1. It is well known that the cliques of a triangulated graph can be arranged into a junction tree [12] [4], [9], [2], [3].

Once the junction tree is constructed, for each pair of neighboring subsets in the tree, C_1, C_2, we define the associated separator $S = C_1 \cap C_2$ and let $\mathcal{S}$ be the collection of $|\mathcal{C}| - 1$ separators (not necessarily distinct). By defining $\phi_S(x_S) \equiv 1$ for each $S \in \mathcal{S}$ we can extend our factorization to the form

$$\phi = \frac{\prod_{C \in \mathcal{C}} \phi_C}{\prod_{S \in \mathcal{S}} \phi_S} \tag{2}$$

where each ϕ_C and ϕ_S are functions depending only on the variables in C and S. We refer to such a representation as a potential representation with potential functions $\{\phi_C\}$ and $\{\phi_S\}$.

The virtue of such a representation is that, through a series of local "message passing" computations, the representation can be transformed into an "equilibrium" representation in which the

ϕ_C and ϕ_S functions are the marginal distributions on the associated variables. The equilibrium is invariant to further message passing operations. The message passing calculation can be defined generically in terms of a "marginalization" operator, $\sum_{G \setminus H}$ and an "extension" operator, E_G where G and H are subsets of V with $H \subset G$. Specifically, for neighboring (in the tree) subsets $C_1, C_2 \in \mathcal{C}$ with separator S and associated potentials $\phi_{C_1}, \phi_{C_2}, \phi_S$, a message is passed from C_1 to C_2 when the current potentials ϕ_{C_2}, ϕ_S, are replaced with

$$\phi'_S = \sum_{C_1 \setminus S} \phi_{C_1} \tag{3}$$

$$\phi'_{C_2} = \phi_{C_2} E_{C_2} \frac{\phi'_S}{\phi_S} \tag{4}$$

It is easy to see that the quotient of products in Eqn. 2 is invariant in the message passing operation. Thus, message passing operations do not affect the joint distribution given by Eqn. 2.

The equilibrium representation is achieved by a regimen of such operations known as the "message passing algorithm." The message passing algorithm consists of two phases, *Collect Evidence* followed by *Distribute Evidence* [3]. In *Collect Evidence* one clique is arbitrarily chosen to be the root. Then each non-leaf clique receives a message from each of its non-root-side neighbors. The sequence of messages is coordinated so that a clique can send a message only when it has received messages from all of its non-root-side neighbors. *Distribute Evidence* begins by having each non-root clique receive a message from its root-side neighbor. These messages are coordinated so that a clique cannot send a message until it has received the message from its root-side neighbor. It is well known that, under a variety of different conditions, any message passing scheme that obeys these constraints will result in the equilibrium distribution.

Conditional marginal distributions can be computed in a similar manner. Suppose we begin with a representation of the joint distribution of the form of Eqn. 2, and then observe that one of the variables, x_v takes on the value o_v. Let $C \in \mathcal{C}$ be a clique containing v. If we multiply ϕ_C by the indicator function $1_{x_v = o_v}$, then the representation of Eqn. 2 is modified to contain the conditional distribution of X given $X_v = o_v$. A round of the message passing algorithm then produces an equilibrium distribution in which the potentials contain conditional marginal distributions.

3 Computing with Potentials

This section describes our representation of a Gaussian potential function which plays the role of the ϕ_C and ϕ_S factors in Eqn. 2. To perform the operations that arise in the computation of marginal and conditional marginal distributions, we consider Gaussian potentials with two kinds of degeneracies. The first kind arises when the concentration matrix, (inverse of the covariance matrix), has a non-trivial null space. This kind of degeneracy is used in representing conditional distributions, and in the extension operation. One can think of this kind of degeneracy as a subspace in which the projection has infinite variance. The second kind of degeneracy arises when the covariance matrix has a non-trivial null space. It is necessary represent such degeneracies to allow the conditioning on observed variables and the representation of linear constraints among the variables. This kind of degeneracy corresponds to a subspace in which the projection has 0 variance.

Our representation of a Gaussian potential function on n variables , ϕ, is as a quintuple:

$$\phi = (U^\infty, U^0, U^+, \Sigma, S, \mu)$$

where $\phi(x)$ is defined by

$$\phi(x) = 1_{\{P_{U^0}x = P_{U^0}\mu\}} \exp\{-\frac{1}{2}(x-\mu)^t S(x-\mu)\} \tag{5}$$

In our representation, Σ and S are non-negative definite matrices with $S = \Sigma^-$, the generalized inverse of S [13],[14]. U^∞, U^0, U^+ are mutually orthogonal subspaces that form a decomposition of $\Re^n$. U^0 gives the support of ϕ through $\text{Supp}(\phi) = \{P_{U^0}x = P_{U^0}\mu\}$, where P_{U^0} is projection onto the subspace U^0; U^+ is the range of Σ (and S); and U^∞ is the space on which ϕ does not depend: $\phi(x) = \phi(x + \lambda z)$ for all real λ if and only if $z \in U^\infty$. We will occasionally represent subspaces by matrices with the meaning that the desired space is the range of the matrix.

We require that μ be chosen so that

$$P_{U^\infty}\mu = 0$$

thus making our potential representation unique. For ϕ to correspond to a probability density we must have that $U^\infty = \{0\}$ in which case ϕ proportional to a, possibly degenerate, Gaussian density (with respect to the natural measure on the support set). Our representation of a potential is defined only up to a multiplicative constant so we write $\phi_1 = \phi_2$ when $\phi_1(x) = c\phi_2(x)$ for all x and some constant c.

The basic operations involved in the message passing computation are marginalizing a potential defined on the variables of G to a subset of variables H: $\sum_{G\setminus H}\phi$; extending a potential defined on variables H to a larger set of variables G: $E_G\phi$; multiplying two potentials; and dividing two potentials. The message passing calculation will be completely specified once we describe how these four basic operations are performed in the remainder of the section. We also describe the potential representation of a conditional distribution, since a common way to represent a joint distribution is as a product of conditional distributions, as in recursive factorization.

3.1 Marginalization

Suppose ϕ is a potential on the variables of $G = (x_1, \ldots, x_n)$ with $H = (x_{\pi_1}, \ldots, x_{\pi_m})$ a permutation of some subset of the variables of G. Then the components of $\phi_M = \sum_{G\setminus H}\phi$ are given by

$$\begin{aligned}
U_M^\infty &= I_H U^\infty \\
U_M^0 &= (I_H U^{0\perp})^\perp \\
U_M^+ &= (U_M^\infty \oplus U_M^0)^\perp \\
\Sigma_M &= P_{U_M^{\infty\perp}} I_H \Sigma I_H^t P_{U_M^{\infty\perp}} \\
S_M &= \Sigma_M^- \\
\mu_M &= P_{U_M^{\infty\perp}} I_H \mu
\end{aligned}$$

where I_H is the $m \times n$ matrix that extracts the components of H by $I_H(x_1, \ldots, x_n)^t = (x_{\pi_1}, \ldots, x_{\pi_m})^t$, $A^\perp$ is the orthogonal complement of a subspace A, and $A \oplus B = \{a + b : a \in A, b \in B\}$ — we do not require that $A \cap B = \{0\}$. It is a straightforward exercise to show that all of the constraints on potentials stated in the remarks after Eqn. 5 are satisfied by ϕ_M.

When ϕ is a genuine Gaussian density then $U^\infty = \{0\}$ and ϕ_M reduces to

$$U_M^\infty = \{0\}$$

$$
\begin{aligned}
U_M^0 &= (I_H U^+)^\perp \\
U_M^+ &= I_H U^+ \\
\Sigma_M &= I_H \Sigma I_H^t \\
S_M &= \Sigma_M^- \\
\mu_M &= I_H \mu
\end{aligned}
$$

which is seen to be a potential representation for the marginal distribution on the variables of H.

3.2 Extension

Suppose $G = (x_1, \ldots, x_n)$ is a collection of variables with $H = (x_{\pi_1}, \ldots, x_{\pi_m})$ a permutation of some subset of the variables of G. If ϕ is a potential defined on H, define the components of $\phi_E = E_G \phi$ by

$$
\begin{aligned}
U_E^\infty &= I_H^t U^\infty \oplus I_{H^c}^t \\
U_E^0 &= I_H^t U^0 \\
U_E^+ &= I_H^t U^+ \\
\Sigma_E &= I_H^t \Sigma I_H \\
S_E &= I_H^t S I_H \\
\mu_E &= I_H^t \mu
\end{aligned}
$$

where I_H is as before and I_{H^c} is the $(n-m) \times n$ matrix that extracts the components *not* in H. $E_G \phi$ is an extension of ϕ to the components of G in the sense of

$$
\phi_E(x) = \phi(I_H x)
$$

which can be verified by direct computation with Eqn. 5.

