BAYESIAN MODEL SELECTION FOR LATENT VARIABLE CAUSAL NETWORKS BY SEQUENTIAL MONTE CARLO

by

Mehmet Burak Kurutmaz

B.S., Computer Engineering, Boğaziçi University, 2016

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Computer Engineering

Boğaziçi University

2019

To my family ...

ACKNOWLEDGEMENTS

First and foremost, I would like to thank to my supervisor Prof. Ali Taylan Cemgil for all his support and guidance throughout the course of this research. I will never forget our academic discussions at Kireçburnu, accompanied with uncountably many cups of tea. I also want to thank to my examiners, Prof. Ahmet Celal Cem Say and Assist. Prof. Mohan Ravichandran for the valuable feedback they have provided for this thesis.

I would especially like to thank my dearest friend, academic partner, and rebuttal comrade Melih Barsbey for his invaluable contributions to this work and for the sleepless nights that we worked before many deadlines. I would also like to thank dear members of our BAM group, Assist. Prof. Umut Şimşekli and Assist. Prof. Sinan Yıldırım for their guidance and collaboration. I have always enjoyed doing research with them and learned a lot from their extensive knowledge. During my master's study, I had the opportunity to work with great people like Onür Poyraz, Gökhan Çapan, Caner Türkmen and Emrullah Dar. It was my pleasure to work with them.

I always felt lucky to share the same laboratory with Doğa Siyli, Cihan Camgöz, Alper Bozkurt and Oğulcan Özdemir; and I cannot imagine how the lab would be if they were not there. Additionally, many thanks to all my dear friends and colleagues in Perceptual Intelligence Laboratory for their endless support and valuable friendship, namely, Alp Kındıroğlu, Berkant Kepez, Çağatay Yıldız, Çağlar Hızlı, Çağrı Sofuoğlu, Gizem Ünlü, Gönül Aycı, Hazal Koptagel, İlhan Adıyaman, İlker Gündoğdu, Metehan Doyran, Mine Öğretir, Nihan Karslıoğlu, Orhan Sönmez, Özge Bozal, Semih Akbayrak, Serhan Daniş, Serkan Buğur, Taha Ceritli, Ufuk Can Biçici, Uras Mutlu and Yavuz Nuzumlalı. I would also like to thank Prof. Bülent Sankur, Prof. İlker Birbil, Prof. Lale Akarun, Dr. Suzan Üsküdarlı, Assist. Prof. Murat Çelik, Assist. Prof. Emre Uğur and the faculty members of CMPE for their help and suggestions.

I am very grateful to Boğaziçi University for all the awesome friendships and wonderful memories I had over the course of eight years. I would like to thank my dear friends Mert Çotuk, Umut Gülsün, Mert Yaşin, Can Güler, Çağlayan Aras, Adnan Cengiz, Oğuzhan Karakaya, Evren Civelek, Ahmet Bağlan, Berkay Varçok, Can Kurtan, Cenkay Arapisaoğlu, Mert Cem Taşdemir, Mert İmre, Mert Tiftikçi, Merve Ünlü, Naci Pekçokgüler, Oğuzhan Aydınlı, Rıdvan Sayar, Faruk Kılıç, and Yavuz Öz for sharing all the stressful and happy moments of my graduate years.

I reserved the last part for my closest friends: my parents Şener Kurutmaz and Ömer Kurutmaz, and my brother Kaan Kurutmaz. You have always been with me for my twenty years of education life and have always been supporting me in every way. I have always felt extremely lucky to have a such family. I cannot thank you enough.

Finally, I thank the Scientific and Technological Research Council of Turkey (TÜBİTAK). This thesis has been supported by the M.Sc. scholarship (2210-A) and by the grant 116E580 (FBIMATRIX) from TÜBİTAK.

ABSTRACT

BAYESIAN MODEL SELECTION FOR LATENT VARIABLE CAUSAL NETWORKS BY SEQUENTIAL MONTE CARLO

Inferring the causal structure of several random variables is a challenging task when interventions are not feasible. Presence of latent confounders further increases the difficulty of the problem, and therefore is neglected in the majority of the causal discovery literature. In this thesis, we adopt a Bayesian approach to causal structure learning by building on the assumption of the independence of cause and effect mechanisms. Without any additional assumptions, we reformulate causal structure learning as a Bayesian model selection problem where we compare appropriate graph structures using the marginal likelihood of associated graphs.

In the presence of confounders, marginal likelihood computation is equivalent to scoring Bayesian networks with latent variables, which is known to be computationally intractable. In order to approximate this quantity, we develop a sequential Monte Carlo algorithm that provides an asymptotically unbiased estimator, along with a Variational Bayes algorithm that provides a variational lower bound for the marginal likelihood. We particularly analyze the mixture of linear basis functions model with Gaussian noise, which is a frequently encountered modelling choice in the empirical literature. In this model, statistical independence of parameters renders Markov equivalent graphs distinguishable, and allows the identification of a unique causal graph. We illustrate the performance of our framework in both synthetic and real data sets, focusing on the bivariate case. Our direct approach seems to perform at the level of state of the art causal discovery methods. The generalizability of our approach makes it a promising framework for large scale causal structure learning.

ÖZET

GİZLİ DEĞİŞKENLİ NEDENSEL AĞLARDA PARÇACIK SÜZGECİ İLE BAYESCİ MODEL SEÇİMİ

Rassal değişkenler arasındaki neden sonuç ilişkilerinin çıkarımı, kontrollü deneylerin mümkün olmadığı durumlarda zorlu bir problemdir. Gizli rassal değişkenlerin varlığı problemin zorluğunu daha da arttırır, ve bundan dolayı nedensellik keşfi yazınının çoğunluğunda gizli değişkenler ihmal edilmektedir. Bu tezde nedensel yapı öğrenilmesi problemine, neden-sonuç mekanizmalarının bağımsızlığı varsayımı üzerine inşa edilmiş Bayesci bir yaklaşım getirmekteyiz. Başka herhangi bir varsayıma ihtiyaç duymadan, nedensel yapı öğrenimi problemini, uygun çizge yapılarının marjinal olabilirliklerinin karşılaştırıldığı bir Bayesçi model seçim problemine indirgemekteyiz.

Gizli değişkenlerin varlığında marjinal olabilirlik hesabı, gizli değişkenli Bayes ağlarını puanlamaya eşdeğerdir ve bu problemin genel olarak çözümlenmesinin zor olduğu bilinmektedir. Bu nedenle, marjinal olabilirliğin kestirimi için asimptotik yansız bir kestirici hesaplayan bir parçacık süzgeci algoritması ile birlikte marjinel olabilirliği alttan sınırlandıran bir varyasyonel Bayes yordaması algoritması geliştirmekteyiz. Bu çalışmada, ampirik araştırmalarda sıklıkla karşılaşılan bir modelleme tercihi olan Gauss gürültülü doğrusal taban fonksiyonları karışımı modelini özellikle analiz etmekteyiz. Bu modelde, parametrelerin istatistiksel bağımsızlığı aynı Markov denkliğindeki çizgeleri ayırt edilebilir kılmakta ve yegane nedensel ağın belirlenmesine olanak sağlamaktadır. İki değişkenli durum üzerinden hem yapay hem de gerçek veri setlerinde yürütülen deneyler sonucunda, yaklaşımımızın başarısının en ileri nedensellik keşfi yöntemleriyle aynı seviyede olduğu görülmektedir. Ayrıca yaklaşımımızın genellenebilirliği, onu daha büyük ölçekli nedensel yapı öğrenimi için de umut verici bir çerçeve haline getirmektedir.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

$\times$	Cartesian product
$=^{+}$	Equality up-to an additive constant
$\equiv$	Equivalence
$\propto$	Proportionality
$\mathbb{1}_{\{\cdot\}}$	Indicator function
$a_{n\|r_{\pi(\boldsymbol{x}_n)}}$	Prior shape parameter of $\rho_{n\|r_{\pi(\boldsymbol{x}_n)}}$
$a^*_{n\|r_{\pi(\boldsymbol{x}_n)}}$	Posterior shape parameter of $\rho_{n\|r_{\pi(\boldsymbol{x}_n)}}$
$\hat{a}_{n\|r_{\pi(\boldsymbol{x}_n)}}$	Variational shape parameter of $\rho_{n\|r_{\pi(\boldsymbol{x}_n)}}$
$b_{n\|r_{\pi(\boldsymbol{x}_n)}}$	Prior rate parameter of $\rho_{n\|r_{\pi(\boldsymbol{x}_n)}}$
$b^*_{n\|r_{\pi(\boldsymbol{x}_n)}}$	Posterior rate parameter of $\rho_{n\|r_{\pi(\boldsymbol{x}_n)}}$
$\hat{b}_{n\|r_{\pi(\boldsymbol{x}_n)}}$	Variational rate parameter of $\rho_{n\|r_{\pi(\boldsymbol{x}_n)}}$
$\mathcal{B}_{\mathcal{P}}[\mathcal{Q}]$	Evidence lower bound of VB
$\mathcal{B}[\mathcal{Q}, r_{1:K}^{1:T}]$	Posterior lower bound of DEM
Categorical$(\cdot)$	Categorical distribution
$\det(\cdot)$	Determinant operator
Dirichlet$(\cdot)$	Dirichlet distribution
$\mathsf{E}\{\cdot\}$	Expectation operator
$E_{\mathcal{G}}$	Edge set of the graph $\mathcal{G}$
$\mathsf{E}_{Q(\boldsymbol{X})}\{f(\boldsymbol{X})\}$	Expectation of $f(\boldsymbol{X})$ over the distribution $Q(\boldsymbol{X})$
$\mathcal{G}$	A directed acyclic graph
Gamma$(\cdot,\cdot)$	Gamma distribution
h_t	t^{th} unnormalized target distribution of SMC
I	Identity matrix
K	Number of latent categorical variables
$\mathrm{KL}(\cdot\|\cdot)$	Kullback-Leibler divergence
$m_{n\|r_{\pi(\boldsymbol{x}_n)}}$	Prior mean of $w_{n\|r_{\pi(\boldsymbol{x}_n)}}$
$m^*_{n\|r_{\pi(\boldsymbol{x}_n)}}$	Posterior mean of $w_{n\|r_{\pi(\boldsymbol{x}_n)}}$
$\hat{m}_{n\|r_{\pi(\boldsymbol{x}_n)}}$	Variational mean of $w_{n\|r_{\pi(\boldsymbol{x}_n)}}$

N	Number of observed continuous variables
$[N]$	Set of first N positive integers
$\mathcal{N}(\cdot, \cdot)$	Normal distribution
$\mathcal{NG}(\cdot, \cdot, \cdot, \cdot)$	Normal-Gamma distribution
$p(\boldsymbol{X})$	Marginal distribution of $\boldsymbol{X}$
$p(\boldsymbol{X} \mid \cdot)$	Conditional distribution of $\boldsymbol{X}$
$p(\boldsymbol{X}, \boldsymbol{Y})$	Joint distribution of $\boldsymbol{X}$ and $\boldsymbol{Y}$
$\hat{p}(\cdot \mid \cdot)$	A manipulated distribution
$\mathcal{P}$	Target posterior
$\mathcal{Q}$	Instrumental distribution
q_t	t^{th} proposal distribution of SMC
$\boldsymbol{r}_k$	k^{th} categorical random variable
r_k^t	t^{th} realization of $\boldsymbol{r}_k$
$\boldsymbol{r}_{\pi(\boldsymbol{v})}$	Set of categorical parents of variable $\boldsymbol{v}$
$r_{\pi(\boldsymbol{v})}^t$	t^{th} realization of $\boldsymbol{r}_{\pi(\boldsymbol{v})}$
$r_{1:K}^{1:T,j}$	j^{th} sample of $\boldsymbol{r}_{1:K}^{1:T}$
$\{r_{1:K}^{1:t,j}\}_{j=1,\ldots,J}$	Set of all particles
$\{r_{1:K}^{1:t,j}, \zeta^j\}_{j=1,\ldots,J}$	Set of weighted particles
$\mathcal{R}_k$	Domain of $\boldsymbol{r}_k$
$\lvert\mathcal{R}_k\rvert$	Cardinality of $\boldsymbol{r}_k$
T	Number of samples
$\text{tr}(\cdot)$	Trace operator
$\bar{U}$	Complement of a set U
$V_{\mathcal{G}}$	Vertex set of the graph $\mathcal{G}$
w_n	Set of all regression weights for $\boldsymbol{x}_n$
$w_{n\lvert r_{\pi(\boldsymbol{x}_n)}}$	Regression weights for $\boldsymbol{x}_n$ when $\boldsymbol{r}_{\boldsymbol{\pi(x_n)}} = r_{\pi(\boldsymbol{x}_n)}$
$\boldsymbol{x}_n$	n^{th} continuous random variable
x_n^t	t^{th} realization of $\boldsymbol{x}_n$
$\boldsymbol{x}_{\pi(\boldsymbol{v})}$	Set of continuous parents of variable $\boldsymbol{v}$
$x_{\pi(\boldsymbol{v})}^t$	t^{th} realization of $\boldsymbol{x}_{\pi(\boldsymbol{v})}$
$\mathcal{X}_n$	Domain of $\boldsymbol{x}_n$

