A Alternative Mixed Graphical Models

It is instructive to compare our class of mixed MRF distributions (5) with the models derived from the marginal distribution \( P(Z) \) and the conditional distribution \( P(Y|Z) \).

Suppose that we model the conditional distribution \( P(Y|Z) \) as in the conditional distribution form of (6). Therefore, this alternative distribution has the same form of conditional distribution \( P(Y|Z) \) as . However, instead of assuming that each node-conditional distribution is drawn from an exponential family, which would then lead to our joint mixed MRF distribution in (5) for \( P(Y, Z) \), we model the random vector \( Z \) separately as following a Markov Random Field (MRF) distribution:

\[
P(Z) = \exp \left\{ \sum_{r' \in V_Z} \theta^{Z}_{rr'} B_Z(Z_{r'}) + \sum_{r' \in V_Z} C_Z(Z_{r'}) + \sum_{(r', t') \in E_Z} \theta^{Z}_{rt'} B_Z(Z_{r'}) B_Z(Z_{t'}) - A_Z(\theta^Z, \theta^{zz}) \right\}.
\]

(10)

Note that the log-partition function \( A_Z(\cdot) \) here is defined as

\[
\log \sum_{Z \in \mathcal{Z}} \exp \left\{ \sum_{r' \in V_Z} \theta^{Z}_{rr'} B_Z(Z_{r'}) + \sum_{r' \in V_Z} C_Z(Z_{r'}) + \sum_{(r', t') \in E_Z} \theta^{Z}_{rt'} B_Z(Z_{r'}) B_Z(Z_{t'}) \right\},
\]

which is dependent only on the parameters \( \theta^Z \) and \( \theta^{zz} \).

Given the specifications of the conditional distribution \( P(Y|Z) \) and the marginal distribution \( P(Z) \), we can then specify the joint distribution simply as \( P(Y, Z) = P(Y|Z)P(Z) \), so that

\[
P(Y, Z; \theta) = \exp \left\{ \sum_{r \in V_Y} \theta^{Y}_{rr} B_Y(Y_r) + \sum_{r' \in V'_Z} \theta^{Z}_{rr'} B_Z(Z_{r'}) + \sum_{(r, t) \in E_Y} \theta^{Z}_{rt} B_Y(Y_r) B_Y(Y_t) + \sum_{(r', t') \in E_Z} \theta^{Z}_{rt'} B_Z(Z_{r'}) B_Z(Z_{t'}) + \sum_{r \in V_Y} C_Y(Y_r) + \sum_{r' \in V_Z} C_Z(Z_{r'}) - A_{Y|Z}(\{\theta_r(Z)\}_{r \in V_Y}, \theta^{yy}) - A_Z(\theta^Z, \theta^{zz}) \right\}.
\]

(11)

Note that this distribution is distinct from our mixed MRF distribution in (5). In particular, the log-partition function of (11) is not \( A_{Y|Z}(\cdot) + A_Z(\cdot) \) as \( A_{Y|Z} \) is a function on random vector \( Z \).

The form of \( P(Y, Z) \) in (11) is thus much more complicated than that in (5) due to the complicated nonlinear term \( A_{Y|Z}(\{\theta_r(Z)\}_{r \in V_Y}, \theta^{yy}) \). On the other hand, an important benefit of this modeling approach is that the conditions for normalizability of (11) can be characterized simply as those on the marginal \( P(Z) \) (10) and those on the conditional \( P(Y|Z) \) (6). In other words, so long as (10) and (6) are well-defined, the joint distribution (11) always exists and is well-defined as well.

B Proof of Theorem 1

This theorem can be understood as the extension of Proposition 2 in (Yang et al., 2012); the only difference here is that we allow the heterogeneous types of node-conditional distributions. We follow the proof policy of that paper: Define \( Q(X) \) as

\[
Q(X) := \log(P(X)/P(0)),
\]

for any \( X = (X_1, \ldots, X_p) \in \mathcal{X}_1 \times \ldots \times \mathcal{X}_p \) where \( 0 \) indicates a zero vector (The number of zeros vary appropriately in the context below). For any \( X \), also denote \( \bar{X}_r := (X_1, \ldots, X_{r-1}, 0, X_{r+1}, \ldots, X_p) \).