3.3 Multiplication

Theorem 1 *Let $\phi_1 = (U_1^\infty, U_1^0, U_1^+, \Sigma_1, S_1, \mu_1)$ and $\phi_2 = (U_2^\infty, U_2^0, U_2^+, \Sigma_2, S_2, \mu_2)$. If the system*

$$
\begin{pmatrix} P_{U_1^0} \\ P_{U_2^0} \end{pmatrix} x = \begin{pmatrix} P_{U_1^0} \mu_1 \\ P_{U_2^0} \mu_2 \end{pmatrix} \tag{6}
$$

is solvable for x then $\phi_P = \phi_1 \phi_2$ has components given by

$$
\begin{aligned}
U_P^\infty &= U_1^\infty \cap U_2^\infty \\
U_P^0 &= U_1^0 \oplus U_2^0 \\
U_P^+ &= (U_P^0 \oplus U_P^\infty)^\perp \\
\Sigma_P &= S_P^- \\
S_P &= P_{U_P^{0\perp}} (S_1 + S_2) P_{U_P^{0\perp}} \\
\mu_P &= \mu_0 + \Sigma_P (S_1(\mu_1 - \mu_0) + S_2(\mu_2 - \mu_0))
\end{aligned}
$$

where μ_0 is the unique solution in U_P^0 to Eqn. 6.

If Eqn. 6 is not solvable then the supports of ϕ_1 and ϕ_2 do not intersect, so ϕ_P is undefined.

3.4 Division

Theorem 2 *Let ϕ_1, ϕ_2 be potentials such that $\phi_1 = \phi_2\phi_3$ for some potential ϕ_3. Let $\phi_Q = \phi_1/\phi_2$ where 0/0 = 0.*

The components of ϕ_Q are given by

$$\begin{aligned} U_Q^\infty &= U_1^\infty \\ U_Q^0 &= U_1^0 \\ U_Q^+ &= U_1^+ \\ \Sigma_Q &= S_Q^- \\ S_Q &= P_{U_Q^{0\perp}}(S_1 - S_2)P_{U_Q^{0\perp}} \\ \mu_Q &= \mu_0 + \Sigma_Q(S_1(\mu_1 - \mu_0) - S_2(\mu_2 - \mu_0)) \end{aligned}$$

where $\mu_0 = P_{U_1^0}\mu_1$.

3.5 Representing Conditional Distributions

A probability *admits recursive factorization* [1] with respect to a directed acyclic graph (DAG) G if it has a density, $\phi(x)$, that can be represented as

$$\phi(x) = \prod_{v \in V} f_v(x_v | x_{\mathrm{pa}(v)}) \tag{7}$$

where V are the nodes of the graph and $\mathrm{pa}(\gamma)$ are the parents in G of γ. Such a factorization leads in a straightforward manner to a representation of the form of Eqn. 1 as discussed above. We show here how to represent the factors in Eqn. 7 as potentials when these factors are densities for degenerate conditional Gaussians distributions.

If two random vectors x and y have a joint Gaussian distribution then they are related by $y = Ax + \xi$ where ξ is is jointly Gaussian. The case of representing conditional Gaussian distributions is summarized as follows.

Theorem 3 *Suppose two random vectors x and y are related by the equation*

$$y = Ax + \xi$$

where ξ is a Gaussian random vector with potential representation $\xi = (\{0\}, U^0, U^+, \Sigma, S, \mu)$. The conditional distribution of y given x can be represented as a potential $\phi_C(z)$ where $z = \begin{pmatrix} x \\ y \end{pmatrix}$. The components of ϕ_C are given by

$$\begin{aligned} U_C^\infty &= \begin{pmatrix} I \\ A \end{pmatrix} \\ U_C^0 &= \begin{pmatrix} -A^t \\ I \end{pmatrix} U^0 \\ U_C^+ &= (U_C^0 \oplus U_C^\infty)^\perp \\ \Sigma_C &= S_C^- \\ S_C &= (-A\ I)^t S(-A\ I) \\ \mu_C &= P_{U_C^{\infty\perp}} \begin{pmatrix} 0 \\ \mu \end{pmatrix} \end{aligned}$$

4 The Message Passing Calculations

4.1 Modified Message Passing

Rather than performing the message passing calculations as in Eqns. 3–4, we use the slight variation

$$\phi'_S = \sum_{C_1 \setminus S} \phi_{C_1} \tag{8}$$

$$\phi'_{C_2} = (\phi_{C_2} E_{C_2} \phi'_S)/E_{C_2}\phi_S \tag{9}$$

since our definition of potential division given in Section 3.4 constrains the numerator and denominator. We pass messages in two distinct scenarios. In *Collect Evidence* all ϕ_S potentials have been initialized to 1 ($U^\infty = I$) so Eqns. 8–9 become

$$\begin{aligned} \phi'_S &= \sum_{C_1 \setminus S} \phi_{C_1} \\ \phi'_{C_2} &= (\phi_{C_2} E_{C_2} \sum_{C_1 \setminus S} \phi_{C_1})/1 \end{aligned}$$

and clearly our constraint on division is met since the numerator is equal to the denominator times a potential. When we encounter the reverse message during *Distribute Evidence*, the potential at C_2 might have received messages from other cliques and hence is of the form $\tilde{\phi}_{C_2} E_{C_2} \sum_{C_1 \setminus S} \phi_{C_1}$. The message passing calculations for this reverse direction now become

$$\begin{aligned} \phi''_S &= \sum_{C_1 \setminus S} \phi_{C_1} \sum_{C_2 \setminus S} \tilde{\phi}_{C_2} \\ \phi''_{C_1} &= (\phi_{C_1} E_{C_1} \sum_{C_1 \setminus S} \phi_{C_1} E_{C_1} \sum_{C_2 \setminus S} \tilde{\phi}_{C_2})/E_{C_1} \sum_{C_1 \setminus S} \phi_{C_1} \end{aligned}$$

where we use the assumption, discussed later, $\sum_{G \setminus H} E_G \phi_H \phi_G = \phi_H \sum_{G \setminus H} \phi_G$ where ϕ_G and ϕ_H depend on the indicated variables. Clearly the division appearing in this message passing calculation also obeys the constraint imposed by our definition of division.

4.2 Entering Evidence

If a variable x_v, $v \in V$, is observed to be some value o_v, this information can be easily incorporated so that the joint potential representation of Eqn. 2 now represents the posterior distribution on all model variables conditioned on the observation $x_v = o_v$. To do this, we simply identify a clique, C, with $v \in C$. Then ϕ_C is modified so that

$$\phi'_C = \phi_C E_C 1_{\{x_v = o_v\}}$$

where the one-dimensional potential $1_{\{x_v = o_v\}}$ is given by

$$1_{\{x_v = o_v\}} = (U^\infty = \{0\}, U^0 = \Re, U^+ = \{0\}, \Sigma = 0, S = 0, \mu = o_v)$$

Unlike as in Lauritzen [9], [10], the evidence does not need to be entered at each clique containing the variable x_v since our treatment of potential functions allows constraints to propagate through the message passing algorithm.

4.3 Correctness of Algorithm

Shenoy and Shafer [15] show that, for generic operations Σ and E, if

$$\sum_{G\backslash K} \phi = \sum_{H\backslash K} \sum_{G\backslash H} \phi \tag{10}$$

where $K \subseteq H \subseteq G$ and ϕ is a potential on the variables of G, and

$$\sum_{G\backslash H} (E_G \phi_H)\phi_G = \phi_H \sum_{G\backslash H} \phi_G \tag{11}$$

where ϕ_G and ϕ_H are potentials defined on the variables of G and H with $G \subseteq H$ then after a round of the message passing algorithm each resulting ϕ_C and ϕ_S from Eqn. 2 will be proportional to $\sum_{V\backslash C} \phi$[1]

Returning to our case, if ϕ in Eqn. 2 is a genuine Gaussian distribution, that is $U^\infty = \{0\}$, then by the comment at the end of Section 3.1, $\sum_{V\backslash C} \phi$ and $\sum_{V\backslash S} \phi$ are the marginal distributions on the associated variables. Thus our machinery for computing with potentials is guaranteed to produce a factorization of ϕ in terms of marginal distributions in the form of Eqn. 2 if we can establish the axioms of Shenoy and Shafer (Eqn. 10-11).

Theorem 4 *Let $K \subseteq H \subseteq G$ be subsets of variables with ϕ defined on G. Then if the marginalization operator is defined as in Section 3.1, then*

$$\sum_{G\backslash K} \phi = \sum_{H\backslash K} \sum_{G\backslash H} \phi$$

For the second axiom of Shenoy and Shafer (Eqn. 11) we provide a partial proof. Section 5 presents the results of numerical experiments providing strong evidence that the missing identities hold.