$\boldsymbol{X}$	A random variable
X	A particular value of $\boldsymbol{X}$
$\boldsymbol{X} \sim p(\cdot)$	$\boldsymbol{X}$ is sampled from $p(\cdot)$
$\boldsymbol{X} \sim p(\cdot \mid \cdot)$	$\boldsymbol{X}$ is conditionally sampled from $p(\cdot \mid \cdot)$
$\boldsymbol{X} \to \boldsymbol{Y}$	$\boldsymbol{X}$ causes $\boldsymbol{Y}$
$\boldsymbol{X} \leftarrow \boldsymbol{Y}$	$\boldsymbol{Y}$ causes $\boldsymbol{X}$
$\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}$	$\boldsymbol{X}$ is independent from $\boldsymbol{Y}$
$\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$	$\boldsymbol{X}$ is conditionally independent from $\boldsymbol{Y}$ given $\boldsymbol{Z}$
$\gamma_{k\|r_{\pi(\boldsymbol{r}_k)}}$	Prior parameter of $\theta_{k\|r_{\pi(\boldsymbol{r}_k)}}$
$\gamma^*_{k\|r_{\pi(\boldsymbol{r}_k)}}$	Posterior parameter of $\theta_{k\|r_{\pi(\boldsymbol{r}_k)}}$
$\hat{\gamma}_{k\|r_{\pi(\boldsymbol{r}_k)}}$	Variational posterior parameter of $\theta_{k\|r_{\pi(\boldsymbol{r}_k)}}$
Γ	Gamma function
$\boldsymbol{\epsilon}$	Gaussian noise
ζ^j	Weight of j^{th} particle
θ_k	Set of all parameters of $\boldsymbol{r}_k$'s conditional distributions
$\theta_{k\|r_{\pi(\boldsymbol{r}_k)}}$	Parameter of $\boldsymbol{r}_k$'s distribution when $\boldsymbol{r}_{\boldsymbol{\pi}(\boldsymbol{r}_k)} = r_{\pi(\boldsymbol{r}_k)}$
$\hat{\theta}^t$	Variational posterior parameter of $\boldsymbol{r}^t_{1:K}$
$\Lambda_{n\|r_{\pi(\boldsymbol{x}_n)}}$	Prior precision scale matrix of $w_{n\|r_{\pi(\boldsymbol{x}_n)}}$
$\Lambda^*_{n\|r_{\pi(\boldsymbol{x}_n)}}$	Posterior precision scale matrix of $w_{n\|r_{\pi(\boldsymbol{x}_n)}}$
$\hat{\Lambda}_{n\|r_{\pi(\boldsymbol{x}_n)}}$	Variational precision scale matrix of $w_{n\|r_{\pi(\boldsymbol{x}_n)}}$
ν^t	t^{th} step incremental weight
$\nu^{t,j}$	t^{th} step incremental weight of j^{th} particle
$\boldsymbol{\xi}_n$	n^{th} random variable of a Bayesian network
$\boldsymbol{\xi}_{1:N}$	Set $\{\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_N\}$
$\boldsymbol{\xi}_U$	Set of random variables indexed by a set U
$\pi(\boldsymbol{\xi}_n)$	Set of parent indices of $\boldsymbol{\xi}_n$
ρ_n	Set of all precision parameters of $\boldsymbol{x}_n$
$\rho_{n\|r_{\pi(\boldsymbol{x}_n)}}$	Precision parameter of $\boldsymbol{x}_n$ when $\boldsymbol{r}_{\boldsymbol{\pi}(\boldsymbol{x}_n)} = r_{\pi(\boldsymbol{x}_n)}$
ϕ	Basis functions
ψ	Digamma function

LIST OF ACRONYMS/ABBREVIATIONS

ANM	Additive Noise Model
AUC	Area Under Curve
BN	Bayesian Network
CEP	Cause-Effect Pairs
CI	Confidence Interval
CN	Causal Network
ELBO	Evidence Lower Bound
EM	Expectation Maximization
FCI	Fast Causal Inference
FG	Factor Graph
DAG	Directed Acyclic Graph
IGCI	Information Geometric Causal Inference
IS	Importance Sampling
KL	Kullback-Leibler
LiNGAM	Linear Non-Gaussian Acyclic Model
MAP	Maximum A Posteriori
MCMC	Markov Chain Monte Carlo
MH	Metropolis-Hastings
MRF	Markov Random Field
NMF	Nonnegative Matrix Factorization
PC	Peter-Clark Algorithm
PG	Particle Gibbs
PGM	Probabilistic Graphical Model
PMCMC	Particle Markov Chain Monte Carlo
PNL	Post Non-Linear Model
ROC	Receiver Operating Characteristic
SEM	Structural Equation Model
SIS	Sequential Importance Sampling

SMC	Sequential Monte Carlo
SSM	State Space Model
VB	Variational Bayes

1. INTRODUCTION

One of the main objectives of scientific research is discovering the cause-effect relationships among several variables, as causal interpretations inhere deep insights about the operational principles of systems. Inferring causal relationships allows us to interact with our environment and manipulate the outcomes of complex systems under various conditions, which makes it a fundamental element of learning and human level intelligence. Given the current bottlenecks of deep learning, and given increased emphasis on accountability and transparency of methods in machine learning, causality is likely to be a topic of ever increasing importance [1].

The conventional practice for inferring causal relations is *randomized experiments* in which the effects of some phenomena are tested on randomly divided control and experimental groups of subjects. However, the experimental methods may not be applicable to the cases where intervening the natural process is not practical, efficient, moral, or even possible. On top of that, collecting experimental data is an active process that necessitates constant interaction with the environment, whereas abundant and readily available "big data" is often collected by passive observation. On this basis, the question of how observational data can be used for causal inference becomes a critical issue.

Bayesian networks (BNs) are the standard tools in modelling domains that contain several interacting variables [2, 3], which also makes them a natural candidate for causal discovery research. BNs are directed acyclic graph (DAG) based representations of conditional independence relations between random variables. The expressive power of BNs is limited to statistical independence properties, and they do not necessarily imply causal relations. A widely accepted proposition about encoding the causal relationships via probabilistic models is proposed by Spirtes *et al.* (1993), who postulate the *causal Markov condition*: "A random variable is independent of its non-effects given its direct causes". This proposition makes it possible to represent causal relations via DAGs structurally similar to BNs, with the difference that the directed edges encode

truly causal relations. In other words, *causal networks* (CNs) [5] are special BNs where all the directed edges are from cause variables to effect variables.

The essential difference between BNs and CNs is that the latter are unique by the nature of cause-effect relationships, whereas there may exist multiple *Markov equivalent* BNs [6, 7] satisfying the same conditional independence statements. For instance, the networks $\boldsymbol{X} \rightarrow \boldsymbol{Z} \rightarrow \boldsymbol{Y}$ and $\boldsymbol{X} \leftarrow \boldsymbol{Z} \rightarrow \boldsymbol{Y}$ are Markov equivalent but they are not causally equivalent. Hence it may not be possible to uniquely identify a CN solely based on a set of statistical independence statements when there are multiple Markov equivalent BNs. Even so, one might still be able to deduce a subset of cause-effect relationships of interest from conditional independence statements [8]. If in all Markov equivalent BNs satisfying the given conditional independence statements, a variable $\boldsymbol{X}$ is an ancestor of $\boldsymbol{Y}$, then we could automatically infer that $\boldsymbol{X}$ is a cause of $\boldsymbol{Y}$ since one of these BNs has to be the true CN. However, such an inference is not applicable to many problems, including the case of two variables where we cannot distinguish the causal graphs $\boldsymbol{X} \rightarrow \boldsymbol{Y}$ and $\boldsymbol{Y} \rightarrow \boldsymbol{X}$.

In general, we only have access to data that is assumed to be generated from a CN, and not to the conditional independence statements themselves which could allow us to make partial inferences regarding the causal relations. Thus, we need extra information to make inference regarding the underlying CN structure. This extra information could be in the form of additional data obtained through interventions [5] (as in experimental methods), or in the form of additional assumptions on the nature of causal processes.

Intervention, that is setting the values of some random variables explicitly to specific values by an experimenter, is the main causal discovery method. Pearl (2009) formalized this approach by devising *do-calculus*. If the causal structure is known, intervening a set of variables is shown to be equivalent to discarding the directed edges coming from their parents. In this approach, the actual causal network is learned through the additional structural information obtained by various interventions.

The alternative route, when intervention is not practical, is assuming additional structural assumptions about the data generating mechanisms for the cause and effect. A common assumption is the independence of the distributions of the cause and effect variables as they correspond to *independent mechanisms* in nature [9]. We will not go into details of the validity of the independence assumption but sensible justifications are provided in the context of *deterministic causal inference* [10] and *semi-supervised learning* [11]. In addition to mechanism independence, most methods in the causal discovery literature share other certain assumptions. Together with the almost ubiquitous *faithfulness* assumption that states every conditional independence should be entailed by the causal DAG [4], many methods also assume *acyclicity* (i.e. observed variables not affecting each other), *no selection bias* (i.e. data collection is not affected by latent variables), and *causal sufficiency* (i.e. there exists no unobserved variables that affect more than one observed variable). In this thesis, we will mainly focus on the mechanism independence and causal sufficiency assumptions.

Being able to account for latent *confounding* variables is surely a crucial capability for a causal inference algorithm. In many empirical studies, certain influential variables are unrecorded, or latent, since it is impossible to ensure that all common causes are measured for a study. Therefore, it is safe to say that in significant amount of cases the causal sufficiency assumption is unrealistic. Moreover, considering the presence of latent variables would be more favorable, if we take account of the superior expressive power of the latent variable models. For instance, *naive Bayes* model encodes structured conditional independence relationships [12], whereas its marginal is a complete graph which encodes no specific independence relationship. This sounds even worse when we consider the causal deductions about the interventions: In the former model manipulating a variable has no effect on the others, whereas in the latter model manipulating a variable may affect all the other variables. Thus, ignoring confounding variables almost always leads to wrong deductions about the potential effects of manipulations.

Indeed, finding cause effect relationships in the presence of unobserved variables is the chief aim of many scientific inquiries [13]. As important as it is however, causal

inference becomes much more computationally challenging when accounting for latent variables. Several attempts have been made to handle latent variables in the causal context [14–18], some of which are extensions to work already described above.

In this thesis, we propose a Bayesian approach to *causal structure learning* problem, which is identifying the causal network that generated the observed data, in the presence of latent confounders. This problem is considered as a challenging task in multiple ways. In the subsequent sections, we describe these particular challenges and our approach to overcome them. Furthermore, we discuss the similarities and contrasts of our approach to existing body of work concerning causal structure learning.

1.1. Approach and Contributions

In this section, we outline our approach to the causal structure learning problem and state its novelties. Above all, we believe that the following two main criteria are the sine qua non of a causal structure learning framework:

(i) It should be generalizable to the arbitrary graph structures including the ones with latent variables.
(ii) It should be able to identify the underlying CN from a set of hypothetical graph structures.

In order to meet the first criterion, we adopt a Bayesian model selection approach where we score arbitrary causal graph structures by estimating marginal likelihood. Since it is generally intractable to calculate the marginal likelihood in latent variable networks, we develop a sequential Monte Carlo algorithm to calculate unbiased estimations of the marginal likelihood, and we further develop an alternative variational inference algorithm for lower bounding the marginal likelihood.

The fulfillment of the second criterion, namely the identification of the underlying CN, also contains two sub-problems:

(i) The method must be able to distinguish among Markov equivalent graphs.

(ii) The identified network should correspond to the actual causal network.

The key concept constituting our solution to the problem of breaking Markov equivalence is the principle of independent cause-effect mechanisms [17]. In Bayesian modelling, this principle is equivalent to the *mutual independence* of the parameters belonging to cause and effect distributions [19]. In the simplest case, consider the joint distribution of two dependent variables $\boldsymbol{X}$ and $\boldsymbol{Y}$, which can be parameterized with respect to two alternative Markov equivalent factorizations:

$$p(\boldsymbol{X};\theta_1)\ p(\boldsymbol{Y} \mid \boldsymbol{X};\theta_2) = p(\boldsymbol{Y};\hat{\theta}_1)\ p(\boldsymbol{X} \mid \boldsymbol{Y};\hat{\theta}_2) \tag{1.1}$$

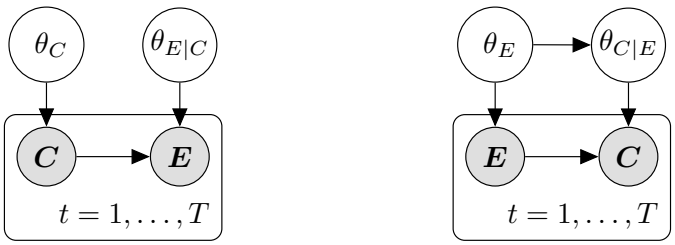
where θ_1, θ_2, $\hat{\theta}_1$ and $\hat{\theta}_2$ are unknown parameters of the conditional distributions. The mechanism independence in the causal model, say $\boldsymbol{X} \to \boldsymbol{Y}$, implies that our knowledge of $p(\boldsymbol{X};\theta_1)$ should not give any additional information about $p(\boldsymbol{Y} \mid \boldsymbol{X};\theta_2)$ or vice versa. This independence condition is satisfied only if the parameters θ_1 and θ_2 are assumed to be independent. However, in the acausal model this is not necessarily true between $\hat{\theta}_1$ and $\hat{\theta}_2$ without further assumptions (e.g. likelihood equivalence [20]), since both distributions in the alternative factorization depend on the same unknown parameters θ_1 and θ_2:

$$p(\boldsymbol{Y};\hat{\theta}_1) = \sum_X p(\boldsymbol{X} = X;\theta_1)\ p(\boldsymbol{Y} \mid \boldsymbol{X} = X;\theta_2) \tag{1.2}$$

$$p(\boldsymbol{X} \mid \boldsymbol{Y};\hat{\theta}_2) = \frac{p(\boldsymbol{X};\theta_1)\ p(\boldsymbol{Y} \mid \boldsymbol{X};\theta_2)}{\sum_X p(\boldsymbol{X} = X;\theta_1)\ p(\boldsymbol{Y} \mid \boldsymbol{X} = X;\theta_2)} \tag{1.3}$$

Therefore, we break the symmetry among Markov equivalent graphs by assuming independent priors on the parameters, which enables us to identify the a-posteriori most likely causal network.

The solution to the second sub-problem, identification of the actual CN, is arguably the most challenging one. Assuming independent priors on the parameters breaks the symmetry among Markov equivalent graphs, but selection of the appropri-



(a) Causal network. (b) Acausal network.

Figure 1.1. Parameter independence of the causal network is not necessarily true for the acausal network.

ate priors to identify true network requires a deep understanding about the philosophy of causality. To circumvent this, we also learn the appropriate priors from the data in a supervised manner. We infer the priors that lead us to the actual causal structures by using labeled data sets.

To sum up briefly, we can list our approach and contributions in seven main headings:

- We treat causal structure learning as a Bayesian model selection problem, that is applicable to general graph structures including *causally insufficient* contexts.
- In order to calculate a tractable estimation of marginal likelihood in the presence of latent confounders, we develop a novel sequential Monte Carlo procedure along with a novel variational inference algorithm.
- When interventions are not feasible, scoring models using marginal likelihood enables us to distinguish among Markov equivalent causal networks through selecting independent priors.
- We also learn those priors from labeled data in a supervised manner.
- We demonstrate that causal direction determination for two observed variables becomes a special case of our model when accounting for the exogenous factors by summarizing them as a single latent confounding variable.
- In addition to improving the inference regarding causal direction, accounting for latent confounders also allows us to detect spurious associations.