Now, consider the following general form for \( Q(X) \):

\[
Q(X) = \sum_{t_1 \in V} X_{t_1} G_{t_1} (X_{t_1}) + \ldots + \sum_{t_1, \ldots, t_k \in V} X_{t_1} \ldots X_{t_k} G_{t_1, \ldots, t_k} (X_{t_1}, \ldots, X_{t_k}),
\]

(12)

since the joint distribution on \( X \) has factors of size \( k \) at most. It can then be seen that

\[
\exp(Q(X) - Q(\bar{X}_r)) = P(X)/P(\bar{X}_r) = \frac{P(X_r|X_1, \ldots, X_{r-1}, X_{r+1}, \ldots, X_p)}{P(0|X_1, \ldots, X_{r-1}, X_{r+1}, \ldots, X_p)},
\]

(13)

where the first equality follows from the definition of \( Q(X) \). Now, consider simplifications of both sides of (13). Given the form of \( Q(X) \) in (12), we have

\[
Q(X) - Q(\bar{X}_r) = \sum_{t_2, \ldots, t_k \in \{2, \ldots, p\}} X_{t_2} \ldots X_{t_k} G_{t_1, t_2, \ldots, t_k} (X_{t_1}, \ldots, X_{t_k}).
\]

(14)

Also, given the exponential family form of the node-conditional distribution specified in the theorem,

\[
\log \frac{P(X_r|X_1, \ldots, X_{r-1}, X_{r+1}, \ldots, X_p)}{P(0|X_1, \ldots, X_{r-1}, X_{r+1}, \ldots, X_p)} = E_r(X_r|B_r(X_r) - B_r(0)) + (C_r(X_r) - C_r(0)).
\]

(15)

Setting \( X_t = 0 \) for all \( t \neq r \) in (13), and using the expressions for the left and right hand sides in (14)
and (15), we obtain,

\[ X_r G_r(X_r) = E_r(0)(B_r(X_r) - B_r(0)) + (C_r(X_r) - C_r(0)). \]

Setting \( X_u = 0 \) for all \( u \not\in \{r, t\} \),

\[ X_r G_r(X_r) + X_r X_t G_{rt}(X_r, X_t) = E_r(0, X_t, 0)(B_r(X_r) - B_r(0)) + (C_r(X_r) - C_r(0)). \]

Combining these two equations yields

\[ X_r X_t G_{rt}(X_r, X_t) = (E_r(0, X_t, 0) - E_r(0))(B_r(X_r) - B_r(0)). \]  \hspace{1cm} (16)

Similarly, from the same reasoning for node \( t \), we have

\[ X_t G_t(X_t) + X_t X_r G_{tr}(X_t, X_r) = E_t(0, X_r, 0)(B_t(X_t) - B_t(0)) + (C_t(X_t) - C_t(0)), \]

and at the same time,

\[ X_r X_t G_{rt}(X_r, X_t) = (E_r(0, X_t, 0) - E_r(0))(B_t(X_t) - B_t(0)). \]  \hspace{1cm} (17)

Therefore, from (16) and (17), we obtain

\[ E_r(0, X_t, 0) - E_r(0) = \frac{E_r(0, X_r, 0) - E_r(0)}{B_t(X_t) - B_t(0)} (B_r(X_r) - B_r(0)). \]  \hspace{1cm} (18)

Since (18) should hold for all possible combinations of \( X_r, X_t \), for any fixed \( X_t \neq 0 \),

\[ E_r(0, X_t, 0) - E_r(0) = \theta_{rt}(B_r(X_r) - B_r(0)). \]  \hspace{1cm} (19)

Plugging (19) back into (17),

\[ X_r X_t G_{rt}(X_r, X_t) = \theta_{rt}(B_r(X_r) - B_r(0))(B_t(X_t) - B_t(0)). \]