Theorem 5 *Suppose $H \subseteq G$ and let*

$$\phi_1 = \sum_{G\backslash H} (E_G \phi_H)\phi_G = (U_1^\infty, U_1^0, U_1^+, \Sigma_1, S_1, \mu_1) \tag{12}$$

and

$$\phi_2 = \phi_H \sum_{G\backslash H} \phi_G = (U_2^\infty, U_2^0, U_2^+, \Sigma_2, S_2, \mu_2) \tag{13}$$

Then $U_1^\infty = U_2^\infty$, $U_1^0 = U_2^0$, and $U_1^+ = U_2^+$.

5 Numerical Implementation

For the purposes of computing, our numerical representation of a potential on n variables, $\phi = (U^\infty, U^0, U^+, \Sigma, S, \mu)$, is a quintuple of matrices. The first three matrices are such that $U = (U^\infty, U^0, U^+)$ is an $n \times n$ unitary matrix; the columns of each of the matrices U^∞, U^0, U^+ form orthonormal bases for the spaces they represent. Σ, S, and μ are exactly as before.

[1] The work in [15] uses a slightly different definition of message passing, but the extension from their definition to the one presented here is straightforward.

Once the U^∞, U^0, U^+ matrices have been computed in any of the operations of Section 3, the Σ, S and μ calculations are straightforward and stable as follows. Once S is known, the calculation of $\Sigma = S^-$ that appears in Sections 3.3, 3.4, and 3.5 is given by

$$\Sigma = S^- = U^+(U^{+t}SU^+)^{-1}U^{+t}$$

where the matrix $U^{+t}SU^+$ is guaranteed to be invertible since U^+ is of full rank with $R(U^+) = R(S)$. A similar calculation holds for $S = \Sigma^-$ of Section 3.1. The projection operations that appear throughout the calculations are computed by, e.g., $P_{U^0} = U^0 U^{0t}$ with similar relations for other projections. The unique solution in U_P^0 of Eqn. 6, μ_0, discussed in Section 3.3, can be found by solving the equivalent system

$$U_P^{0\,t}(P_{U_1^0} + P_{U_2^0})U_P^0 x = U_P^0(P_{U_1^0}\mu_1 + P_{U_2^0}\mu_2)$$

and letting $\mu_0 = U_P^0 x$ where the matrix $U_P^{0\,t}(P_{U_1^0} + P_{U_2^0})U_P^0$ is guaranteed to be invertible.

To compute the matrices U^∞, U^0, U^+ that appear in the operations of Section 3, note that any of the desired "U" spaces can be seen as either $N(A)$ or $N(A)^\perp$ for some matrix A. For instance, taking U_P^∞, U_P^0, U_P^+ of Section 3.3 we have

$$\begin{array}{lcll} U_P^\infty & = & U_1^\infty \cap U_2^\infty & = & N((U_1^0, U_1^+, U_2^0, U_2^+)^t) \\ U_P^0 & = & U_1^0 \oplus U_2^0 & = & N((U_1^0, U_2^0)^t)^\perp \\ U_P^+ & = & (U_P^0 \oplus U_P^\infty)^\perp & = & N((U_P^0, U_P^\infty)^t) \end{array}$$

We can find $N(A)$ and $N(A)^\perp$ by using a numerical singular value decomposition which represents any matrix A as

$$A = WDV^t \tag{14}$$

where W has orthonormal columns, D is diagonal, and V is unitary [16]. Clearly the columns of V corresponding to the 0 diagonal entries of D span $N(A)$ while the remaining columns span $N(A)^\perp$. The numerical determination of the 0 singular values can be problematic in some cases since floating point computations are not exact. In practice we must choose the 0 singular values by thresholding.

We performed an experiment designed both to demonstrate the correctness of the equations $S_1 = S_2$ and $\mu_1 = \mu_2$ appearing in Eqns. 12 and 13, and to explore the numerical reliability of our approach. We randomly chose pairs of test potentials, ϕ_G and ϕ_H, to be of random dimensions N_G, N_H with

$$\begin{array}{rcl} N_G & \sim & \text{UNIF}(2, 3, \ldots, 15) \\ N_H & \sim & \text{UNIF}(1, 2, \ldots, N_G - 1) \end{array}$$

Care must be taken in choosing ϕ_G and ϕ_H since a haphazard random assignment will miss certain important special cases, as follows. In the computation of ϕ_1 of Eqn. 12 we must compute the intermediate result

$$\phi_I = (E_G \phi_H)\phi_G \tag{15}$$

in which we compute

$$\begin{array}{rcl} U_I^0 & = & (I_H^t U_H^0) \oplus U_G^0 \\ U_I^\infty & = & (I_H^t U_H^\infty \oplus I_{H^c}) \cap U_G^\infty \end{array}$$

	Average $\| e \|$	Maximum $\| e \|$	Average $\| \phi_H \|$	Average $\| \phi_G \|$
μ	1.157352e-08	5.751196e-05	3.530415e-01	4.423257e-01
S	1.549258e-08	7.694086e-05	2.258726e-01	3.525940e-01

Table 1: In 4970 out of 5000 of randomly chosen pairs ϕ_G and ϕ_H in which no ambiguous singular values occurred the average and worst case errors in the computations of μ and S.

Suppose the decompositions U_G^∞, U_G^0, U_G^+ and U_H^∞, U_H^0, U_H^+ are chosen randomly and independently, e.g. by randomly partitioning the unitary "V" matrix of Eqn. 14 when A is a $N_G - 1 \times N_G$ (or $N_H - 1 \times N_H$) matrix with independent UNIF$(0,1)$ components. Then, with probability 1, U_I^0 has dimension $\min(N_G, \dim(U_G^0) + \dim(U_H^0))$ and U_I^∞ has dimension $\max(\dim(U_H^\infty) + \dim(U_G^\infty) - N_H, 0)$. Thus, for instance, we only encounter a non-trivial intersection between $I_H^t U_H^0$ and U_G^0 if the sum of the dimensions of the spaces exceeds N_G. Instead we choose U_G^∞, U_G^0, U_G^+ and U_H^∞, U_H^0, U_H^+ so that the pairs $I_H^t U_H^\infty$ and U_G^∞ as well as $I_H^t U_H^0$ and U_G^0 have intersection of randomly chosen dimension. The means μ_G, μ_H were chosen so that the product in Eqn. 15 is defined (cf. Eqn. 6).

For 5000 random pairs of potentials ϕ_H and ϕ_G, chosen as described above we performed the calculations of Eqns. 12 and 13 in double precision using a library of routines written in c. For each pair we computed the norm, $(\| \cdot \|)$, of the error matrices

$$e_\mu = \mu_1 - \mu_2$$
$$e_S = S_1 - S_2$$

where for a $r \times c$ matrix A

$$\| A \| = \left(\frac{1}{rc} \sum_{i,j} A_{ij}^2 \right)^{1/2}$$

Occasionally during the computation of ϕ_1 or ϕ_2 a singular value appears that is "ambiguous" which we have operationally defined as being between 5.000000e-05 and 1.000000e-07. In our 5000-pair test set we encountered ambiguous singular values in 30 cases. Of the remaining 4790 pairs Table 1 summarizes our results. The table shows that in all of cases not involving ambiguous singular values no (normed) error occurred greater than .00001 and the average errors were much smaller. We believe these results argue strongly for the correctness of the potential operations described in Section 3.

Actually, the results were similar among the 30 cases in which ambiguous singular values were detected suggesting that an incorrect assessment of an ambiguous singular value is usually not of consequence. Two exceptions occurred in which we observed errors of (e_μ, e_S) of (1.088484e-01, 7.911990e-02) and (2.674409e+01,3.371012e-04).

Lauritzen [9] shows that the message passing computations can be performed for Gaussian potentials that do not have U^0 spaces. Our method is more costly computationally, due to U^∞, U^0, U^+ spaces that must be tracked throughout the message passing algorithm. These computations are relatively costly due to their reliance on a numerical singular value decomposition. However, in many situations this burden can be significantly decreased or even eliminated. Note that in operations of Section 3, the computations of U^∞, U^0, U^+ do not depend on the values of Σ, S, and μ. For each message passed from C_1 to C_2 we need only retain all past calculations of

U_S^∞, U_S^0, U_S^+ and $U_{C_2}^\infty, U_{C_2}^0, U_{C_2}^+$. In many applications there will be very few distinct such computations performed at each pair of adjacent cliques and these can be stored ahead of time. The "online" calculation of the Σ, S and μ components involves no singular value decompositions and can thus be computed efficiently.