- Experiments on synthetic and real-world data demonstrate that we perform on par with the state of the art methods in the literature.

The canonical example of causal discovery is identifying the cause variable in a bivariate setting, yet it is a relatively difficult case. Indeed, conventional causal discovery methods require at least three variables to identify the true causal network [21]. This is because partial or complete inference in a larger data set could sometimes be easier as the existence structural information increases [8]. For instance, consider the causal network structure $\boldsymbol{X} \rightarrow \boldsymbol{Y} \leftarrow \boldsymbol{H}$. If we have access to observations from all of the variables, then we can easily identify this network via conditional independence properties, since there is no other Markov equivalent network encoding the same independence relationships. However, if the variable $\boldsymbol{H}$ is latent, then the conditional independence properties alone do not help distinguish among the bivariate Markov equivalent graphs, namely $\boldsymbol{X} \rightarrow \boldsymbol{Y}$ and $\boldsymbol{X} \leftarrow \boldsymbol{Y}$. Moreover, bivariate scenario can be considered as a building block of the causal discovery problem at large.

Since it would be impossible for a study to measure all related random quantities, we can safely assume that two observed variables, say $\boldsymbol{X}$ and $\boldsymbol{Y}$, actually belong to a larger graph with many unobserved variables. Fortunately, we can sum up the effects of these exogenous variables into a single latent variable and might still infer the causal relation between them. By doing so, the relationship of those two dependent variables boils down to one of three cases [22]:

(i) $\boldsymbol{X}$ causes $\boldsymbol{Y}$
(ii) $\boldsymbol{Y}$ causes $\boldsymbol{X}$
(iii) The relationship is spurious (an unobserved variable causes them both).

We therefore view the modeling of confounding variables indispensable even in the bivariate case, and demonstrate our approach in this setting.

In this thesis, we provide a *variational Bayes* (VB) formulation to our latent variable model to come up with a lower bound to the true marginal likelihood, on top

of the unbiased marginal likelihood estimations of SMC, and demonstrate the accuracy of these two approaches on bivariate scenario. The output of the SMC algorithm also provides an approximation to the posterior distribution of the latent variables, but this distribution may highly be degenerated due to resampling, while the variational posterior distribution in VB does not even belong to the same family as the true posterior. Therefore, we additionally develop a *particle Gibbs* sampler [23] and a *dual Expectation-Maximization* [24] algorithm, in order for inferring the posterior of the latent confounding variables.

1.2. Related Work

Previous attempts at causal discovery research consist of the constraint-based search algorithms such as *PC* [25] and *FCI* [4]. However, these methods exploit the conditional independence relationships observed in the data, and by construction they can identify a network up to a Markov equivalence class of DAGs. Several other methods allow distinguishing among Markov equivalent graphs by restricting the functional relationship between cause and effect variables, so that they are able to choose a single graph as the causal one. The family of such models is called *structural equation models* (SEMs) [5]. In a SEM every effect variable $\boldsymbol{Y}$ is assumed to be a function of its direct causes $\boldsymbol{X}$ and some independent error term $\boldsymbol{\epsilon}$:

$$\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{\epsilon} \qquad\qquad \boldsymbol{Y} \equiv f_\theta(\boldsymbol{X}, \boldsymbol{\epsilon}) \tag{1.4}$$

where f is a member of appropriately constrained functional class, and θ denotes the parameters of the function f. Linear non-Gaussian acyclic model (LiNGAM) [26], additive noise model (ANM) [27], and post-nonlinear model (PNL) [28] are the foremost examples belonging to this family. By restricting the functional relationship between the cause, noise, and effect variables; and/or their distributions, they are able to choose a unique causal graph in most of the cases [29].

Another family of algorithms exploits the idea of independent cause-effect mechanisms, an idea that is long existed in the causal discovery literature [20]. Janzing and

Scholkopf (2010) propose *algorithmic Markov condition* that states the marginal distribution of the cause and the conditional distribution of the effect variables should be *algorithmically independent* which is measured in terms of uncomputable *Kolmogorov complexity* [30]. The class of inference methods based on this assumption is called *Information Geometric Causal Inference* (ICGI). Similar applications, that are developed based on the mechanism independence, include the methods developed by Janzing *et al.* (2012), Zhang *et al.* (2015) and Mooij *et al.* (2016).

Our approach can be considered in the intersection of the work mentioned above in many aspects:

- It is based on the notion of mechanism independence as the methods developed by Janzing *et al.* (2012) and Zhang *et al.* (2015).
- We utilize Bayesian model selection as do the methods described by Stegle *et al.* (2010), Shimizu and Bollen (2014), Zhang *et al.* (2015), and Zhang *et al.* (2016).
- We do not assume causal sufficiency similar to methods developed by Silva *et al.* (2006), Shimizu *et al.* (2009), Janzing *et al.* (2009), Janzing *et al.* (2012), Zhang *et al.* (2015), and Schölkopf *et al.* (2016).
- Our inference algorithm can work on arbitrary graph structures including latent variables similar to algorithm proposed by Spirtes *et al.* (1993), and Shimizu and Bollen (2014).
- We can discriminate among Markov equivalent graphs as the methods proposed by Shimizu *et al.* (2006), Zhang and Hyvärinen (2008), Hoyer *et al.* (2009), Janzing *et al.* (2012), Zhang *et al.* (2015).

A more comprehensive review including the recent developments in causal discovery is written by Spirtes and Zhang (2016). For further information related to aforementioned methods, reader may refer to this seminal review.

1.3. Organization of the Thesis

The rest of the thesis is organized as follows: In Chapter 2, we provide the necessary background information to remind the related concepts that our model is built upon. We next go on to describe our model in Chapter 3 and we introduce our model selection and inference methods in Chapter 4. In Chapter 5 we compare the performance of our algorithms on both synthetic and real data experiments and Chapter 6 concludes the thesis.

2. THEORETICAL BACKGROUND

This chapter is devoted to summarize the theoretical background needed to understand subsequent chapters of the thesis. We first give a brief overview of Bayesian networks and causal networks. In the following section, we give the technical definitions of Markov equivalence and distribution equivalence of statistical models. Finally, we summarize the Bayesian approach of model selection in its full generality.

2.1. Bayesian Networks and Their Causal Extension

Probabilistic graphical models (PGMs) are the graph based representations of joint probability distributions of several random variables. For a set of random variables $\boldsymbol{\xi}_1$, $\boldsymbol{\xi}_2$, ..., $\boldsymbol{\xi}_N$, number of all potential settings that they can influence each other is exponentially large, so simplifying assumptions are inevitable for creating feasible models. Even if these variables were binary, independently specifying all the probability values $p(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_N)$ for each possible setting would require $O(2^N)$ space. In order to enable tractable inference in such systems, graph based representations of joint distributions are of great practical interest. Specifying independence properties of random variables is the fundamental idea of PGMs, which in turn yields structured factorizations of the joint probability distributions.

*Markov random field*s (MRFs) and *Factor Graphs* (FGs) are some of the most-known PGM alternatives [36], which make use of undirected graphs and bipartite graphs respectively. *Bayesian networks* (BNs) [37], on the other hand, are *directed acyclic graph* (DAG) based representations of joint probability distributions, which make them suitable to represent causal relationships as well. In this representation, random variables are denoted by the vertices, while the conditional dependencies among them are denoted by the directed edges.

The technical description of a Bayesian network $\mathcal{G}$ consists of a tuple $(V_\mathcal{G}, E_\mathcal{G})$ where $V_\mathcal{G} = \{\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_N\}$ denotes the set of random variables, and $E_\mathcal{G} \subseteq V_\mathcal{G} \times V_\mathcal{G}$ denotes

the set of directed edges. This graph $\mathcal{G}$ is a DAG, meaning that the set of directed edges is not allowed to have any directed cycle. If there exists a directed edge from ξ_i to ξ_j, i.e. $(\xi_i, \xi_j) \in E_{\mathcal{G}}$, vertex ξ_i is referred to as a *parent* of ξ_j and vertex ξ_j is referred to as a *child* of ξ_i. Similarly, when there is a directed path from ξ_i to ξ_j, vertex ξ_i is referred to as an *ancestor* of ξ_j and vertex ξ_j is referred to as a *descendant* of ξ_i.

A BN encodes conditional independences via *local Markov property*, which states "A variable is conditionally independent of its non-descendants given its parents." [38]. The immediate implication of this property is a certain factorization of the joint probability distribution. Let's define $\pi(\boldsymbol{\xi}_n)$ as the set of indices that the parents of the random variable $\boldsymbol{\xi}_n$ have

$$\pi(\boldsymbol{\xi}_n) \equiv \{i \in [N] \mid (\boldsymbol{\xi}_i, \boldsymbol{\xi}_n) \in E_{\mathcal{G}}\} \tag{2.1}$$

where $[N] \equiv \{1, 2, \ldots, N\}$ is the set of first N positive integers. Also let $\Omega : [N] \to [N]$ be any *topological order* of the variables with respect to the network $\mathcal{G}$, so that whenever $i < j$, the one-to-one mapping Ω satisfies $(\xi_{\Omega(i)}, \xi_{\Omega(j)}) \notin E_{\mathcal{G}}$. Then, the chain rule of probability allows us to factorize the joint probability distribution as follows

$$p(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_N) = \prod_{n=1}^{N} p(\boldsymbol{\xi}_{\Omega(n)} \mid \boldsymbol{\xi}_{\Omega(n+1)}, \ldots, \boldsymbol{\xi}_{\Omega(N)}) \tag{2.2}$$

Since Ω is a topological order; $\boldsymbol{\xi}_{\Omega(n+1)}, \ldots, \boldsymbol{\xi}_{\Omega(N)}$ contain all of the parents but none of the descendants of $\boldsymbol{\xi}_{\Omega(n)}$. As the consequence of local Markov property, $\boldsymbol{\xi}_{\Omega(n)}$ is only dependent to its parents and is independent of the rest of the $\boldsymbol{\xi}_{\Omega(n+1)}, \ldots, \boldsymbol{\xi}_{\Omega(N)}$, i.e.

$$p(\boldsymbol{\xi}_{\Omega(n)} \mid \boldsymbol{\xi}_{\Omega(n+1)}, \ldots, \boldsymbol{\xi}_{\Omega(N)}) = p(\boldsymbol{\xi}_{\Omega(n)} \mid \boldsymbol{\xi}_{\pi(\boldsymbol{\xi}_{\Omega(n)})}) \tag{2.3}$$

where the notation $\boldsymbol{\xi}_U$ refers to a collection of random variables, indexed by the set U as $\boldsymbol{\xi}_U = \{\boldsymbol{\xi}_u \in V_{\mathcal{G}} \mid u \in U\}$ for any $U \subseteq [N]$. By rearranging the terms in Equation 2.2,

we obtain the following factorization for the joint distribution

$$p(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_N) = \prod_{n=1}^{N} p(\boldsymbol{\xi}_n \mid \boldsymbol{\xi}_{\pi(\boldsymbol{\xi}_n)}) \tag{2.4}$$

BNs enable effective computation of the joint probability distribution, which in turn allows arbitrary queries to be answered owing to special factorization in Equation 2.4. Inference problem, that is the computation of marginal distributions conditioned on the event that a subset of random variables $\boldsymbol{\xi}_U$ are observed in a specific state ξ_U, is one of the fundamental uses of the PGMs. In other words, the goal of inference is computing the posterior marginals of form

$$p(\boldsymbol{\xi}_S = \xi_S \mid \boldsymbol{\xi}_U = \xi_U\} = \sum_{i_{\bar{S}}} \frac{p(\boldsymbol{\xi}_{\bar{U}} = \xi_{\bar{U}}, \boldsymbol{\xi}_U = \xi_U)}{p(\boldsymbol{\xi}_U = \xi_U)} \mathbb{1}_{\{\boldsymbol{\xi}_S = \xi_S\}} \tag{2.5}$$

where $S, U \subseteq [N]$ and $\bar{U}, \bar{S}$ denote their complements with respect to $[N]$. Exact calculation of the right hand side expression is feasible depending on the structure of the graph $\mathcal{G}$ and the particular sets U and S. *Junction tree algorithm* [39] is one of the methods for exact inference in graphical models. By moralization and triangulation steps, certain group of nodes are combined into cliques to construct a tree representation, which leads to the following factorization of the joint distribution

$$p(\boldsymbol{\xi}_{1:N} = \xi_{1:N}) = \frac{\prod_{C \in \mathcal{C}} p(\boldsymbol{\xi}_C = \xi_C)}{\prod_{D \in \mathcal{D}} p(\boldsymbol{\xi}_D = \xi_D)}, \tag{2.6}$$

where $\mathcal{C}$ and $\mathcal{D}$ are collection of sets named *cliques* and *separators* satisfying the running intersection property. This factorization allows a propagation algorithm on a tree for efficiently computing desired marginals. In a sense, the junction tree can be viewed as a compact representation of the joint distribution from which desired posterior marginals can still be computed efficiently. We will not further delve into the technical details of the junction tree algorithm here but refer the reader to the relevant literature [38, 40, 41].

Note that on the contrary of conditional independence statements, the dependencies (the lack of independence assertions) deducted from a BN does not necessarily hold for a joint probability distribution. Assume we have bivariate network $\boldsymbol{\xi}_1 \rightarrow \boldsymbol{\xi}_2$, and in this network a specific conditional distribution could be of the form

$$p(\boldsymbol{\xi}_2 \mid \boldsymbol{\xi}_1) = p(\boldsymbol{\xi}_2) \tag{2.7}$$

for which $\boldsymbol{\xi}_1 \perp\!\!\!\perp \boldsymbol{\xi}_2$, i.e. $\boldsymbol{\xi}_1$ are $\boldsymbol{\xi}_2$ unconditionally independent. However, all the independence assertions of the network, which is nothing, still holds for the joint distribution. So, even though the DAG shows a *graphical dependence*, there exists distributions for which this dependence does not hold. If in addition to the independence assertions encoded in the graph, the dependencies are also valid for the distribution, graph and distribution are said to be *faithful* to one another [4].