More generally, we can show that

\[ X_{t_1} \ldots X_{t_k} G_{t_1, \ldots, t_k}(X_{t_1}, \ldots, X_{t_k}) = \theta_{t_1, \ldots, t_k}(B_{t_1}(X_{t_1}) - B_{t_1}(0)) \ldots (B_{t_k}(X_{t_k}) - B_{t_k}(0)). \]

Thus, the \( k \)-th order factors in the joint distribution as specified in (12) are tensor products of \( (B_r(X_r) - B_r(0)) \), thus proving the statement of the theorem.

### C Proof of Theorem 2

We can simplify start from the definition of the log partition function in the Manichean MRF joint distribution in (5):

\[
A(\theta) = \sum_{Y \in \mathcal{Y}} \exp \left\{ \sum_{r \in \mathcal{V}_r} \theta^{\mu}_r B_r(Y_r) + \sum_{r' \in \mathcal{V}_{r'}} \theta^{\mu'}_{r'} B_{r'}(Z_{r'}) + \right. \\
\left. \sum_{(r,r') \in E_{Y}} \theta^{\mu}_{rt} B_r(Y_r) B_{r'}(Z_{r'}) + \sum_{(r',r) \in E_Z} \theta^{\mu'}_{r't} B_{r'}(Z_{r'}) B_{r}(Z_{r'}) \right\}.
\]

Since (18) should hold for all possible combinations of \( Y_r, X_t \), for any fixed \( Y_t \neq 0 \),

\[
\theta^{\mu}_r B_r(Y_r) + \sum_{r' \in \mathcal{V}_{r'}} \theta^{\mu'}_{r'} B_{r'}(Z_{r'}) + \sum_{(r,r') \in E_{Y}} \theta^{\mu}_{rt} B_r(Y_r) B_{r'}(Z_{r'}) + \sum_{(r',r) \in E_Z} \theta^{\mu'}_{r't} B_{r'}(Z_{r'}) B_{r}(Z_{r'}) = \theta^{\mu}_r B_r(Y_r) + \sum_{r' \in \mathcal{V}_{r'}} \theta^{\mu'}_{r'} B_{r'}(Y_{r'}) + \sum_{r' \in \mathcal{V}_{r'}} \theta^{\mu'}_{r'} B_{r'}(Z_{r'}) B_{r}(Z_{r'}) \]

Similarly this can be represented as

\[
\sum_{Y \in \mathcal{Y}} \exp \left\{ \sum_{r \in \mathcal{V}_r} \theta^{\mu}_r B_r(Y_r) + \sum_{r' \in \mathcal{V}_{r'}} \theta^{\mu'}_{r'} B_{r'}(Y_{r'}) + \sum_{(r,r') \in E_{Y}} \theta^{\mu}_{rt} B_r(Y_r) B_{r'}(Y_{r'}) + \sum_{(r',r) \in E_{Y}} \theta^{\mu'}_{r't} B_{r'}(Y_{r'}) B_{r}(Y_{r'}) \right\}.
\]

Hence, we can conclude as in the statement since the term

\[
\sum_{r \in \mathcal{V}_r} \theta^{\mu}_r B_r(Y_r) + \sum_{r' \in \mathcal{V}_{r'}} \theta^{\mu'}_{r'} B_{r'}(Y_{r'}) + \sum_{(r,r') \in E_{Y}} \theta^{\mu}_{rt} B_r(Y_r) B_{r'}(Y_{r'}) + \sum_{(r',r) \in E_{Y}} \theta^{\mu'}_{r't} B_{r'}(Y_{r'}) B_{r}(Y_{r'}) \]

is the conditional log-partition function \( A_{Y|Z}(\theta(Z), \theta^{\mu}) \) by definition.