6 An Example: Automated Musical Accompaniment

We have developed an expert system that plays the role of a sensitive, responsive, and trainable musical accompanist in a piece of music composed for soloist and accompaniment. We will discuss this system here while demonstrating the application of our methodology to this problem.

We deal here with non-improvisatory music in which there are two distinguished parts: a part played by the live musician which we call the "solo" part, and a part played by the computer which we call the "accompaniment." Our system makes the assumption that the live player plays the solo part and serves the role of worthy leader and teacher, while the role of the computer is to follow and learn from the soloist.

We have partitioned the accompaniment problem into two components, "Listen" and "Play." Listen takes as input the acoustic signal generated by the soloist and, using a hidden Markov model, performs a real-time analysis of the signal. The output of Listen is essentially a running commentary on the acoustic input which identifies note boundaries in the solo part and communicates these events with variable latency. We focus here on the more difficult "Play" component which attempts to create an accompaniment using the output from Listen. The task of Play is, at the most basic level, one of knowledge fusion since there are a number of disparate knowledge sources that must be synthesized in any plausible attempt at the musical accompaniment task. These include the musical score, the output from Listen, the training data from past "rehearsals," and also training data for the accompaniment part. A more detailed description of this work can be found in [17]

We have developed a probabilistic model, a Bayesian network, that represents all of these knowledge sources through a jointly Gaussian distribution that contains hundreds of random variables. The observable variables in this model are the estimated soloist note onset times produced by Listen and the directly observable times for the accompaniment notes. Between these observable variables lie several layers of hidden variables that describe unobservable quantities such as local tempo, change in tempo, and rhythmic stress.

6.1 The Model

We model the time evolution of the solo part as follows. For each of the solo notes, indexed by $n = 0, \ldots, N$, we define a random vector representing the time, t_n, (in seconds) and the "tempo," s_n, (in secs. per beat) for the note. We model this sequence of random vectors through a random difference equation:

$$\begin{pmatrix} t_{n+1} \\ s_{n+1} \end{pmatrix} = \begin{pmatrix} 1 & l_n \\ 0 & 1 \end{pmatrix} \begin{pmatrix} t_n \\ s_n \end{pmatrix} + \begin{pmatrix} \tau_n \\ \sigma_n \end{pmatrix} \tag{16}$$

$n = 0, \ldots, N-1$, where l_n is the musical length of the n^{th} note in beats and the $\{(\tau_n, \sigma_n)^t\}$ and $(t_0, s_0)^t$ are mutually independent Gaussian random vectors.

The means and variances of the $\{\sigma_n\}$ show where the soloist is speeding-up (negative mean), slowing-down (positive mean), and tell us if these tempo changes are nearly deterministic (low variance), or quite variable (high variance). The $\{\tau_n\}$ variables describe stretches (positive mean)

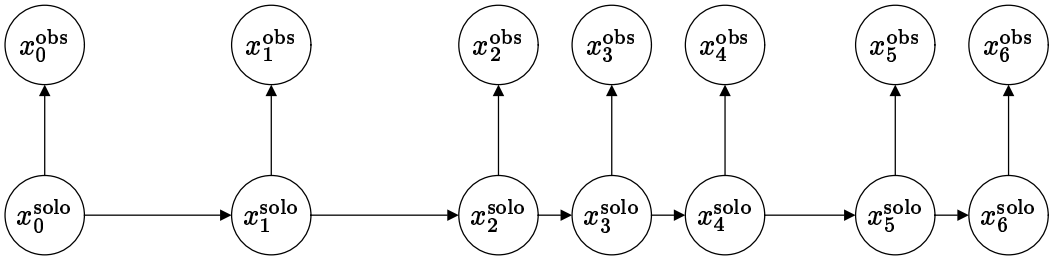

Figure 1: The dependency structure of the $\{x_n^{\text{solo}}\}$ and and $\{x_n^{\text{obs}}\}$ variables. The horizontal placement of graph vertices in the figure corresponds to their times, in beats, as indicated by the score.

or compressions (negative mean) in the music that occur without any actual change in tempo. Thus, the distributions of the $(\tau_n, \sigma_n)^t$ vectors characterize the solo player's rhythmic interpretation. Both overall tendencies (means) and the repeatability of these tendencies (covariances) are expressed by these vectors. While we do not discuss the learning problem here, the mean and covariance of the $(\tau_n, \sigma_n)^t$ vectors are be estimated from past rehearsal data using the EM algorithm.

The solo model can be summarized as

$$x_{n+1}^{\text{solo}} = A_n x_n^{\text{solo}} + \xi_n^{\text{solo}} \tag{17}$$

for $n = 0, \ldots, N-1$ where $x_n^{\text{solo}} = (t_n, s_n)^t$, $\xi_n^{\text{solo}} = (\tau_n, \sigma_n)^t$ and A_n is the 2x2 matrix in Eqn. 16. In Eqn. 17 the $\{\xi_n^{\text{solo}}\}$ and x_0^{solo} are mutually independent Gaussian random vectors.

We model the relationship between the true times of solo events $\{s_n\}$ and the observed times estimated by the Listen process as

$$x_n^{\text{obs}} = B x_n^{\text{solo}} + \xi_n^{\text{obs}} \tag{18}$$

where the matrix $B = (1, 0)$ and the $\{\xi_n^{\text{obs}}\}$ are independent 0-mean Gaussian variables with known variances. The $\{x_n^{\text{solo}}\}$ and $\{x_n^{\text{obs}}\}$ variables have a dependency structure expressed in the directed acyclic graph (DAG) of Figure 1 which qualitatively describes Eqns. 17 and 18.

We need to include a collection of variables into our model that account for the accompaniment part. We begin by defining a model for the accompaniment part alone that is completely analogous to the solo model. Specifically, we define a process

$$x_{m+1}^{\text{accom}} = C_m x_m^{\text{accom}} + \xi_m^{\text{accom}}$$

for $m = 0, \ldots, M-1$ where the $\{x_m^{\text{accom}}\}$ are (time,tempo) variables for the accompaniment notes, where x_0^{accom} and the $\{\xi_m^{\text{accom}}\}$ are mutually independent Gaussian vectors that express the accompaniment's rhythmic interpretation, and where the $\{C_m\}$ are matrices analogous to the $\{A_n\}$ of Eqn. 17. The means and covariances of the x_0^{accom} and $\{\xi_m^{\text{accom}}\}$ variables are then learned from actual performances of the accompaniment using the EM algorithm as with solo model. One might think of the x^{accom} process as representing the "practice room" distribution on the accompaniment part — that is, the way the accompaniment plays when issues of synchronizing with the soloist are not relevant.

We then combine our solo and accompaniment models into a joint model containing the variables of both parts. In doing so, the solo and accompaniment models play asymmetric roles since we model the notion that the accompaniment must follow the soloist. To this end we begin with