Causal networks (CNs), similar to BNs, encode the causal relationships via *causal Markov condition*: "For a causally sufficient network, a random variable is statistically independent of its non-effects given its direct causes." [4]. This statistical independence condition for cause-effect relationship, suggests representing causal relations via Bayesian networks where all the directed edges indicate immediate cause-effect relationships. Therefore, all the properties of BNs, including the factorization in Equation 2.4, also apply to CNs. Furthermore, CNs also model various *manipulations* to random variables. Manipulation of a variable $\boldsymbol{\xi}_n$ is defined as perturbing its distribution conditional to the observed values of other variables $\boldsymbol{\xi}_U$. Under causal Markov assumption, manipulating $\boldsymbol{\xi}_i$ to $\hat{p}(\boldsymbol{\xi}_i \mid \boldsymbol{\xi}_U)$, where $\boldsymbol{\xi}_U$ does not contain any descendant of $\boldsymbol{\xi}_i$, corresponds to replacing the original term $p(\boldsymbol{\xi}_i \mid \boldsymbol{\xi}_{\pi(\boldsymbol{\xi}_i)})$ in the factorization of the joint density by the manipulated density $\hat{p}(\boldsymbol{\xi}_i \mid \boldsymbol{\xi}_U)$:

$$\hat{p}(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_N) = \hat{p}(\boldsymbol{\xi}_i \mid \boldsymbol{\xi}_U) \prod_{n \in [N] \setminus \{i\}} p(\boldsymbol{\xi}_n \mid \boldsymbol{\xi}_{\pi(\boldsymbol{\xi}_n)}) \tag{2.8}$$

This update of the joint distribution is called the *manipulation rule*. Thanks to manipulation rule, it is possible to predict the effect of a previously unobserved manipulation

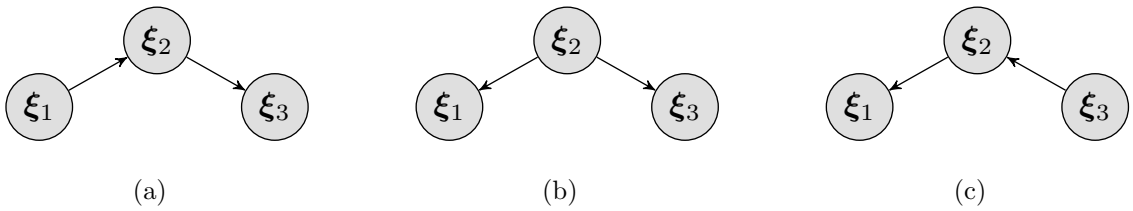


Figure 2.1. Examples of Markov equivalent Bayesian networks.

from the causal DAG structure and unmanipulated density.

2.2. Markov and Distribution Equivalence

In general, two different BNs may encode the same conditional independence assertions. If this is the case, these two graphs are called *Markov equivalent* [42]. For instance, all three networks depicted in Figure 2.1 are Markov equivalent, since they only represent the independence assertion that the variables $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_3$ are conditionally independent given $\boldsymbol{\xi}_2$. Another example of Markov equivalence is the complete DAGs, since each complete DAG represents no conditional independence assertion.

Even though Markov equivalent graphs share the same statistical properties, they state causally distinct assertions, i.e. they model manipulations differently. For instance, manipulating $\boldsymbol{\xi}_1$ influences the distributions of all the variables for the graph in Figure 2.1(a), while it has no effect on the other variables for the rest of the Markov equivalent graphs.

Another equivalence concept for the graphical models is *distribution equivalence*. Assume all the local likelihoods for the variables $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_N$ in graph $\mathcal{G}$ are constrained to a family of distributions $\mathcal{F}$, i.e.

$$\forall n: \qquad p(\boldsymbol{\xi}_n \mid \boldsymbol{\xi}_{\pi(\boldsymbol{\xi}_n)}) \equiv f_n(\boldsymbol{\xi}_n, \boldsymbol{\xi}_{\pi(\boldsymbol{\xi}_n)}) \qquad f_n \in \mathcal{F} \tag{2.9}$$

where $F = \{f_1, \ldots, f_N\}$ is the set of specific likelihood functions of the model. Let $\hat{\mathcal{G}}$ be another graph in consideration, for which the set of parents denoted by $\hat{\pi}(\boldsymbol{\xi}_n)$.

Assume the local likelihoods in this graph are also from the same family $\mathcal{F}$

$$\forall n: \qquad p(\boldsymbol{\xi}_n \mid \boldsymbol{\xi}_{\hat{\pi}(\boldsymbol{\xi}_n)}) \equiv \hat{f}_n(\boldsymbol{\xi}_n, \boldsymbol{\xi}_{\hat{\pi}(\boldsymbol{\xi}_n)}) \qquad \hat{f}_n \in \mathcal{F} \tag{2.10}$$

where $\hat{F} = \{\hat{f}_1, \ldots, \hat{f}_N\}$ denotes the set of likelihood functions for the alternative model. If for every F, there exists an $\hat{F}$ that specifies the same joint distribution, and vice versa, these two models are distribution equivalent with respect to $\mathcal{F}$. In other words, for two graphical models to be distribution equivalent, the following criteria should be satisfied:

$$\forall F, \exists \hat{F}: \qquad \prod_{n=1}^{N} f_n(\boldsymbol{\xi}_n, \boldsymbol{\xi}_{\pi(\boldsymbol{\xi}_n)}) = \prod_{n=1}^{N} \hat{f}_n(\boldsymbol{\xi}_n, \boldsymbol{\xi}_{\hat{\pi}(\boldsymbol{\xi}_n)}) \qquad f_n, \hat{f}_n \in \mathcal{F} \tag{2.11}$$

Distribution equivalence is the special case of Markov equivalence, i.e. distribution equivalence implies Markov equivalence, but the converse does not necessarily hold. There are families where the Markov equivalence directly implies the distribution equivalence [20], e.g. the family of multinomial likelihoods, but this is not valid in general. For instance, Markov equivalent complete graphs for generalized linear-regression model are shown to represent different sets of joint distributions [43]. In the literature, it is often thought that the distribution equivalent causal models are not distinguishable through observational data. In this thesis, however, we show that assigning priors to the family of distributions $\mathcal{F}$ makes distribution equivalent graphs identifiable.

2.3. Structure Learning for Bayesian Networks

Given a set of random variables and a set of conditional independence statements, constructing the underlying directed graphical model (or its Markov equivalent graphs) is a problem with a known solution [21]. However, if we have only access to data $\mathcal{D}$

that is sampled from a graphical model we have to find a graph $\mathcal{G}$ that maximizes:

$$p(\mathcal{G} \mid \mathcal{D}) \propto p(\mathcal{D} \mid \mathcal{G})\ p(\mathcal{G})$$

The prior $p(\mathcal{G})$ can be chosen suitably, for example as uninformative. Then the model posterior becomes proportional to the marginal likelihood, i.e.

$$p(\mathcal{G} \mid \mathcal{D}) \propto p(\mathcal{D} \mid \mathcal{G}) = \int p(\mathcal{D} \mid \theta, \mathcal{G})\ p(\theta \mid \mathcal{G})\ \mathrm{d}\theta \tag{2.12}$$

where θ is the model parameters of a BN. The above integral can be computed in closed form for certain important special cases, in particular with Dirichlet priors $p(\theta \mid \mathcal{G})$ and with no missing values in $\mathcal{D}$ [7, 20]. When searching the underlying graph, two potential computational difficulties may arise: firstly, the number of different graph structures quickly increases with the number of variables, rendering the problem a combinatorial search problem. Secondly, if data are missing or equivalently if the graph has hidden variables, even the computation of the marginal likelihood for a single graph may become quickly intractable. There are a number of methods that allow approximation of this value [44], in Chapter 4 we will describe our chosen methods.

In this thesis, we will deal with relatively small number of alternative graph structures. So only the second problem, approximation of the marginal likelihood becomes a computational obstacle.

3. A MIXTURE OF LINEAR BASIS FUNCTIONS MODEL

For the sake of generality, we will first define our model for general latent causal graph structures, and then present the potential causal relationships between two variables as the instances of it. A general causal graph $\mathcal{G}(V_{\mathcal{G}}, E_{\mathcal{G}})$ is a combination of a vertex set $V_{\mathcal{G}}$, which is the set of observed and latent random variables, and a set of directed edges $E_{\mathcal{G}} \subseteq V_{\mathcal{G}} \times V_{\mathcal{G}}$ where directed edges imply immediate cause-effect relationships between these variables. Let $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \ldots, \boldsymbol{x}_N\} \subseteq V_{\mathcal{G}}$ denote the set of continuous random variables, and $\{\boldsymbol{r}_1 \ldots, \boldsymbol{r}_k, \ldots, \boldsymbol{r}_K\} \subseteq V_{\mathcal{G}}$ denote the discrete latent variables of the network where each $\boldsymbol{x}_n$ and each $\boldsymbol{r}_k$ are defined in the domains $\mathcal{X}_n$ and $\mathcal{R}_k$, respectively. The set of parent vertices of a vertex $\boldsymbol{v} \in V_{\mathcal{G}}$ is denoted by $\pi(\boldsymbol{v})$, while we denote its continuous parents by $\boldsymbol{x}_{\pi(\boldsymbol{v})}$, and discrete parents by $\boldsymbol{r}_{\pi(\boldsymbol{v})}$.

A causal network, unlike Bayesian networks, provides a unique recipe about the nature of the data generating process rather than just specifying the conditional independence properties. Namely, every child variable is assumed to be generated according to the realizations of its parents. By specifying the conditional distributions of the child variables, we end up with a unique causal generative model. For the scope of this thesis, we specify these conditional distributions as follows: we assume *categorical* distributions on the discrete variables $\boldsymbol{r}_{1:K}$ and linear basis functions models with *Gaussian* noise on the continuous variables $\boldsymbol{x}_{1:N}$. Though these choices are by no means mandatory for our framework[1] , we define latent variables as categorical because latent categorical variables are more convenient for empirical studies (these variables might correspond to existence of a gene or having some rare illness), and provides more interpretable results. Although this choice may seem to be restrictive, it is possible to approximate any continuous or discrete distribution in arbitrary precision by increasing the cardinality of a categorical variable. Similarly, linear basis functions models with *Gaussian* noise is also a frequently encountered modelling choice in the empirical literature. By restricting our attention to the graphical structures that do not include a continuous variable as a parent of a categorical variable for inferential convenience

[1] See Chapter 6 for further discussion.

[20], we construct the following generative model for the T independent and identically distributed observations from the network $\mathcal{G}$:

$$\forall k, t: \qquad r_k^t \mid r^t_{\pi(\boldsymbol{r}_k)} \sim \text{Categorical}(\theta_{k|r^t_{\pi(\boldsymbol{r}_k)}}) \tag{3.1}$$

$$\forall n, t: \qquad x_n^t \mid x^t_{\pi(\boldsymbol{x}_n)}, r^t_{\pi(\boldsymbol{x}_n)} \sim \mathcal{N}({w_{n|r^t_{\pi(\boldsymbol{x}_n)}}}^{\mathrm{T}} \phi(x^t_{\pi(\boldsymbol{x}_n)}), {\rho_{n|r^t_{\pi(\boldsymbol{x}_n)}}}^{-1}) \tag{3.2}$$

where $1 \leq t \leq T$, ϕ is an arbitrary basis function with the convention $\phi(\{\}) = 1$, and $\theta_{k|r^t_{\pi(\boldsymbol{r}_k)}}$, $w_{n|r^t_{\pi(\boldsymbol{x}_n)}}$, $\rho_{n|r^t_{\pi(\boldsymbol{x}_n)}}$'s are the parameters of the conditional distributions. Namely, θ_k is the conditional distribution table of $\boldsymbol{r}_k$, w_n is the weights of the basis functions, and ρ_n is the precision parameter of the conditional distribution of $\boldsymbol{x}_n$.

In Bayesian analysis, it is natural to view parameters as random variables [20, 45–47]. One can justify this practice via relating the source of stochasticity with the lack of related information[2] . Since these parameters are unknown, we encode our knowledge by assigning probability distributions to them. As discussed earlier, declaring parameters as random variables simplifies the notion of independent cause-effect mechanisms as follows: Since the conditional distributions are the functions of the parameters, independence of the conditional distributions boils down to the independence of the parameters. Therefore, we complete our generative model by defining independent conjugate prior distributions on the parameters

$$\forall k, r_{\pi(\boldsymbol{r}_k)}: \qquad \theta_{k|r_{\pi(\boldsymbol{r}_k)}} \sim \text{Dirichlet}(\gamma_{k|r_{\pi(\boldsymbol{r}_k)}}) \tag{3.3}$$

$$\forall n, r_{\pi(\boldsymbol{x}_n)}: \qquad w_{n|r_{\pi(\boldsymbol{x}_n)}}, \rho_{n|r_{\pi(\boldsymbol{x}_n)}} \sim \mathcal{NG}(m_{n|r_{\pi(\boldsymbol{x}_n)}}, \Lambda_{n|r_{\pi(\boldsymbol{x}_n)}}, a_{n|r_{\pi(\boldsymbol{x}_n)}}, b_{n|r_{\pi(\boldsymbol{x}_n)}}) \tag{3.4}$$

where $\gamma_{k|r_{\pi(\boldsymbol{r}_k)}}$, $m_{n|r_{\pi(\boldsymbol{x}_n)}}$, $\Lambda_{n|r_{\pi(\boldsymbol{x}_n)}}$, $a_{n|r_{\pi(\boldsymbol{x}_n)}}$, $b_{n|r_{\pi(\boldsymbol{x}_n)}}$ are the prior parameters, i.e. hyperparameters, of our generative model.

[2] Although there may exist some quantum theory experiments that seems the contradict with this view [48].