### D Proof of Corollary 1

The conditional distribution \( P(Y|Z = z) \) for any particular assignment of the random variables \( Z \) is normalizable by assumption. It can then be shown that the log-partition function of the joint distribution is precisely given by

\[
E_{Z'} \left\{ \sum_{r \in \mathcal{V}_r} \theta^{\mu}_r B_r(Y_r) + \sum_{r' \in \mathcal{V}_{r'}} \theta^{\mu'}_{r'} B_{r'}(Y_{r'}) + \sum_{(r,r') \in E_{Y}} \theta^{\mu}_{rt} B_r(Y_r) B_{r'}(Y_{r'}) + \sum_{(r',r) \in E_{Y}} \theta^{\mu'}_{r't} B_{r'}(Y_{r'}) B_{r}(Y_{r'}) \right\}.
\]

This expression is also finite and well-defined since there are only finitely many configurations of \( Z \).

### E Proof of Theorem 3

Suppose that neither conditions (a) nor (b) are satisfied. Then, either \( X_r \) or \( X_t \) can possibly take values
approaching both $\infty$ and $-\infty$. Also, for some $\alpha, \beta \geq 0$ such that $-C_r(X_r) = O(X_r^\alpha)$ and $-C_t(X_t) = O(X_t^\beta)$, we have $(\alpha - 1)(\beta - 1) < 1$. We will show that under these conditions, the necessary condition for normalizability detailed in Proposition 1 will be violated, that is:

$$C_r(X_r) + \theta_r X_r X_t + C_t(X_t) \geq 0, \quad (20)$$

for sufficiently large $X_r$ and $X_t$, from which we can conclude that the joint (5) is not normalizable. Note that we ignore the node-wise terms $\theta_r X_r$ and $\theta_t X_t$ without loss of generality in our asymptotic argument since they are asymptotically smaller than the quadratic term.

Consider the following sequences of values taken by the random variables $X_r, X_t$, where $X_r = a^\gamma$ and $X_t = a^\delta$ for arbitrary positive $a$ and some fixed positive constants $\gamma$ and $\delta$. We then have $X_r X_t = a^{\gamma+\delta}$ and $X_r^\alpha + X_t^\beta = a^{\alpha\gamma} + a^{\beta\delta}$. As we increase $a$, $X_r$ and $X_t$ will approach infinity, however, if $\gamma + \delta > \max\{\alpha \gamma, \beta \delta\}$, then $C_r(X_r) + \theta_r X_r X_t + C_t(X_t)$ will not be less than or equal to 0; in other words, the necessary condition for normalizability detailed in Proposition 1 will be violated.

**Case 1: $\alpha$ or $\beta$ is less than or equal to 1**

Consider the case where $\alpha \leq 1$. If we simply set $\gamma = \max\{\beta, 1\}$ and $\delta = 1$, then $\gamma + \delta > \max\{\alpha \gamma, \beta \delta\}$, so that the necessary condition for normalizability detailed in Proposition 1 will be violated as discussed above. By symmetry, the same will hold when $\beta \leq 1$. Thus, in this case, (20) always holds.

**Case 2: Both $\alpha$ and $\beta$ is larger than 1** In this case, the condition $\gamma + \delta > \max\{\alpha \gamma, \beta \delta\}$ can be rewritten as $\delta > (\alpha - 1)\gamma$ and $\frac{\gamma}{\beta - 1} > \delta$. Hence, as long as $(\alpha - 1)\gamma < \frac{\gamma}{\beta - 1}$, we can always find $\gamma$ and $\delta$ satisfying $\gamma + \delta > \max\{\alpha \gamma, \beta \delta\}$, so that the necessary condition for normalizability detailed in Proposition 1 will be violated. By symmetry, the same will hold when $\beta \leq 1$. The earlier (case 1) also can be absorbed in this condition $(\alpha - 1)\gamma < \frac{\gamma}{\beta - 1}$, which is equivalent as $(\alpha - 1)(\beta - 1) < 1$.

Therefore, if $(\alpha - 1)(\beta - 1) < 1$, then the condition (20) always holds, so that from Proposition 1, the joint distribution in (5) will not be normalizable.