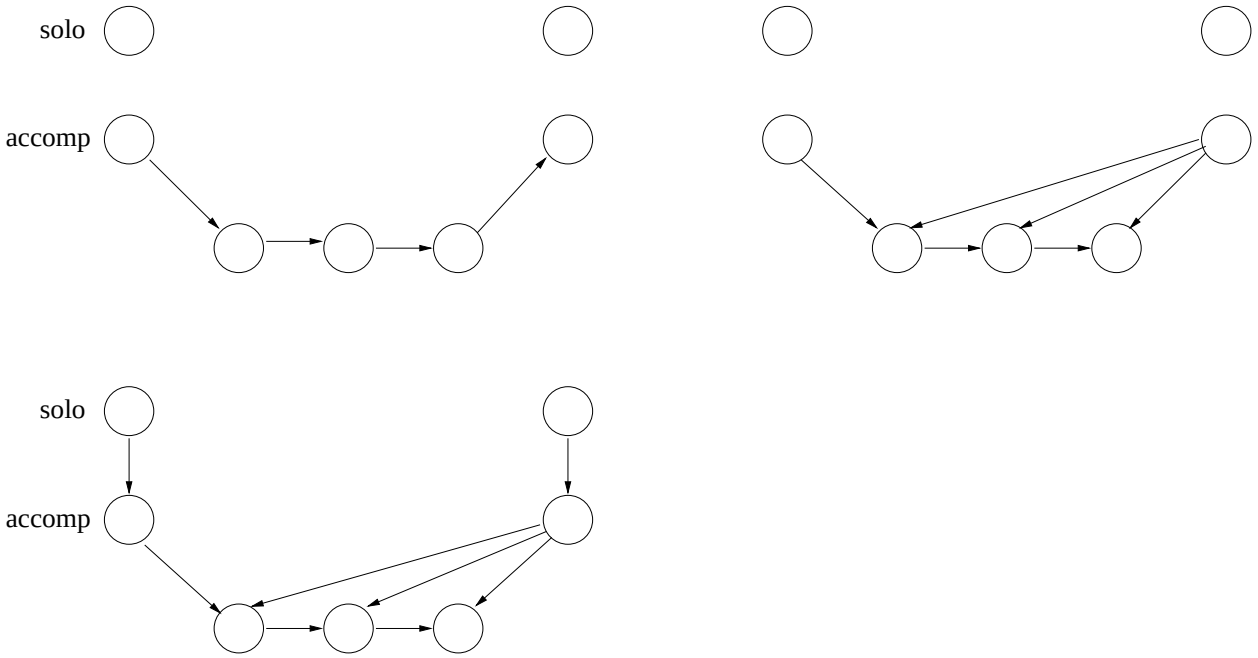

Figure 2: **Upper Left:** A sequence of 5 accompaniment notes, the first and last of which, $x_{m_l}^{\text{accom}}$ and $x_{m_r}^{\text{accom}}$, coincide with the solo notes $x_{n(m_l)}^{\text{solo}}$ and $x_{n(m_r)}^{\text{solo}}$. The conditional distribution of each vector given its predecessor is learned during a training phrase. **Upper Right:** An undirected graph of the same variables used for computing the joint distribution on $x_{m_l}^{\text{accom}}$ and $x_{m_r}^{\text{accom}}$. **Lower Left:** A directed graph showing the dependency structure for the conditional distribution of the $x_{m_l}^{\text{cond}}, \ldots, x_{r_l}^{\text{cond}}$ given $x_{n(m_l)}^{\text{solo}}$ and $x_{n(m_r)}^{\text{solo}}$.

the solo model exactly as it has been trained from examples as in Eqn. 17. We then define the conditional distribution of the accompaniment part *given* the solo part in a way that integrates the rhythmic interpretation of the accompaniment as represented in the x^{accom} process *and* the desire for synchronicity, as follows.

Consider a section of the accompaniment part "sandwiched" between two solo notes as in the upper left panel of Figure 2. For simplicity we assume that m_l and m_r are the indices of the leftmost and rightmost accompaniment notes and that $n(m_l)$ and $n(m_r)$ are the indices of the coincident solo notes of Figure 2. Under the practice room distribution, the accompaniment notes $x_{m_l+1}^{\text{accom}}, \ldots, x_{m_r-1}^{\text{accom}}$ have a conditional distribution given $x_{m_l}^{\text{accom}}$ and $x_{m_r}^{\text{accom}}$ that can be represented as follows.

Since we are conditioning on $x_{m_l}^{\text{accom}}$ and $x_{m_r}^{\text{accom}}$, $x_{m_l+1}^{\text{accom}}$ must depend on both of these variables, thus the arrows leading from $x_{m_l}^{\text{accom}}$ and $x_{m_r}^{\text{accom}}$ to $x_{m_l+1}^{\text{accom}}$ in Figure 2. Then, due to the Markov structure on $x_{m_l+1}^{\text{accom}}, \ldots, x_{m_r-1}^{\text{accom}}$, the conditional distribution of $x_{m_l+2}^{\text{accom}}$ given $x_{m_l}^{\text{accom}}, x_{m_r}^{\text{accom}}$, and $x_{m_l+1}^{\text{accom}}$ only depends on $x_{m_l+1}^{\text{accom}}$ and $x_{m_r}^{\text{accom}}$, thus justifying the arrows leading to $x_{m_l+2}^{\text{accom}}$ in Figure 2. Reasoning similarly, we produce a graphical representation of the conditional distribution of $x_{m_l+1}^{\text{accom}}, \ldots, x_{m_r-1}^{\text{accom}}$ given $x_{m_l}^{\text{accom}}$ and $x_{m_r}^{\text{accom}}$ as in Figure 2. We relate these variables to the coincident solo variables by assuming that

$$\begin{aligned} x_{m_l}^{\text{accom}} &= x_{n(m_l)}^{\text{solo}} + \xi_{m_l}^{\text{link}} \\ x_{m_r}^{\text{accom}} &= x_{n(m_r)}^{\text{solo}} + \xi_{m_r}^{\text{link}} \end{aligned}$$

where $\xi^{\text{link}}_{m_l}$ and $\xi^{\text{link}}_{m_r}$ are independent variables with 0 mean and small covariance. The resulting dependency structure can be seen in the bottom panel of Figure 2.

Situations arise in which accompaniment notes cannot be sandwiched between a pair of coincident solo notes leading to several other cases that employ the basic idea described above. We will not describe these cases here. Figure 3 shows a DAG describing the dependency structure of a model corresponding to the opening measure of the Sinfonia of J. S. Bach's Cantata 12. The 2nd and 1st layers of the graph are the solo process and the output of Listen as described by Eqns 17 and 18. The 3rd layer denotes "phantom" nodes which arise when accompaniment notes are sandwiched between solo notes yet no coincident solo notes exist. The 4th layer shows the accompaniment notes that are coincident with solo notes. The 5th layer shows the sandwiched accompaniment notes. Finally, for each accompaniment vector (the 4th and 5th layers) we define a variable that deterministically "picks off" the time component of the vector. These variable compose the 6th layer of the graph. Only the top and bottom layers in this graph are directly observable.

6.2 Real-time Accompaniment

The above model is expressed in terms of a directed acyclic graph (DAG) and leads to a factorization of the joint density

$$\phi(x) = \prod_{v \in V} f_v(x_v | x_{\text{pa}}(v))$$

where V is the entire collection of variables in our model and $\text{pa}(v)$ are the parents of v in the DAG. In this representation $f_v(x_v|x_{\text{pa}}(v))$ represents the conditional distribution on x_v given its parents in the graph.

Suppose we construct the undirected graph $\mathcal{G}$ by "moralizing" the DAG (dropping the directions of the edges and adding edges between all nodes that share a child), and then triangulating (adding edges until there are no chordless cycles of length greater than three). Then for each $v \in V$ we have $v \bigcup \text{pa}(v) \subset C(v)$ where $C(v) \in \mathcal{C}$ and $\mathcal{C}$ are the cliques of $\mathcal{G}$. Hence $\phi(x)$ can be written as

$$\phi(x) = \prod_{C \in \mathcal{C}} \phi_C(x_C)$$

where $\phi_C(x_C) = \prod_{v:C(v)=C} f_v(x_v|x_{\text{pa}}(v))$. Note that each of the $\phi_C(x_C)$ factors can be represented as a Gaussian potential by first representing the conditional densities as in Section 3.5, extending the potential to C as in Section 3.2, and multiplying the potentials as in Section 3.3.

We now have a representation of the joint density as in Eqn.1. Before our real-time application begins, we transform the representation to one as in Eqn. 2 by setting the separator potentials identically equal to 1 ($\phi_S = (U^0 = \{0\}, U^+ = \{0\}, U^\infty = \Re^{|S|}, \Sigma = 0, S = 0, \mu = 0)$. Next we compute the equilibrium representation by performing the operations of *Collect Evidence* and *Distribute Evidence* with some arbitrarily chosen root clique. On termination of these two procedures, each ϕ_C and ϕ_S represents the marginal distribution of the indicated variables. Note that all of original potential functions are obtained as products of conditional densities whose potential representations involve degeneracies (nontrivial U^∞ spaces) as in Section 3.5.