3.1. Identifiability of Markov Equivalent Graphs

Remember that when constructing a generative model for causal inference, our aim is making Markov equivalent graph structures identifiable. However, the model that is described only by Equations 3.1 and 3.2 is not necessarily identifiable [26, 27]. To be more precise, consider the case where we have two continuous variables and no latent categorical variable, which is equivalent to the following structural equation model:

$$x_1 = w_1(1) + \rho_1^{-1/2}\epsilon_1 \qquad \epsilon_1 \sim \mathcal{N}(0,1)$$
$$x_2 = w_2(1)x_1 + w_2(2) + \rho_2^{-1/2}\epsilon_2 \qquad \epsilon_2 \sim \mathcal{N}(0,1)$$

One can also construct the following equivalent structural equation model in which the dependence structure is reversed:

$$x_2 = w_1(1)w_2(1) + w_2(2) + \hat{\rho}_2^{-1/2}\hat{\epsilon}_2 = \hat{w}_2(1) + \hat{\rho}_2^{-1/2}\hat{\epsilon}_2 \qquad \hat{\epsilon}_2 \sim \mathcal{N}(0,1)$$
$$x_1 = \frac{1}{w_2(1)}x_2 - \frac{w_2(2)}{w_2(1)} + \hat{\rho}_1^{-1/2}\hat{\epsilon}_1 = \hat{w}_1(1)x_2 - \hat{w}_1(2) + \hat{\rho}_1^{-1/2}\hat{\epsilon}_1 \qquad \hat{\epsilon}_1 \sim \mathcal{N}(0,1)$$

These two models are not identifiable with the descriptions above, since they both correspond to linear models with Gaussian noise. However, by assuming priors on the parameters we can break the symmetry and make these Markov equivalent models identifiable. For instance, assuming Gaussian priors on the weights of the first model implies non-Gaussian priors on the second model, which makes these two models *distribution inequivalent* [35]. Moreover, even when two Markov equivalent models are also distribution equivalent, choosing appropriate prior parameters that violate *likelihood equivalence* still makes them identifiable [20]. Indeed, for a model with a parameterization as described, only a very specific choice of priors leads to likelihood equivalence between the Markov equivalent models [47, 49], and we will avoid following such a constraint. Choosing arbitrary priors almost always leads to likelihood inequivalent, hence identifiable models.

3.2. Posterior Distribution of Parameters

The priors in Equation 3.3 and Equation 3.4 are not selected arbitrarily, we choose these priors because they are conjugate to categorical and Gaussian likelihoods. This is because assuming independent conjugate priors results in independent posteriors from the same family due to Dirichlet-Categorical and NormalGamma-Normal conjugacy [7].

$$\begin{aligned}
&\forall k, r_{\pi(\boldsymbol{r}_k)} && \theta_{k|r_{\pi(\boldsymbol{r}_k)}} \mid r_{1:K}^{1:T}, x_{1:N}^{1:T} \sim \text{Dirichlet}(\gamma^*_{k|r_{\pi(\boldsymbol{r}_k)}}) \\
&\forall n, r_{\pi(\boldsymbol{x}_n)} && w_{n|r_{\pi(\boldsymbol{x}_n)}}, \rho_{n|r_{\pi(\boldsymbol{x}_n)}} \mid r_{1:K}^{1:T}, x_{1:N}^{1:T} \sim \mathcal{NG}(m^*_{n|r_{\pi(\boldsymbol{x}_n)}}, \Lambda^*_{n|r_{\pi(\boldsymbol{x}_n)}}, a^*_{n|r_{\pi(\boldsymbol{x}_n)}}, b^*_{n|r_{\pi(\boldsymbol{x}_n)}})
\end{aligned}$$

where $\gamma^*_{k|r_{\pi(\boldsymbol{r}_k)}}$, $m^*_{n|r_{\pi(\boldsymbol{x}_n)}}$, $\Lambda^*_{n|r_{\pi(\boldsymbol{x}_n)}}$, $a^*_{n|r_{\pi(\boldsymbol{x}_n)}}$, $b^*_{n|r_{\pi(\boldsymbol{x}_n)}}$ are the posterior parameters as their form is standard in Bayesian statistics:

$$\begin{aligned}
\gamma^*_{k|r_{\pi(\boldsymbol{r}_k)}}(r_k) &\equiv \gamma_{k|r_{\pi(\boldsymbol{r}_k)}} + \sum_{t=1}^{T} \mathbb{1}_{\left\{r^t_{\pi(\boldsymbol{r}_k)} = r_{\pi(\boldsymbol{r}_k)}\right\}} \mathbb{1}_{\left\{r^t_k = r_k\right\}} \\
\Lambda^*_{n|r_{\pi(\boldsymbol{x}_n)}} &\equiv \Lambda_{n|r_{\pi(\boldsymbol{x}_n)}} + \sum_{t=1}^{T} \mathbb{1}_{\left\{r^t_{\pi(\boldsymbol{x}_n)} = r_{\pi(\boldsymbol{x}_n)}\right\}} \phi(x^t_{\pi(\boldsymbol{x}_n)}) \phi(x^t_{\pi(\boldsymbol{x}_n)})^{\mathrm{T}} \\
m^*_{n|r_{\pi(\boldsymbol{x}_n)}} &\equiv {\Lambda^*_{n|r_{\pi(\boldsymbol{x}_n)}}}^{-1} \Big(\Lambda_{n|r_{\pi(\boldsymbol{x}_n)}} m_{n|r_{\pi(\boldsymbol{x}_n)}} + \sum_{t=1}^{T} \mathbb{1}_{\left\{r^t_{\pi(\boldsymbol{x}_n)} = r_{\pi(\boldsymbol{x}_n)}\right\}} x^t_n \phi(x^t_{\pi(\boldsymbol{x}_n)}) \Big) \\
a^*_{n|r_{\pi(\boldsymbol{x}_n)}} &\equiv a_{n|r_{\pi(\boldsymbol{x}_n)}} + \frac{1}{2} \sum_{t=1}^{T} \mathbb{1}_{\left\{r^t_{\pi(\boldsymbol{x}_n)} = r_{\pi(\boldsymbol{x}_n)}\right\}} \\
b^*_{n|r_{\pi(\boldsymbol{x}_n)}} &\equiv b_{n|r_{\pi(\boldsymbol{x}_n)}} + \frac{1}{2} \Big({m_{n|r_{\pi(\boldsymbol{x}_n)}}}^{\mathrm{T}} \Lambda_{n|r_{\pi(\boldsymbol{x}_n)}} m_{n|r_{\pi(\boldsymbol{x}_n)}} \\
&\qquad - {m^*_{n|r_{\pi(\boldsymbol{x}_n)}}}^{\mathrm{T}} \Lambda^*_{n|r_{\pi(\boldsymbol{x}_n)}} m^*_{n|r_{\pi(\boldsymbol{x}_n)}} + \sum_{t=1}^{T} \mathbb{1}_{\left\{r^t_{\pi(\boldsymbol{x}_n)} = r_{\pi(\boldsymbol{x}_n)}\right\}} (x^t_n)^2 \Big)
\end{aligned}$$

Knowing the closed form expressions of the posterior is of particular importance when developing inference algorithms in Chapter 4. To design proposal distribution for the SMC algorithm in Section 4.1, marginal distribution of $r_{1:K}^{1:T}, x_{1:N}^{1:T}$ needs to be known. As an immediate consequence of the availability of the posterior, a closed form formula

for this marginal distribution is obtained by using Bayes rule:

$$
\begin{aligned}
p(r_{1:K}^{1:T}, x_{1:N}^{1:T}) &= \frac{p(\theta_{1:K})\; p(w_{1:N}, \rho_{1:N})\; p(r_{1:K}^{1:T}, x_{1:N}^{1:T} \mid \theta_{1:K}, \rho_{1:N}, w_{1:N})}{p(\theta_{1:K}, \rho_{1:N}, w_{1:N} \mid r_{1:K}^{1:T}, x_{1:N}^{1:T})} \\
&= \frac{1}{(2\pi)^{NT/2}} \left(\prod_{k=1}^{K} \prod_{r_{\pi(\boldsymbol{r}_k)}} \frac{\Gamma\big(\sum_{r_k} \gamma_{k|r_{\pi(\boldsymbol{r}_k)}}(r_k)\big)}{\Gamma\big(\sum_{r_k} \gamma^*_{k|r_{\pi(\boldsymbol{r}_k)}}(r_k)\big)} \prod_{r_k} \frac{\Gamma\big(\gamma^*_{k|r_{\pi(\boldsymbol{r}_k)}}(r_k)\big)}{\Gamma\big(\gamma_{k|r_{\pi(\boldsymbol{r}_k)}}(r_k)\big)} \right) \\
&\left(\prod_{n=1}^{N} \prod_{r_{\pi(\boldsymbol{x}_n)}} \frac{{b_{n|r_{\pi(\boldsymbol{x}_n)}}}^{a_{n|r_{\pi(\boldsymbol{x}_n)}}}}{{b^*_{n|r_{\pi(\boldsymbol{x}_n)}}}^{a^*_{n|r_{\pi(\boldsymbol{x}_n)}}}} \frac{\Gamma(a^*_{n|r_{\pi(\boldsymbol{x}_n)}})}{\Gamma(a_{n|r_{\pi(\boldsymbol{x}_n)}})} \sqrt{\frac{\det(\Lambda_{n|r_{\pi(\boldsymbol{x}_n)}})}{\det(\Lambda^*_{n|r_{\pi(\boldsymbol{x}_n)}})}} \right)
\end{aligned}
$$

Similarly in Section 4.2, choosing these priors for our model enables us to derive a closed form lower bound for the marginal likelihood. For the sake of brevity, we conclude our discussion here about the inferential properties of our model, but we continue to discuss it in the next chapters. Additionally, most of the technical details omitted in this section are presented in Appendix B including the derivation of the posterior distribution.

3.3. Identifiable Graphical Models for Bivariate Causality

In this section, we define the appropriate graphical structures for causal structure learning in the bivariate case. As we stated in Chapter 1, we do not assume causal sufficiency and allow the existence of possibly many exogenous variables. Luckily, we can combine the effects of exogenous variables into a single latent variable with an arbitrary cardinality. As a result, the relationship between two observable dependent variables $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ boils down to one of three cases due to causal Markov condition [22]:

(i) $\boldsymbol{x}_1$ causes $\boldsymbol{x}_2$,

(ii) $\boldsymbol{x}_2$ causes $\boldsymbol{x}_1$,

(iii) they do not cause each other, but a latent variable $\boldsymbol{r}_1$ causes both of them.

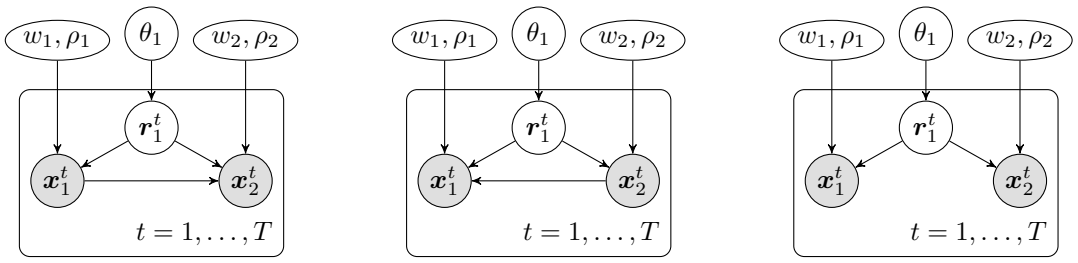


Figure 3.1. Graphical models for causality. (a) $\boldsymbol{x}_1$ causes $\boldsymbol{x}_2$. (b) $\boldsymbol{x}_2$ causes $\boldsymbol{x}_1$. (c) The relationship is spurious.

Associated causal networks corresponding to each of these hypotheses are depicted in Figure 3.1, where latent variable $\boldsymbol{r}_1$ represents the overall effect of the all unobserved variables. For the spurious relationship (Figure 3.1(c)), marginally correlated variables $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ become independent once the latent common cause variable $\boldsymbol{r}_1$ is known. However in direct causal relationships (Figures 3.1(a) and 3.1(b)), even when the latent common cause is known, two variables are still dependent and the direction of cause-effect relationship is implicit in the parameterization of the models.

The identifiability of these models resides in the fact that modelling parameters explicitly as random variables makes these graphs Markov inequivalent. If we were considering only the marginal models of the observed variables, then we would end up with three Markov equivalent graphs. However, including latent variables and independent parameters renders distinctive conditional independence properties for each graph. For instance, when $\boldsymbol{x}_2$ and $\boldsymbol{r}_1$ are known, $\boldsymbol{x}_1$ and the parameters of $\boldsymbol{x}_2$ are dependent only in the case of $\boldsymbol{x}_1 \rightarrow \boldsymbol{x}_2$, or knowing $\boldsymbol{r}_1$ makes $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ independent only if they have a spurious relationship. These distinctive conditional independence properties are the underlying reasons making all of these graphs identifiable.

4. MODEL SELECTION AND INFERENCE

In the previous chapter, we introduced a general modelling framework for representing causal relationships among continuous variables in the presence of categorical confounding variables. At this point, our goal is to derive generic computational procedures that are able to score this class of models in terms of the marginal likelihood of the observed data, and to sample from the posterior of the confounding variables.

4.1. Sequential Monte Carlo

Exact calculation of the marginal likelihood $p(x_{1:N}^{1:T})$ requires summing $p(r_{1:K}^{1:T}, x_{1:N}^{1:T})$ over all possible combinations of $r_{1:K}^{1:T}$, which ends up with a huge sum with exponentially many terms. Therefore, we will pursue another strategy: Instead of exact calculation, we will approximate the marginal likelihood via *Monte Carlo* simulation. A naive attempt to estimate marginal likelihood would be performing *importance sampling* (IS), in which we express the marginal likelihood as an expectation under a proposal distribution $\mathcal{Q}(r_{1:K}^{1:T})$ and estimate this expectation by sampling J independent configurations of $r_{1:K}^{1:T}$

$$
\begin{aligned}
p(x_{1:N}^{1:T}) &= \sum_{r_{1:K}^{1:T}} p(r_{1:K}^{1:T}, x_{1:K}^{1:T}) && (4.1)\\
&= \sum_{r_{1:K}^{1:T}} \frac{p(r_{1:K}^{1:T}, x_{1:K}^{1:T})}{\mathcal{Q}(r_{1:K}^{1:T})} \mathcal{Q}(r_{1:K}^{1:T}) && (4.2)\\
&= \mathsf{E}_{\mathcal{Q}} \left\{ \frac{p(r_{1:K}^{1:T}, x_{1:K}^{1:T})}{\mathcal{Q}(r_{1:K}^{1:T})} \right\} && (4.3)\\
&\approx \frac{1}{J} \sum_{j=1}^{J} \frac{p(r_{1:K}^{1:T,j}, x_{1:K}^{1:T})}{\mathcal{Q}(r_{1:K}^{1:T,j})} && (4.4)
\end{aligned}
$$

However, under poor choices of proposal distribution, the variance of such an estimator would be extremely high relative to its mean, which in turn leads to poor estimations of marginal likelihood. The high variance arises due to the fact that a carelessly chosen proposal is likely to assign small mass to the points where $p(\boldsymbol{r}_{1:K}^{1:T}, x_{1:K}^{1:T})$ is large, hence

vast amount of the terms in the summation are evaluated to zero. In order to avoid unnecessary computation of zero weights in Equation 4.4, we use the more efficient *sequential Monte Carlo* (SMC) technique to construct a sequentially built proposal that eliminates zero weighted samples via *resampling*.

SMC covers a wide range of simulation based techniques to approximate a sequence of posterior distributions and marginal likelihoods, which became very popular in the literature of non-linear non-Gaussian *state space models* (SSMs) [50]. Their parallelizable nature and generality make them a primary choice in the contexts where real-time inference is indispensable [51].

For our model, we utilize a *Rao-Blackwellized* variant of *sequential importance sampling* (SIS) method [50] which is recently adapted to the context of model selection for Dirichlet-Categorical networks [52]. We show that the approach of Cemgil *et al.* (2019) is indeed generalizable to a wide range of conjugate exponential family networks. In order for this to be possible, it is sufficient that the marginal density $p(r_{1:K}^{1:t}, x_{1:N}^{1:t})$ to have a closed form solution, which is the case for the model introduced in Chapter 3. We start to describe our inference algorithm by first forming the following sequence of unnormalized target distributions $h_t(r_{1:K}^{1:t})$ on the latent variables:

$$h_t(r_{1:K}^{1:t}) = p(r_{1:K}^{1:t}, x_{1:N}^{1:t}) \tag{4.5}$$

$$\propto p(r_{1:K}^{1:t} \mid x_{1:N}^{1:t}) \tag{4.6}$$

with the condition that the normalizing constants of h_t's are the marginal likelihoods of the observations. We next construct a proposal distribution $\mathcal{Q}$ for $r_{1:K}^{1:T}$, by choosing filtering distributions as the sequence of proposal distributions $q_1, \ldots, q_T$:

$$\mathcal{Q}(r_{1:K}^{1:T}) \equiv \prod_t q_t(r_{1:K}^t \mid r_{1:K}^{1:t-1}) \tag{4.7}$$

$$q_t(r_{1:K}^t \mid r_{1:K}^{1:t-1}) \equiv p(r_{1:K}^t \mid r_{1:K}^{1:t-1}, x_{1:N}^{1:t}) \tag{4.8}$$

$$= \frac{p(r_{1:K}^{1:t}, x_{1:N}^{1:t})}{\sum_{r_{1:K}} p(\boldsymbol{r}_{1:K}^t = r_{1:K}, r^{1:t-1}, x_{1:N}^{1:t})} \tag{4.9}$$

Note that each proposal distribution admits a form that can be evaluated just by the calculation of the marginal densities in the form $p(r_{1:K}^{1:t}, x_{1:N}^{1:t})$. Then, we express the marginal likelihood as an expectation of a product under $\mathcal{Q}(\boldsymbol{r}_{1:K}^{1:t})$ by using the following telescoping product:

$$p(x_{1:N}^{1:T}) = \sum_{r_{1:K}^{1:T}} p(r_{1:K}^{1:T}, x_{1:N}^{1:T}) = \sum_{r_{1:K}^{1:T}} h_T(r_{1:K}^{1:T}) = \sum_{r_{1:K}^{1:T}} \prod_{t=1}^{T} \frac{h_t(r_{1:K}^{1:t})}{h_{t-1}(r_{1:K}^{1:t-1})} \tag{4.10}$$

$$= \frac{q(r_{1:K}^{1:T})}{q(r_{1:K}^{1:T})} \sum_{r_{1:K}^{1:T}} \prod_{t=1}^{T} \frac{h_t(r_{1:K}^{1:t})}{h_{t-1}(r_{1:K}^{1:t-1})} \tag{4.11}$$

$$= \mathsf{E}_{\mathcal{Q}} \left\{ \prod_{t=1}^{T} \frac{h_t(\boldsymbol{r}_{1:K}^{1:t})}{h_{t-1}(\boldsymbol{r}_{1:K}^{1:t-1})\ q_t(\boldsymbol{r}_1^t \mid \boldsymbol{r}_{1:K}^{1:t-1})} \right\} \tag{4.12}$$

$$= \mathsf{E}_{\mathcal{Q}} \left\{ \prod_{t=1}^{T} p(x_{1:N}^t \mid \boldsymbol{r}_{1:K}^{1:t-1}, x_{1:N}^{1:t-1}) \right\} \tag{4.13}$$

$$= \mathsf{E}_{\mathcal{Q}} \left\{ \prod_{t} \frac{\sum_{r_{1:K}} p(\boldsymbol{r}_{1:K}^t = r_{1:K}, \boldsymbol{r}_{1:K}^{1:t-1}, x_{1:N}^{1:t})}{p(\boldsymbol{r}_{1:K}^{1:t-1}, x_{1:N}^{1:t-1})} \right\} \tag{4.14}$$

Since our goal is to estimate the marginal likelihood, all we need to do is sampling the latent variables $r_{1:K}^{1:T}$ from the distribution $\mathcal{Q}(r_{1:K}^{1:T})$ in a sequential manner, and then dynamically evaluate the *incremental weights*, which are also comprised of marginal densities, and finally multiply them.

$$\nu^t \equiv \frac{h_t(\boldsymbol{r}_{1:K}^{1:t})}{h_{t-1}(\boldsymbol{r}_{1:K}^{1:t-1})\ q_t(\boldsymbol{r}_1^t \mid \boldsymbol{r}_{1:K}^{1:t-1})} \tag{4.15}$$

$$= \frac{\sum_{r_{1:K}} p(\boldsymbol{r}_{1:K}^t = r_{1:K}, r_{1:K}^{1:t-1}, x_{1:N}^{1:t})}{p(r_{1:K}^{1:t-1}, x_{1:N}^{1:t-1})} \tag{4.16}$$

Notice that the product of the incremental weights is an unbiased estimator of the expectation in Equation 4.14. Hence, we can draw independent sequences of $r_{1:K}^{1:T}$, calculate the estimators $\zeta \equiv \prod_{t=1}^{T} \nu^t$ and average them to find the classical *Monte Carlo* estimation. However, the variance of such an estimator also increases as the sample size T increases, since the dimensionality of latent variables depends on T. The canonical approach to reduce the variance of this estimator is sampling these sequences in parallel and performing resampling.

Require: $\mathcal{G}$,$x_{1:N}^{1:T}$

Initialize $Z = 1$

for $j = 1 \ldots, J$ **do**

Initialize $\zeta^j = 1$

end for

for $t = 1, \ldots, T$ **do**

for $j = 1, \ldots, J$ **do**

$q_t(\boldsymbol{r}_{1:K}^t \mid r_{1:K}^{1:t-1,j}) \leftarrow \frac{p(\boldsymbol{r}_{1:K}^t, r_{1:K}^{1:t-1,j}, x_{1:N}^{1:t})}{\sum_{r_{1:K}} p(\boldsymbol{r}_{1:K}^t = r_{1:K}, r^{1:t-1,j}, x_{1:N}^{1:t})}$

Sample $r_{1:K}^{t,j} \sim q_t(\boldsymbol{r}_{1:K}^t \mid r_{1:K}^{1:t-1,j})$

Calculate $\nu^{t,j} \leftarrow \frac{\sum_{r_{1:K}} p(\boldsymbol{r}_{1:K}^t = r_{1:K}, \boldsymbol{r}_{1:K}^{1:t-1,j}, x_{1:N}^{1:t})}{p(\boldsymbol{r}_{1:K}^{1:t-1,j}, x_{1:N}^{1:t-1})}$

Update $\zeta^j \leftarrow \zeta^j \times \nu^{t,j}$

end for

if resampling is on **then**

Update $Z \leftarrow Z \times \frac{1}{J} \sum_{j=1}^{J} \zeta^j$.

$\{r_{1:K}^{1:t,j}, \zeta^j = 1\}_{j=1,\ldots,J} \leftarrow \text{Resample}(\{r_{1:K}^{1:t,j}, \zeta^j\}_{j=1,\ldots,J})$

end if

end for

return weighted samples $\{r_{1:K}^{1:T,j}, \zeta^j\}_{j=1,\ldots,J}$.

return the marginal likelihood estimate Z.

Figure 4.1. SIS-CN: Sequential importance sampling for latent variable causal networks.

In the resampling steps, samples with zero weights are replaced with the highest weighted ones for ensuring the balance of the weight distribution. Here we omit the detailed discussion of resampling steps as these are standard [50]. Another practical consideration that we do not go much into detail is the fact that filtering distributions and importance weights can be calculated by the *junction tree factorization* [53] of the causal network. As a result, a simplified sketch of our inference algorithm *SIS-CN* can be found in Figure 4.1.

As we obtained the marginal densities in closed form, we can apply this inference method to the general (and the bivariate) models described in Chapter 3. The

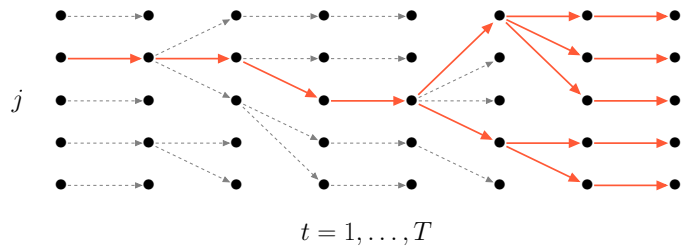


Figure 4.2. An example scheme of successive resampling steps. All the resulting particles at the final step T share the same ancestors in earlier steps.

explicit forms of the formulas related to this method and the aforementioned posterior parameters specifically for the bivariate case can be found in Appendix B.

4.1.1. Particle Gibbs Sampler

In addition to approximating marginal likelihood, one might also be interested in inferring latent confounding variables in practice. A Bayesian way of doing that is sampling these latent variables, and representing the posterior distribution by its samples. Note that the weighted samples of the SMC method provides an IS distribution that is an approximation to the posterior distribution, i.e.

$$p(r_{1:K}^{1:T} \mid x_{1:N}^{1:T}) \approx \sum_{j=1}^{J} \zeta^j \mathbb{1}_{\{r_{1:K}^{1:T} = r_{1:K}^{1:T,j}\}} \tag{4.17}$$

However, the IS distribution of SMC is likely to be poor due to successive resampling steps. Such an approximation of posterior distribution becomes degenerated for the samples $r_{1:K}^t$, especially when $t \ll T$, since most of the distinct values for $r_{1:K}^t$ are destroyed in earlier stages by resampling (see Figure 4.2), and therefore could not be carried to the final step.

A popular choice to sample from a target distribution, which corresponds to posterior distribution in our case, relies on designing *Markov Chain Monte Carlo* (MCMC) kernels [54]. Starting from an arbitrary configuration of $r_{1:K}^{1:T}$, in each step the latent variables are updated successively by an MCMC transition kernel, so that the final dis-

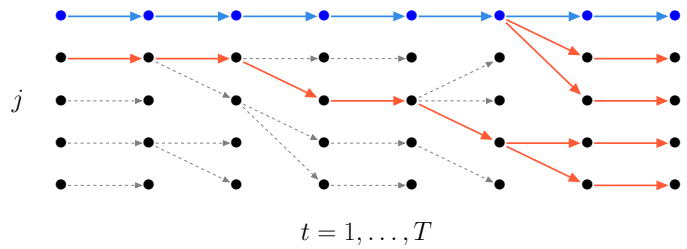


Figure 4.3. An example scheme of successive conditional SMC update steps. A prespecified particle (blue) is preserved in all resampling steps.

tribution of the latent variables converges to the target density of interest. A typical example of MCMC methods, which actually constitutes the majority of the practical methods in MCMC literature, is *Metropolis-Hastings* (MH) algorithm [55]. The MCMC kernel defined by an MH algorithm proposes an update on the latent variables in each step, and this update is either accepted or rejected according to calculated acceptance ratio. However, for this class of MCMC algorithms it is generally difficult to come up with an efficient proposal that ensures fast convergence.

In order to eliminate the drawbacks of the both algorithms, Andrieu *et al.* (2010) melt them in the same pot, and propose *Particle MCMC* (PMCMC) algorithm. PMCMC algorithm employs the IS distribution of SMC as the proposal distribution of an MH kernel where the ratio of marginal likelihood estimations are used to approximate the acceptance ratio. It is shown that such a transition leaves the target density of interest invariant [23].

Andrieu *et al.* (2010) also propose a particle approximation to the *Gibbs* sampler [56], which is the well-known special case of MH algorithm, and they named this algorithm as *Particle Gibss* (PG) sampler. PG samplers rely on a special type of PMCMC update which is called the *conditional SMC* update. Unlike the resampling steps in SMC algorithm, conditional SMC update preserve a prespecified particle in all resampling steps, say $r_{1:K}^{1:T,1}$, while the rest of particles are resampled as before (see Figure 4.3). At the end of each iteration, this prespecified particle is replaced by a randomly selected particle.

Require: $\mathcal{G}$,$x_{1:N}^{1:T}$

Initialize $r_{1:K}^{1:T,1}$

for $j = 1 \ldots, J$ **do**

Initialize $\zeta^j = 1$

end for

repeat

for $t = 1, \ldots, T$ **do**

for $j = 1, \ldots, J$ **do**

$q_t(\boldsymbol{r}_{1:K}^t \mid r_{1:K}^{1:t-1,j}) \leftarrow \frac{p(\boldsymbol{r}_{1:K}^t, r_{1:K}^{1:t-1,j}, x_{1:N}^{1:t})}{\sum_{r_{1:K}} p(\boldsymbol{r}_{1:K}^t = r_{1:K}, r^{1:t-1,j}, x_{1:N}^{1:t})}$

if $j \neq 1$ **then**

Sample $r_{1:K}^{t,j} \sim q_t(\boldsymbol{r}_{1:K}^t \mid r_{1:K}^{1:t-1,j})$

end if

Calculate $\nu^{t,j} \leftarrow \frac{\sum_{r_{1:K}} p(\boldsymbol{r}_{1:K}^t = r_{1:K}, \boldsymbol{r}_{1:K}^{1:t-1,j}, x_{1:N}^{1:t})}{p(\boldsymbol{r}_{1:K}^{1:t-1,j}, x_{1:N}^{1:t-1})}$

Update $\zeta^j \leftarrow \zeta^j \times \nu^{t,j}$

end for

$\{r_{1:K}^{1:t,1}, \zeta^1\} \leftarrow \{r_{1:K}^{1:t,1}, 1\}$

$\{r_{1:K}^{1:t,j}, \zeta^j = 1\}_{j=2,\ldots,J} \leftarrow \text{Resample}(\{r_{1:K}^{1:t,j}, \zeta^j\}_{j=1,\ldots,J})$

end for

Sample $r_{1:K}^{1:T,1} \sim \text{Categorical}(\{r_{1:K}^{1:T,j}\}_{j=1,\ldots,J}; \zeta^{1:J})$

until convergence

return $r_{1:K}^{1:T,1}$.