We are now ready to perform the real-time accompaniment procedure. The methodological key to our real-time accompaniment algorithm is the computation of conditional marginal distributions facilitated by the message-passing algorithm. At any point during the performance some collection of solo notes and accompaniment notes will have been observed. Conditioned on this information we can compute the distribution on the next unplayed accompaniment note. Our

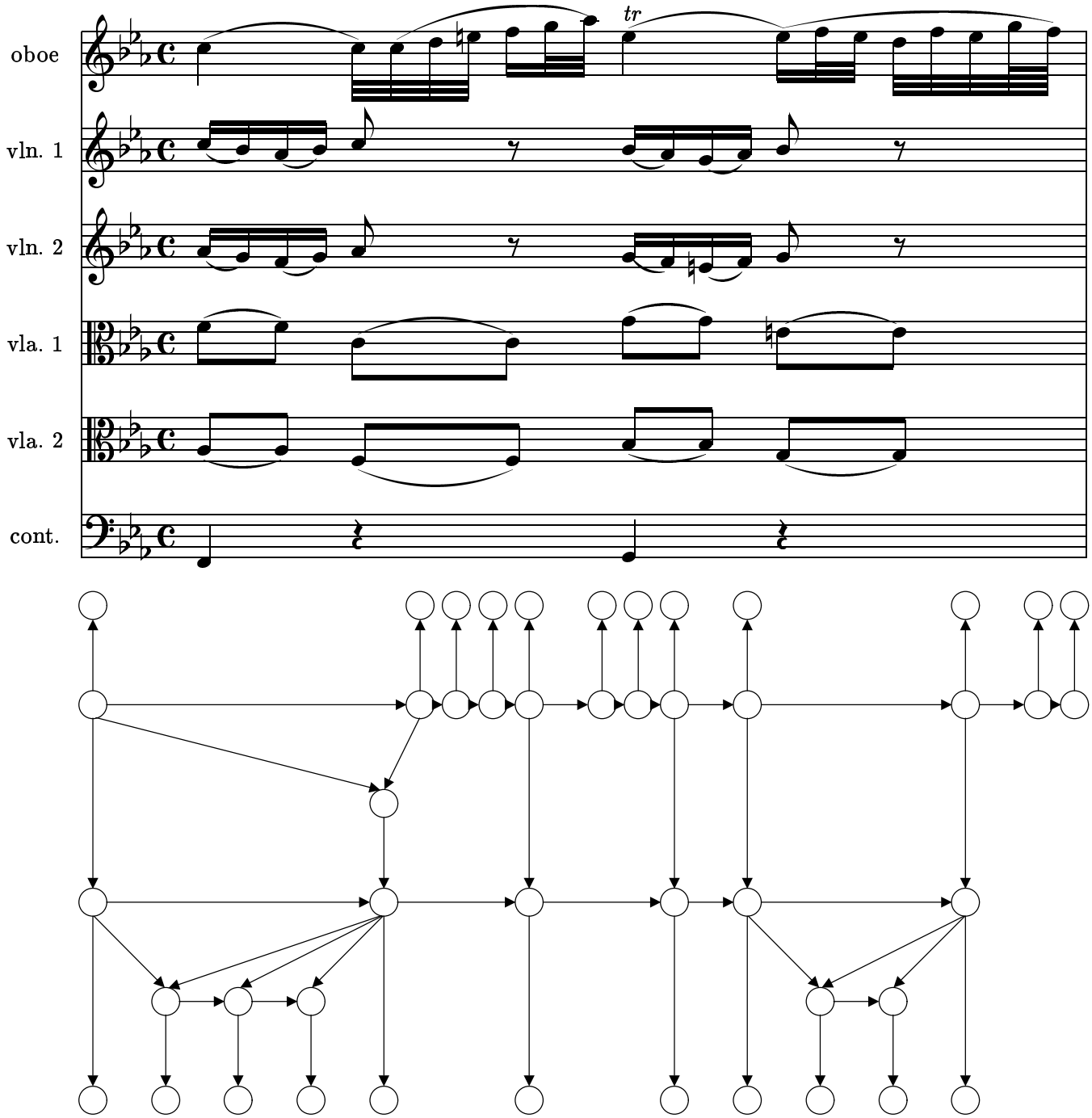

Figure 3: *Top:* The opening measure of the Sinfonia from J.S. Bach's Cantata 12. *Bottom:* The graph corresponding to the first 7/8 of the first measure for this music. The nodes in the 1st (top) layer correspond to the estimated solo note times that come from our real-time analysis. The bottom layer represents the actual times at which accompaniment notes occur. The intermediate layers describe a network of hidden (unobservable) variables including musically meaningful quantities such as local tempo and stress.

potential representation can be modified to represent the conditional distribution, given our observations, by entering each observed variable as in Section 4.2. Then performing *Collect Evidence* with any clique containing the next unplayed accompaniment note as root will leave the root clique potential containing the conditional marginal distribution of the clique variables. Note that the conditioning operation involves multiplication with a degenerate potential (nontrivial U^0 space) as in Section 4.2

Once the marginal of the pending accompaniment note is calculated we schedule the note accordingly. Currently we schedule the note to be played at the posterior mean time given all observed information, however other reasonable choices are possible. Note that this posterior distribution depends on all of the sources of information included in our model: The score information, all currently observed solo and accompaniment note times, the predicted evolution of future solo note times learned during the training phase, and the learned rhythmic interpretation of the accompaniment part.

The initial scheduling of each accompaniment note takes place immediately after the previous accompaniment note is played. It is possible that a solo note will be detected before the pending accompaniment is played; in this event the pending accompaniment note is rescheduled based on the new available information. The pending accompaniment note is rescheduled each time an additional solo note is detected until its current schedule time arrives, at which time it is finally played. In this way our accompaniment makes use of all currently available information.

A demonstration of our real-time accompaniment system in action can be heard on the web page http://fafner.math.umass.edu/music_plus_one.

7 Proofs of Theorems

Theorem 1 *Let* $\phi_1 = (U_1^\infty, U_1^0, U_1^+, \Sigma_1, S_1, \mu_1)$ *and* $\phi_2 = (U_2^\infty, U_2^0, U_2^+, \Sigma_2, S_2, \mu_2)$. *If the system*

$$\begin{pmatrix} P_{U_1^0} \\ P_{U_2^0} \end{pmatrix} x = \begin{pmatrix} P_{U_1^0}\mu_1 \\ P_{U_2^0}\mu_2 \end{pmatrix} \tag{19}$$

is solvable for x *then* $\phi_P = \phi_1\phi_2$ *is given by* $\phi_P = (U_P^\infty, U_P^0, U_P^+, \Sigma_P, S_P, \mu_P)$ *where*

$$\begin{aligned} U_P^\infty &= U_1^\infty \cap U_2^\infty \\ U_P^0 &= U_1^0 \oplus U_2^0 \\ U_P^+ &= (U_P^0 \oplus U_P^\infty)^\perp \\ \Sigma_P &= S_P^- \\ S_P &= P_{U_P^{0\perp}}(S_1 + S_2)P_{U_P^{0\perp}} \\ \mu_P &= \mu_0 + \Sigma_P(S_1(\mu_1 - \mu_0) + S_2(\mu_2 - \mu_0)) \end{aligned}$$

where μ_0 *is the unique solution in* U_P^0 *to Eqn. 19.*

Proof

Clearly Supp(ϕ_P) is the solution set of Eqn. 6. Let μ_0 be the unique solution to Eqn. 6 in

$$N\begin{pmatrix} P_{U_1^0} \\ P_{U_2^0} \end{pmatrix}^\perp = R(P_{U_1^0}\ P_{U_2^0}) = U_1^0 \oplus U_2^0$$

Then Eqn. 6 is equivalent to

$$P_{U_1^0 \oplus U_2^0} x = \mu_0 \tag{20}$$

since both Eqn. 6 and Eqn. 20 have the same null space and both share μ_0 as a solution. Then $\mathrm{Supp}(\phi_P)$ is the solution space of Eqn. 20 so $U_P^0 = U_1^0 \oplus U_2^0$ and

$$P_{U_P^0} \mu_P = \mu_0 \tag{21}$$

The equations for U_P^∞ holds from the definition and hence U_P^+ follows.

S_P must satisfy

$$(x - \mu_P)^t S_P (x - \mu_P) = (x - \mu_1)^t S_1 (x - \mu_1) + (x - \mu_2)^t S_2 (x - \mu_2) + c \tag{22}$$

for all $x \in \mathrm{Supp}(\phi_P) = \{P_{U_P^0} x = \mu_0\}$ and some c. Substituting

$$x = P_{U_P^{0\perp}} y + \mu_0 \tag{23}$$

and equating the quadratic parts of Eqn. 22 gives

$$S_P = P_{U_P^{0\perp}} S_P P_{U_P^{0\perp}} = P_{U_P^{0\perp}} (S_1 + S_2) P_{U_P^{0\perp}}$$

where the leftmost equation follows since S_P can have no component in U_P^0. Using Eqn. 23 and equating the linear parts of Eqn. 22 gives

$$P_{U_P^{0\perp}} S_P (\mu_0 - \mu_P) = P_{U_P^{0\perp}} S_1 (\mu_0 - \mu_1) + P_{U_P^{0\perp}} S_2 (\mu_0 - \mu_2)$$

which will be satisfied if

$$S_P (\mu_0 - \mu_P) = S_1 (\mu_0 - \mu_1) + S_2 (\mu_0 - \mu_2)$$

Since we require $P_{U_P^\infty} \mu_P = 0$ then by Eqn. 21 $\mu_0 - \mu_P \in U_P^+$ so

$$\begin{aligned} \mu_0 - \mu_P &= P_{U_P^+} (\mu_0 - \mu_P) \\ &= \Sigma_P S_P (\mu_0 - \mu_P) \\ &= \Sigma_P (S_1 (\mu_0 - \mu_1) + S_2 (\mu_0 - \mu_2)) \end{aligned}$$

and μ_P is as claimed □

Theorem 2 *Let ϕ_1, ϕ_2 be potentials such that $\phi_1 = \phi_2 \phi_3$ for some potential ϕ_3. Let $\phi_Q = \phi_1 / \phi_2$ where 0/0 = 0.*

$\phi_Q = (U_Q^\infty, U_Q^0, U_Q^+, \Sigma_Q, S_Q, \mu_Q)$ is given by

$$\begin{aligned} U_Q^\infty &= U_1^\infty \\ U_Q^0 &= U_1^0 \\ U_Q^+ &= U_1^+ \\ \Sigma_Q &= S_Q^- \\ S_Q &= P_{U_Q^{0\perp}} (S_1 - S_2) P_{U_Q^{0\perp}} \\ \mu_Q &= \mu_0 + \Sigma_Q (S_1 (\mu_1 - \mu_0) - S_2 (\mu_2 - \mu_0)) \end{aligned}$$

where $\mu_0 = P_{U_1^0} \mu_1$.

Proof

The relationship $\phi_1 = \phi_2 * \phi_3$ guarantees that

$$\begin{aligned} \mathrm{Supp}(\phi_1) &\subseteq \mathrm{Supp}(\phi_2) && (24) \\ U_1^\infty &\subseteq U_2^\infty && (25) \end{aligned}$$

From Eqn. 24 we have $\{\phi_2(x) = 0\} \subseteq \{\phi_1(x) = 0\}$ and the quotient is well-defined. To see that the equations for ϕ_Q are correct we need only verify that

$$\phi_Q \phi_2 = \phi_1 \tag{26}$$

and that we have correctly observed the definition of $0/0$=0.

To verify the Eqn. 26 let $\phi_P = \phi_Q \phi_2$ and observe that Eqn.24 implies

$$U_2^0 \subseteq U_1^0 \tag{27}$$

(and do we need this? $P_{U_2^0}\mu_1 = P_{U_2^0}\mu_2$). Then

$$U_P^\infty = U_Q^\infty \cap U_2^\infty = U_1^\infty \cap U_2^\infty = U_1^\infty$$

by Eqn. 25,

$$U_P^0 = U_Q^0 \oplus U_2^\infty = U_1^0 \oplus U_2^\infty = U_1^0$$

by Eqn. 27, and $U_P^+ = U_1^+$ follows. Then note

$$\begin{aligned} S_P &= P_{U_P^{0\perp}}(P_{U_Q^{0\perp}}(S_1 - S_2)P_{U_Q^{0\perp}} + S_2)P_{U_P^{0\perp}} \\ &= P_{U_1^{0\perp}} S_1 P_{U_1^{0\perp}} \\ &= S_1 \end{aligned}$$

from which $\Sigma_P = \Sigma_1$ follows. Finally, noting that $P_{U_P^0}\mu_P = P_{U_1^0}\mu_1 = \mu_0$,

$$\begin{aligned} \mu_P &= \mu_0 + \Sigma_P(S_Q(\mu_Q - \mu_0) + S_2(\mu_2 - \mu_0)) \\ &= \mu_0 + \Sigma_1(S_Q\Sigma_Q(S_1(\mu_1 - \mu_0) - S_2(\mu_2 - \mu_0)) + S_2(\mu_2 - \mu_0)) \\ &= \mu_0 + \Sigma_1(S_1(\mu_1 - \mu_0) - S_2(\mu_2 - \mu_0)) + S_2(\mu_2 - \mu_0)) \\ &= \mu_0 + \Sigma_1(S_1(\mu_1 - \mu_0)) \\ &= \mu_0 + \mu_1 - \mu_0 \\ &= \mu_1 \end{aligned}$$

where we have used the relation $S^- S = P_{R(S)}$.

To see that the definition $0/0 = 0$ has been observed note that the relations $U_Q^0 = U_1^0$ and $P_{U_Q^0}\mu_Q = P_{U_1^0}\mu_1$ imply that $\mathrm{Supp}(\phi_Q) = \mathrm{Supp}(\phi_1)$ □

Theorem 3 *Suppose two random vectors x and y are related by the equation*

$$y = Ax + \xi$$

where ξ is a Gaussian random vector with potential representation $\xi = (\{0\}, U^0, U^+, \Sigma, S, \mu)$. The conditional distribution of y given x can be represented as a potential $\phi_C(z)$ where $z = \begin{pmatrix} x \\ y \end{pmatrix}$. The components of ϕ_C are given by

$$\begin{aligned} U_C^\infty &= \begin{pmatrix} I \\ A \end{pmatrix} \\ U_C^0 &= \begin{pmatrix} -A^t \\ I \end{pmatrix} U^0 \\ U_C^+ &= (U_C^0 \oplus U_C^\infty)^\perp \\ \Sigma_C &= S_C^- \\ S_C &= (-A\ I)^t S (-A\ I) \\ \mu_C &= P_{U_C^{\infty}{}^\perp} \begin{pmatrix} 0 \\ \mu \end{pmatrix} \end{aligned}$$

Proof

First note since $U_C^\infty \perp U_C^0$, U_C^∞, U_C^0, U_C^+ are mutually orthogonal. Also we have $N(S_C) = U_C^\infty \oplus U_C^0$ since

$$N(-A\ I) = \begin{pmatrix} I \\ A \end{pmatrix} = U_C^\infty$$

and

$$U_C^{0\perp} = \{U^{0t}(-A\ I)x = 0\} = \{(-A\ I)x \in U_0{}^\perp = N(S)^\perp\}$$

gives

$$\begin{aligned} N(S_C)^\perp &= N(-A\ I)^\perp \cap \{(-A\ I)x \in N(S)^\perp\} \\ &= U_C^{\infty\perp} \cap U_C^{0\perp} \end{aligned}$$

Thus

$$N(S_C) = U_C^\infty \oplus U_C^0.$$

Using the relation

$$P_U B x = 0 \Longleftrightarrow U^t B x = 0 \Longleftrightarrow P_{B^t U} x = 0$$

we have

$$\begin{aligned} P_{U^0}(y - Ax) = P_{U^0}\mu &\iff P_{U^0}(-A\ I)z = P_{U^0}(-A\ I) \begin{pmatrix} 0 \\ \mu \end{pmatrix} \\ &\iff P_{U_C^0} z = P_{U_C^0} \begin{pmatrix} 0 \\ \mu \end{pmatrix} \\ &\iff P_{U_C^0} z = P_{U_C^0} P_{U_C^{\infty}{}^\perp} \begin{pmatrix} 0 \\ \mu \end{pmatrix} \\ &\iff P_{U_C^0} z = P_{U_C^0} \mu_C \end{aligned}$$

and

$$\begin{aligned}(y-(Ax+\mu))^t S(y-(Ax+\mu)) &= (z-\begin{pmatrix}0\\ \mu\end{pmatrix})^t(-A\ I)^t S(-A\ I)(z-\begin{pmatrix}0\\ \mu\end{pmatrix})\} \\ &= (z-\begin{pmatrix}0\\ \mu\end{pmatrix})^t S_C(z-\begin{pmatrix}0\\ \mu\end{pmatrix})\} \\ &= (z-\mu_C)^t S_c(z-\mu_C)\end{aligned}$$

since $S_C = P_{U_C^{\infty\perp}} S_C P_{U_C^{\infty\perp}}$. It follows then that

$$\begin{aligned}p(y|x) &= 1_{\{P_{U^0}(y-Ax)=P_{U^0}\mu)\}} \exp\{-\frac{1}{2}(y-(Ax+\mu))^t S(y-(Ax+\mu))\} \\ &= 1_{\{P_{U_C^0} z=P_{U_C^0}\mu_C\}} \exp\{-\frac{1}{2}(z-\mu_C)^t S_c(z-\mu_C)\}\end{aligned}$$

□

Theorem 4 *Let $K \subseteq H \subseteq G$ be subsets of variables with ϕ defined on G. Then if the marginalization operator is defined as in Section 3.1, then*

$$\sum_{G\backslash K} \phi = \sum_{H\backslash K} \sum_{G\backslash H} \phi$$

Proof

Let

$$\begin{aligned}\phi &= (U^\infty, U^0, U^+, \Sigma, S, \mu) \\ \sum_{G\backslash K} \phi &= (U_1^\infty, U_1^0, U_1^+, \Sigma_1, S_1, \mu_1) \\ \sum_{H\backslash K} \sum_{G\backslash H} \phi &= (U_2^\infty, U_2^0, U_2^+, \Sigma_2, S_2, \mu_2)\end{aligned}$$

Then

$$U_2^\infty = I_K I_H U^\infty = I_K U^\infty = U_1^\infty$$

and

$$\begin{aligned}U_2^0 &= (I_K((I_H U^{0\perp})^\perp)^\perp)^\perp \\ &= (I_K I_H U^{0\perp})^\perp \\ &= (I_K U^{0\perp})^\perp \\ &= U_1^0\end{aligned}$$

so we must also have $U_2^+ = U_1^+$.