Figure 4.4. PG-CN: Particle Gibbs sampler for latent variable causal networks.

To sum up, we sample from the posterior distribution to perform inference regarding the latent confounders $\boldsymbol{r}_{1:K}^{1:T}$. Given the difficulties of sampling from high dimensional intractable posteriors, we adapt the PG framework to our causal discovery problem by slightly modifying the resampling step of *SIS-CN* algorithm. We present the sketch of resulting PG sampler in Figure 4.4.

4.2. Variational Inference

An alternative methodology for approximating marginal likelihood in latent variable models is *Variational Bayesian inference* (VB), which is a well known technique in the context of *probabilistic topic modeling* [57, 58] and nonnegative matrix factorization (NMF) models [59]. In this section, for the sake of completeness, we develop variational algorithms for the causal structure learning problem. The derivation is technical but straightforward, so we save most of the technical details for Appendix C, and here we mainly state the results.

VB [60] is a technique where an intractable posterior distribution $\mathcal{P}$ is approximated by a variational distribution $\mathcal{Q}$ via minimizing *Kullback-Leibler* divergence $\mathrm{KL}(\mathcal{Q}||\mathcal{P})$. In the context of Bayesian model selection, minimization of the $\mathrm{KL}(\mathcal{Q}||\mathcal{P})$ corresponds to establishing a tight lower bound for the marginal log-likelihood, which we refer to as *evidence lower bound* (ELBO). This correspondence is due to the following decomposition of marginal log-likelihood

$$\log p(x_{1:N}^{1:T}) = \mathsf{E}_{\mathcal{Q}}\left\{\log \frac{\mathcal{Q}(\boldsymbol{r}_{1:K}^{1:T}, \boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N})}{p(\boldsymbol{r}_{1:K}^{1:T}, \boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N} \mid x_{1:N}^{1:T})}\right\} \tag{4.18}$$

$$+ \mathsf{E}_{\mathcal{Q}}\left\{\log \frac{p(\boldsymbol{r}_{1:K}^{1:T}, x_{1:N}^{1:T}, \boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N})}{\mathcal{Q}(\boldsymbol{r}_{1:K}^{1:T}, \boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N})}\right\} \tag{4.19}$$

$$\equiv \mathrm{KL}(\mathcal{Q}\|\mathcal{P}) + \mathcal{B}_{\mathcal{P}}[\mathcal{Q}] \tag{4.20}$$

$$\geq \mathcal{B}_{\mathcal{P}}[\mathcal{Q}] \tag{4.21}$$

where $\mathcal{P} = p(\boldsymbol{r}_{1:K}^{1:T}, \boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N} \mid x_{1:N}^{1:T})$ is the full posterior distribution, the RHS expression in Equation 4.18 is equal to $\mathrm{KL}(\mathcal{Q}||\mathcal{P})$, and the expression in Equation 4.19 is ELBO which is denoted by $\mathcal{B}_{\mathcal{P}}[\mathcal{Q}]$. The inequality in Equation 4.21 follows from the fact that KL divergence is always nonnegative, which can be shown by *Jensen's inequality* [61], since it can be written as an expectation of a convex function:

$$\mathrm{KL}(\mathcal{Q}\|\mathcal{P}) \equiv \mathsf{E}_{\mathcal{Q}}\left\{\log \frac{\mathcal{Q}}{\mathcal{P}}\right\} = \mathsf{E}_{\mathcal{Q}}\left\{-\log \frac{\mathcal{P}}{\mathcal{Q}}\right\} \geq -\log \mathsf{E}_{\mathcal{Q}}\left\{\frac{\mathcal{P}}{\mathcal{Q}}\right\} = 0 \tag{4.22}$$

In a typical scenario of VB, variational distribution $\mathcal{Q}$ is assumed to be a member of a restricted family of distributions. In its most common form, also known as *mean-field* approximation, $\mathcal{Q}$ is assumed to factorize over some partition of the latent variables, in a way that is reminiscent to a rank-one approximation in the space of distributions

$$\mathcal{Q}(\boldsymbol{r}_{1:K}^{1:T}, \boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N}) = q(\boldsymbol{r}_{1:K}^{1:T})\ q(\boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N}) \tag{4.23}$$

ELBO is then maximized with respect to $\mathcal{Q}$ which is restricted to the class of factorized distributions. Maximization of ELBO serves for two purposes:

(i) ELBO is used in model selection as a substitute of the intractable marginal log-likelihood. The higher the ELBO is the better it approximates the marginal log-likelihood as a lower bound.

(ii) As the marginal log-likelihood does not depend on $\mathcal{Q}$, maximization of ELBO minimizes $\mathrm{KL}(\mathcal{Q}||\mathcal{P})$, which in turn provides a better approximation to the posterior distribution.

To maximize ELBO, i.e. to minimize the $\mathrm{KL}(\mathcal{Q}||\mathcal{P})$, we can employ a coordinate ascent algorithm on the distributions $q(\boldsymbol{r}_{1:K}^{1:T})$ and $q(\boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N})$. By solving a set of *Euler-Lagrange* equations [62], we derive a fixed point iteration algorithm where the updates are in the form of

$$q(\boldsymbol{r}_{1:K}^{1:T}) \propto \exp\big(\mathsf{E}_{q(\boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N})} \left\{\log p(\boldsymbol{r}_{1:K}^{1:T}, x_{1:N}^{1:T}, \boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N})\right\}\big) \tag{4.24}$$

$$q(\boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N}) \propto \exp\big(\mathsf{E}_{q(\boldsymbol{r}_{1:K}^{1:T})} \left\{\log p(\boldsymbol{r}_{1:K}^{1:T}, x_{1:N}^{1:T}, \boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N})\right\}\big) \tag{4.25}$$

and explicit evaluation of the equations above implies the following set of further factorized marginal variational distributions

$$q(\boldsymbol{r}_{1:K}^{1:T}) = \prod_{t=1}^{T} q(\boldsymbol{r}_{1:K}^{t}) \tag{4.26}$$

$$q(\boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N}) = \Big(\prod_{k=1}^{K} \prod_{r_{\pi(\boldsymbol{r}_k)}} q(\boldsymbol{\theta}_{k|r_{\pi(\boldsymbol{r}_k)}})\Big)\Big(\prod_{n=1}^{N} \prod_{r_{\pi(\boldsymbol{x}_n)}} q(\boldsymbol{w}_{n|r_{\pi(\boldsymbol{x}_n)}}, \boldsymbol{\rho}_{n|r_{\pi(\boldsymbol{x}_n)}})\Big) \tag{4.27}$$

where each individual factor is sometimes called *variational posterior*. Due to conjugacy, they also belong to the same family as the posterior

$$\begin{aligned}
\forall t \in [T]: \quad & q(\boldsymbol{r}_{1:K}^{t}) = \text{Categorical}(\boldsymbol{r}_{1:K}^{t}; \hat{\theta}^{t}) \\
\forall k, r_{\pi(\boldsymbol{r}_k)}: \quad & q(\boldsymbol{\theta}_{k|r_{\pi(\boldsymbol{r}_k)}}) = \text{Dirichlet}(\boldsymbol{\theta}_{k|r_{\pi(\boldsymbol{r}_k)}}; \hat{\gamma}_{k|r_{\pi(\boldsymbol{r}_k)}}) \\
\forall n, r_{\pi(\boldsymbol{x}_n)}: \quad & q(\boldsymbol{w}_{n|r_{\pi(\boldsymbol{x}_n)}}, \boldsymbol{\rho}_{n|r_{\pi(\boldsymbol{x}_n)}}) = \mathcal{NG}(\hat{m}_{n|r_{\pi(\boldsymbol{x}_n)}}, \hat{\Lambda}_{n|r_{\pi(\boldsymbol{x}_n)}}, \hat{a}_{n|r_{\pi(\boldsymbol{x}_n)}}, \hat{b}_{n|r_{\pi(\boldsymbol{x}_n)}})
\end{aligned}$$

Here $\hat{\theta}^t$, $\hat{\gamma}_{k|r_{\pi(\boldsymbol{r}_k)}}$, $\hat{m}_{n|r_{\pi(\boldsymbol{x}_n)}}$, $\hat{\Lambda}_{n|r_{\pi(\boldsymbol{x}_n)}}$, $\hat{a}_{n|r_{\pi(\boldsymbol{x}_n)}}$, $\hat{b}_{n|r_{\pi(\boldsymbol{x}_n)}}$ represent the *variational parameters*, and they have algebraically similar forms as the posterior parameters.

$$\begin{aligned}
\log \hat{\theta}^t(r_{1:K}^t) =^{+}\ & \frac{1}{2}\sum_{n=1}^{N} \mathsf{E}_{\mathcal{Q}}\left\{\log \rho_{n|r_{\pi(\boldsymbol{x}_n)}^t}\right\} - \frac{1}{2}\sum_{n=1}^{N}\big(\hat{m}_{n|r_{\pi(\boldsymbol{x}_n)}^t}^{\mathrm{T}}\phi(x_{\pi(\boldsymbol{x}_n)}^t) - x_n^t\big)^2 \mathsf{E}_{\mathcal{Q}}\left\{\rho_{n|r_{\pi(\boldsymbol{x}_n)}^t}\right\} \\
& - \frac{1}{2}\sum_{n=1}^{N}\phi(x_{\pi(\boldsymbol{x}_n)}^t)^{\mathrm{T}}\hat{\Lambda}_{n|r_{\pi(\boldsymbol{x}_n)}^t}^{-1}\phi(x_{\pi(\boldsymbol{x}_n)}^t) + \sum_{k=1}^{K}\mathsf{E}_{\mathcal{Q}}\left\{\log \theta_{k|r_{\pi(\boldsymbol{r}_k)}^t}(r_k^t)\right\} \\
\hat{\gamma}_{k|r_{\pi(\boldsymbol{r}_k)}}(r_k) =\ & \gamma_{k|r_{\pi(\boldsymbol{r}_k)}} + \sum_{t=1}^{T}\mathsf{E}_{\mathcal{Q}}\left\{\mathbb{1}_{\left\{r_{\pi(\boldsymbol{r}_k)}^t = r_{\pi(\boldsymbol{r}_k)}\right\}}\mathbb{1}_{\left\{r_k^t = r_k\right\}}\right\} \\
\hat{\Lambda}_{n|r_{\pi(\boldsymbol{x}_n)}} =\ & \Lambda_{n|r_{\pi(\boldsymbol{x}_n)}} + \sum_{t=1}^{T}\mathsf{E}_{\mathcal{Q}}\left\{\mathbb{1}_{\left\{r_{\pi(\boldsymbol{x}_n)}^t = r_{\pi(\boldsymbol{x}_n)}\right\}}\right\}\phi(x_{\pi(\boldsymbol{x}_n)}^t)\phi(x_{\pi(\boldsymbol{x}_n)}^t)^{\mathrm{T}} \\
\hat{m}_{n|r_{\pi(\boldsymbol{x}_n)}} =\ & \hat{\Lambda}_{n|r_{\pi(\boldsymbol{x}_n)}}^{-1}\big(\Lambda_{n|r_{\pi(\boldsymbol{x}_n)}}m_{n|r_{\pi(\boldsymbol{x}_n)}} + \sum_{t=1}^{T}\mathsf{E}_{\mathcal{Q}}\left\{\mathbb{1}_{\left\{r_{\pi(\boldsymbol{x}_n)}^t = r_{\pi(\boldsymbol{x}_n)}\right\}}\right\}x_n^t\phi(x_{\pi(\boldsymbol{x}_n)}^t)\big) \\
\hat{a}_{n|r_{\pi(\boldsymbol{x}_n)}} =\ & a_{n|r_{\pi(\boldsymbol{x}_n)}} + \frac{1}{2}\sum_{t=1}^{T}\mathsf{E}_{\mathcal{Q}}\left\{\mathbb{1}_{\left\{r_{\pi(\boldsymbol{x}_n)}^t = r_{\pi(\boldsymbol{x}_n)}\right\}}\right\} \\
\hat{b}_{n|r_{\pi(\boldsymbol{x}_n)}} =\ & b_{n|r_{\pi(\boldsymbol{x}_n)}} + \frac{1}{2}\Big(m_{n|r_{\pi(\boldsymbol{x}_n)}}{}^{\mathrm{T}}\Lambda_{n|r_{\pi(\boldsymbol{x}_n)}}m_{n|r_{\pi(\boldsymbol{x}_n)}} \\
& - \hat{m}_{n|r_{\pi(\boldsymbol{x}_n)}}^{\mathrm{T}}\hat{\Lambda}_{n|r_{\pi(\boldsymbol{x}_n)}}\hat{m}_{n|r_{\pi(\boldsymbol{x}_n)}} + \sum_{t=1}^{T}\mathsf{E}_{\mathcal{Q}}\left\{\mathbb{1}_{\left\{r_{\pi(\boldsymbol{x}_n)}^t = r_{\pi(\boldsymbol{x}_n)}\right\}}\right\}(x_n^t)^2\Big)
\end{aligned}$$

Require: $\mathcal{G}$,$x_{1:N}^{1:T}$

Initialize $\hat{\gamma}_{1:K}$, $\hat{m}_{1:N}$, $\hat{\Lambda}_{1:N}$, $\hat{a}_{1:N}$, $\hat{b}_{1:N}$

repeat

Update expected sufficient statistics

- $\mathsf{E}_{\mathcal{Q}}\left\{\log \boldsymbol{\rho}_{n|r_{\pi(\boldsymbol{x}_n)}}\right\} \leftarrow \psi(\hat{a}_{n|r_{\pi(\boldsymbol{x}_n)}}) - \log \hat{b}_{n|r_{\pi(\boldsymbol{x}_n)}}$
- $\mathsf{E}_{\mathcal{Q}}\left\{\boldsymbol{\rho}_{n|r_{\pi(\boldsymbol{x}_n)}}\right\} \leftarrow \frac{\hat{a}_{n|r_{\pi(\boldsymbol{x}_n)}}}{\hat{b}_{n|r_{\pi(\boldsymbol{x}_n)}}}$
- $\mathsf{E}_{\mathcal{Q}}\left\{\log \theta_{k|r_{\pi(\boldsymbol{r}_k)}}(r_k)\right\} \leftarrow \psi\big(\hat{\gamma}_{k|r_{\pi(\boldsymbol{r}_k)}}(r_k)\big) - \psi\big(\sum_{r'_k} \hat{\gamma}_{k|r_{\pi(\boldsymbol{r}_k)}}(r'_k)\big)$

for $t = 1, \ldots, T$ **do**

Update $\log \hat{\theta}^t$

Update expected sufficient statistics

- $\mathsf{E}_{\mathcal{Q}}\left\{\mathbb{1}_{\left\{r^t_{\pi(\boldsymbol{r}_k)}=r_{\pi(\boldsymbol{r}_k)}\right\}}\mathbb{1}_{\left\{r^t_k=r_k\right\}}\right\} \leftarrow \sum_{r_{\neg U}} \hat{\theta}^t(r_{1:K})$ where $U = \pi(\boldsymbol{r}_k) \cup \{k\}$
- $\mathsf{E}_{\mathcal{Q}}\left\{\mathbb{1}_{\left\{r^t_{\pi(\boldsymbol{x}_n)}=r_{\pi(\boldsymbol{x}_n)}\right\}}\right\} \leftarrow \sum_{r_{\neg U}} \hat{\theta}^t(r_{1:K})$ where $U = \pi(\boldsymbol{x}_n)$

end for

Update $\hat{\gamma}_{1:K}$, $\hat{m}_{1:N}$, $\hat{\Lambda}_{1:N}$, $\hat{a}_{1:N}$, $\hat{b}_{1:N}$ w.r.t. the expected sufficient statistics.