Let $U_H^\infty = I_H U^\infty$ and $U_K^\infty = I_K U^\infty = I_K U_H^\infty$. The results $\Sigma_1 = \Sigma_2$, $S_1 = S_2$ and $\mu_1 = \mu_2$ will follow immediately once we demonstrate

$$P_{U_K^{\infty\perp}} I_K P_{U_H^{\infty\perp}} I_H = P_{U_K^{\infty\perp}} I_K$$

For convenience of notation we assume that the components of H form the first several components of G and similarly for K and H. Then we can write

$$\begin{aligned} U_H^{\infty\perp} &= \begin{pmatrix} U_K^{\infty\perp} \\ 0 \end{pmatrix} \oplus \begin{pmatrix} 0 \\ U_{H\setminus K}^{\infty\perp} \end{pmatrix} \oplus \left(\left(\begin{pmatrix} U_K^{\infty} \\ 0 \end{pmatrix} \oplus \begin{pmatrix} 0 \\ U_{H\setminus K}^{\infty} \end{pmatrix} \right) \cap U_H^{\infty\perp} \right) \\ &\stackrel{\text{def}}{=} U_1 \oplus U_2 \oplus U_3 \end{aligned}$$

where $U_{H\setminus K}^{\infty} = I_{H\setminus K} U^{\infty}$. This follows since $U_1, U_2 \subseteq U_H^{\infty\perp}$ and U_3 is the orthogonal complement of $U_1 \oplus U_2$ in $U_H^{\infty\perp}$. Any x with components in G can then be written as

$$x = \begin{pmatrix} x_K^{\infty\perp} & + & x_K^{\infty} & + & \tilde{x}_K \\ x_{H\setminus K}^{\infty\perp} & + & x_{H\setminus K}^{\infty} & + & \tilde{x}_{H\setminus K} \\ x_{G\setminus H} & & & & \end{pmatrix}$$

where $x_K^{\infty\perp} \in U_K^{\infty\perp}$, $x_{H\setminus K}^{\infty\perp} \in U_{H\setminus K}^{\infty\perp}$,

$$\begin{pmatrix} x_K^{\infty} \\ x_{H\setminus K}^{\infty} \end{pmatrix} \in U_3$$

$$\begin{pmatrix} \tilde{x}_K \\ \tilde{x}_{H\setminus K} \end{pmatrix} \in U_H^{\infty}$$

$x_{G\setminus H} \in \Re^{|G\setminus H|}$. Then

$$P_{U_K^{\infty}{}^{\perp}} I_K P_{U_H^{\infty}{}^{\perp}} I_H x = P_{U_K^{\infty}{}^{\perp}} I_K \begin{pmatrix} x_K^{\infty\perp} & + & x_K^{\infty} \\ x_{H\setminus K}^{\infty\perp} & + & x_{H\setminus K}^{\infty} \end{pmatrix} = x_K^{\infty\perp}$$

and

$$P_{U_K^{\infty}{}^{\perp}} I_K x = P_{U_K^{\infty}{}^{\perp}} (x_K^{\infty\perp} + x_K^{\infty} + \tilde{x}_K) = x_K^{\infty\perp}$$

since

$$\begin{pmatrix} \tilde{x}_K \\ \tilde{x}_{H\setminus K} \end{pmatrix} \in U_H^{\infty} \Rightarrow \tilde{x}_K \in U_K^{\infty}$$

□

Theorem 5 *Suppose $H \subseteq G$ and let*

$$\phi_1 = \sum_{G\setminus H} (E_G \phi_H) \phi_G = (U_1^{\infty}, U_1^0, U_1^+, \Sigma_1, S_1, \mu_1)$$

and

$$\phi_2 = \phi_H \sum_{G\setminus H} \phi_G = (U_2^{\infty}, U_2^0, U_2^+, \Sigma_2, S_2, \mu_2)$$

Then $U_1^{\infty} = U_2^{\infty}$, $U_1^0 = U_2^0$, and $U_1^+ = U_2^+$.

Proof

For ease of notation assume that the variables in H are the first several variables of G. Then

$$\begin{aligned} U_1^\infty &= I_H\left(\left(\begin{pmatrix} U_H^\infty \\ 0 \end{pmatrix} \oplus \begin{pmatrix} 0 \\ I \end{pmatrix}\right) \cap U_G^\infty\right) \\ &= U_H^\infty \cap I_H U_G^\infty \\ &= U_2^\infty \end{aligned}$$

where I_H is the matrix that extracts the components of H. Also

$$\begin{aligned} U_1^0 &= \left(I_H\left(\begin{pmatrix} U_H^0 \\ 0 \end{pmatrix} \cap U_G^0\right)^\perp\right)^\perp \\ &= \left(I_H\left(\begin{pmatrix} {U_H^0}^\perp \\ 0 \end{pmatrix} \oplus {U_G^0}^\perp\right)\right)^\perp \\ &= ({U_H^0}^\perp \oplus I_H {U_G^0}^\perp)^\perp \\ &= U_H^0 \cap (I_H {U_G^0}^\perp)^\perp \\ &= U_2^0 \end{aligned}$$

$U_1^+ = U_2^+$ follows from the preceding equations. □

References

[1] Lauritzen S. L., (1996), "Graphical Models," Oxford University Press, New York.

[2] Spiegelhalter D., Dawid A. P., Lauritzen S., Cowell R. (1993), "Bayesian Analysis in Expert Systems," *Statistical Science,* Vol. 8, No. 3, pp. 219–283.

[3] Jensen F., (1996), "An Introduction to Bayesian Networks," Springer-Verlag, New York.

[4] Cowell R., (1999), "Probabilistic Networks and Expert Systems," Springer, New York.

[5] Dawid, A. P. (1992), "Applications of a General Propagation Algorithm for Expert Systems," *Statistics and Computing* vol. 2, pp. 25–36.

[6] Shachter R., Kenley, R. (1989), "Gaussian Influence Diagrams," *Management Science,* Vol. 35, No. 5, pp. 527–550.

[7] Howard R., (1971), "Proximal Decision Analysis," *Management Science,* Vol. 17, No. 9, pp. 507–541; reprinted in Howard R., Matheson J., (eds.) *The Principles and Applications of Decision Analysis*, Vol. II, Strategic Decisions Group, Menlo Park, CA, 1984.

[8] Howard R., Matheson, J., (1981), "Influence Diagrams," in Howard R., Matheson J., (eds.), *The Principles and Applications of Decision Analysis*, Vol. II, Strategic Decisions Group, Menlo Park, CA, 1984.

[9] Lauritzen S. (1992), "Propagation of Probabilities, Means, and Variances in Mixed Graphical Association Models," *Journal of the American Statistical Association*, Vol. 87, No. 420, (Theory and Methods), pp. 1098–1108.

[10] Lauritzen S. L., Jensen F. (1999), "Stable Local Computation with Conditional Gaussian Distributions," *Technical Report R-99-2014*, Department of Mathematic Sciences, Aalborg University.

[11] Lauritzen S. (1995), "The EM Algorithm for Graphical Association Models with Missing Data," *Computational Statistics and Data Analysis* vol. 19,pp. 191–201.

[12] Tarjan R. E., Yannakakis M. (1984), "Simple Linear-Time Algorithms to Test Chordality of Graphs, Test Acyclcity of Hypergraphs, and Selectively Reduce Acyclic Hypergraphs," *SIAM Journal on Computing*, vol. 13, 566–79.

[13] Rao C. R., Mitra S. K., (1971), "Generalized Inverse of Matrices and its Applications," John Wiley and Sons, New York.

[14] Graybill, F., (1969), "Matrices with Applications in Statistics," Wadsworth International Group, Belmont, CA.

[15] Shenoy P., Shafer G., (1990), "Axioms for Probability and Belief-Function Propagation," in *Uncertainty in Artificial Intelligence*, 4, eds. Shachter, R, Levitt, T, Kanal L, Lemmer F., Amsterdam: North-Holland pp. 169-198.

[16] Press W., Teukolsky S., Vetterling W., Flannery B., (1992), "Numerical Recipes in C ", 2nd ed. Cambridge University Press, Cambridge.

[17] Raphael C. "A Probabilistic Expert System for Automatic Musical Accompaniment," to appear in *Journal of Computational and Graphical Statistics.*