Update $\mathcal{B}_{\mathcal{P}}[\mathcal{Q}]$ via Equation 4.32

until $\mathcal{B}_{\mathcal{P}}[\mathcal{Q}]$ converges

return Variational parameters $\hat{\theta}^{1:T}$, $\hat{\gamma}_{1:K}$, $\hat{m}_{1:N}$, $\hat{\Lambda}_{1:N}$, $\hat{a}_{1:N}$, $\hat{b}_{1:N}$

return The evidence lower bound $\mathcal{B}_{\mathcal{P}}[\mathcal{Q}]$.

Figure 4.5. VB-CN: Variational inference for latent variable causal networks.

Notice that evaluating Equation 4.24 and Equation 4.25, takes the form of updating the variational parameters. For updating parameters, first we need to calculate the following *expected sufficient statistics*:

$$\mathsf{E}_{\mathcal{Q}}\left\{\log \boldsymbol{\rho}_{n|r_{\pi(\boldsymbol{x}_n)}}\right\} = \psi(\hat{a}_{n|r_{\pi(\boldsymbol{x}_n)}}) - \log \hat{b}_{n|r_{\pi(\boldsymbol{x}_n)}} \tag{4.28}$$

$$\mathsf{E}_{\mathcal{Q}}\left\{\boldsymbol{\rho}_{n|r_{\pi(\boldsymbol{x}_n)}}\right\} = \frac{\hat{a}_{n|r_{\pi(\boldsymbol{x}_n)}}}{\hat{b}_{n|r_{\pi(\boldsymbol{x}_n)}}} \tag{4.29}$$

$$\mathsf{E}_{\mathcal{Q}}\left\{\log \theta_{k|r_{\pi(\boldsymbol{r}_k)}}(r_k)\right\} = \psi\big(\hat{\gamma}_{k|r_{\pi(\boldsymbol{r}_k)}}(r_k)\big) - \psi\Big(\sum_{r'_k} \hat{\gamma}_{k|r_{\pi(\boldsymbol{r}_k)}}(r'_k)\Big) \tag{4.30}$$

$$\mathsf{E}_{\mathcal{Q}}\left\{\mathbb{1}_{\left\{\boldsymbol{r}^t_U=r_U\right\}}\right\} = \sum_{r_{\neg U}} \hat{\theta}^t(r_{1:K}) \tag{4.31}$$

where the expectations are to be taken with respect to $\mathcal{Q}$ with the most recently updated parameters. In its final form, our variational algorithm becomes equivalent to iteratively calculating the expected sufficient statistics and updating the parameters. This is the reason why the factorization in Equation 4.23 is referred to as "mean-field" approximation. A simplified sketch of our variational inference algorithm *VB-CN* is also presented in Figure 4.5.

As a side note, calculating the sufficient statistics of form $\mathsf{E}_{\mathcal{Q}}\left\{\mathbb{1}_{\left\{\boldsymbol{r}_U^t=r_U\right\}}\right\}$ naively would require storing an intractable tensor of size $\prod_{k=1}^{K}|\mathcal{R}_k|$. Yet, as $\hat{\theta}^t$ respects a factorization implied by the DAG $\mathcal{G}$, and depending on the structure of the graph $\mathcal{G}$ and the set of indices U, it is possible to calculate the required sufficient statistics exactly by the junction tree algorithm. This is very attractive because we do not need to explicitly store or construct the tensors $\mathsf{E}_{\mathcal{Q}}\left\{\mathbb{1}_{\left\{\boldsymbol{r}_U^t=r_U\right\}}\right\}$ but only typically much lower dimensional clique potentials.

4.2.1. Calculation of Evidence Lower Bound

In Bayesian analysis, calculation of ELBO has a great practical importance in two respects: First, since it is the objective function of the variational inference, it serves as the convergence indicator of a variational algorithm. Secondly, and more importantly, it is used to assess the model fit, i.e. model selection and hyperparameter optimization is done through monitoring ELBO [63].

Previously, we have shown that the optimal form of the variational distribution obeys a certain factorization. Here our aim is to derive a closed form expression for the ELBO by making use of this factorization. We start our derivation by first expressing ELBO as a sum of simple expectation terms and group them as follows

$$\mathcal{B}_{\mathcal{P}}[\mathcal{Q}] \equiv \mathsf{E}_{\mathcal{Q}}\left\{\log p(\boldsymbol{r}_{1:K}^{1:T}, x_{1:N}^{1:T}, \boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N}) - \log \mathcal{Q}(\boldsymbol{r}_{1:K}^{1:T}, \boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N})\right\} \quad (4.32)$$

$$= \sum_{t=1}^{T}\sum_{n=1}^{N} \mathsf{E}_{\mathcal{Q}}\left\{\log p(x_n^t \mid x_{\pi(\boldsymbol{x}_n)}^t, \boldsymbol{r}_{\pi(\boldsymbol{x}_n)}^t, \boldsymbol{w}_n, \boldsymbol{\rho}_n)\right\} \quad (4.33)$$

$$+ \sum_{t=1}^{T}\Big(\sum_{k=1}^{K} \mathsf{E}_{\mathcal{Q}}\left\{\log p(\boldsymbol{r}_k^t \mid \boldsymbol{r}_{\pi(\boldsymbol{r}_k)}^t, \boldsymbol{\theta}_k)\right\} - \mathsf{E}_{\mathcal{Q}}\left\{\log q(\boldsymbol{r}_{1:K}^t)\right\}\Big) \quad (4.34)$$

$$+ \sum_{k=1}^{K}\sum_{r_{\pi(\boldsymbol{r}_k)}} \Big(\mathsf{E}_{\mathcal{Q}}\left\{\log p(\theta_{k|r_{\pi(\boldsymbol{r}_k)}})\right\} - \mathsf{E}_{\mathcal{Q}}\left\{\log q(\theta_{k|r_{\pi(\boldsymbol{r}_k)}})\right\}\Big) \quad (4.35)$$

$$+ \sum_{n=1}^{N}\sum_{r_{\pi(\boldsymbol{x}_n)}} \Big(\mathsf{E}_{\mathcal{Q}}\left\{\log p(\boldsymbol{w}_{n|r_{\pi(\boldsymbol{x}_n)}}, \boldsymbol{\rho}_{n|r_{\pi(\boldsymbol{x}_n)}})\right\} - \mathsf{E}_{\mathcal{Q}}\left\{\log q(\boldsymbol{w}_{n|r_{\pi(\boldsymbol{x}_n)}}, \boldsymbol{\rho}_{n|r_{\pi(\boldsymbol{x}_n)}})\right\}\Big) \quad (4.36)$$

where we employ the factorization property of the joint distribution, and the factorization property of the variational posterior. In that form, evaluation of each individual expectation term is straightforward, so we conclude our description by stating the results of each expectation. The necessary derivation steps are given in Appendix C.

(i) Expectation of the Gaussian log-likelihood terms in Equation 4.33 are found to be

$$\begin{aligned}\mathsf{E}_{\mathcal{Q}}\left\{\log p(x_n^t \mid x_{\pi(\boldsymbol{x}_n)}^t, \boldsymbol{r}_{\pi(\boldsymbol{x}_n)}^t, \boldsymbol{w}_n, \boldsymbol{\rho}_n)\right\} = \\ \frac{1}{2}\sum_{r_{\pi(\boldsymbol{x}_n)}} \mathsf{E}_{\mathcal{Q}}\left\{\mathbb{1}_{\left\{\boldsymbol{r}_{\pi(\boldsymbol{x}_n)}^t = r_{\pi(\boldsymbol{x}_n)}\right\}}\right\}\Big(-\big(x_n^t - \hat{m}_{n|r_{\pi(\boldsymbol{x}_n)}}^{\mathrm{T}}\phi(x_{\pi(\boldsymbol{x}_n)}^t)\big)^2 \mathsf{E}_{\mathcal{Q}}\left\{\boldsymbol{\rho}_{n|r_{\pi(\boldsymbol{x}_n)}}\right\} \\ + \mathsf{E}_{\mathcal{Q}}\left\{\log \boldsymbol{\rho}_{n|r_{\pi(\boldsymbol{x}_n)}}\right\} - \phi(x_{\pi(\boldsymbol{x}_n)}^t)^{\mathrm{T}}\hat{\Lambda}_{n|r_{\pi(\boldsymbol{x}_n)}}^{-1}\phi(x_{\pi(\boldsymbol{x}_n)}^t) - \log 2\pi\Big)\end{aligned}$$

(ii) Expectation of the categorical log-likelihood and variational entropy terms in Equation 4.34 are given as

$$\mathsf{E}_{\mathcal{Q}}\left\{\log p(\boldsymbol{r}_k^t \mid \boldsymbol{r}_{\pi(\boldsymbol{r}_k)}^t, \boldsymbol{\theta}_k)\right\} = \sum_{r_k, r_{\pi(\boldsymbol{r}_k)}} \mathsf{E}_{\mathcal{Q}}\left\{\mathbb{1}_{\left\{\boldsymbol{r}_k^t = r_k\right\}}\mathbb{1}_{\left\{\boldsymbol{r}_{\pi(\boldsymbol{r}_k)}^t = r_{\pi(\boldsymbol{r}_k)}\right\}}\right\} \mathsf{E}_{\mathcal{Q}}\left\{\log \theta_{k|r_{\pi(\boldsymbol{r}_k)}}(r_k)\right\}$$

$$\mathsf{E}_{\mathcal{Q}}\left\{\log q(\boldsymbol{r}_{1:K}^t)\right\} = \sum_{r_{1:K}} \hat{\theta}^t(r_{1:K}) \log \hat{\theta}^t(r_{1:K})$$

(iii) Negative cross entropy and negative entropy terms of Dirichlet distributions in Equation 4.35 are given as

$$\begin{aligned}
\mathsf{E}_{\mathcal{Q}}\Big\{\log p(\boldsymbol{\theta}_{k|r_{\pi(\boldsymbol{r}_k)}})\Big\} &= \log\Gamma\big(\sum_{r_k}\gamma_{k|r_{\pi(\boldsymbol{r}_k)}}(r_k)\big) - \sum_{r_k}\log\Gamma\big(\gamma_{k|r_{\pi(\boldsymbol{r}_k)}}(r_k)\big) \\
&\quad + \sum_{r_k}\big(\gamma_{k|r_{\pi(\boldsymbol{r}_k)}}(r_k)-1\big)\mathsf{E}_{\mathcal{Q}}\Big\{\log\boldsymbol{\theta}_{k|r_{\pi(\boldsymbol{r}_k)}}(r_k)\Big\} \\
\mathsf{E}_{\mathcal{Q}}\Big\{\log q(\boldsymbol{\theta}_{k|r_{\pi(\boldsymbol{r}_k)}})\Big\} &= \log\Gamma\big(\sum_{r_k}\hat{\gamma}_{k|r_{\pi(\boldsymbol{r}_k)}}(r_k)\big) - \sum_{r_k}\log\Gamma\big(\hat{\gamma}_{k|r_{\pi(\boldsymbol{r}_k)}}(r_k)\big) \\
&\quad + \sum_{r_k}\big(\hat{\gamma}_{k|r_{\pi(\boldsymbol{r}_k)}}(r_k)-1\big)\mathsf{E}_{\mathcal{Q}}\Big\{\log\boldsymbol{\theta}_{k|r_{\pi(\boldsymbol{r}_k)}}(r_k)\Big\}
\end{aligned}$$

(iv) Finally, negative cross entropy and negative entropy terms of Normal-Gamma distributions in Equation 4.36 are given as

$$\begin{aligned}
&\mathsf{E}_{\mathcal{Q}}\Big\{\log p(\boldsymbol{w}_{n|r_{\pi(\boldsymbol{x}_n)}},\boldsymbol{\rho}_{n|r_{\pi(\boldsymbol{x}_n)}})\Big\} = \\
&\quad a_{n|r_{\pi(\boldsymbol{x}_n)}}\log b_{n|r_{\pi(\boldsymbol{x}_n)}} - \log\Gamma(a_{n|r_{\pi(\boldsymbol{x}_n)}}) + \frac{1}{2}\big(\log\det(\Lambda_{n|r_{\pi(\boldsymbol{x}_n)}}) - \mathrm{tr}(\hat{\Lambda}^{-1}_{n|r_{\pi(\boldsymbol{x}_n)}}\Lambda_{n|r_{\pi(\boldsymbol{x}_n)}})\big) \\
&\quad - \frac{M}{2}\log 2\pi + \Big(a_{n|r_{\pi(\boldsymbol{x}_n)}} + \frac{M}{2} - 1\Big)\mathsf{E}_{\mathcal{Q}}\Big\{\log\boldsymbol{\rho}_{n|r_{\pi(\boldsymbol{x}_n)}}\Big\} - b_{n|r_{\pi(\boldsymbol{x}_n)}}\mathsf{E}_{\mathcal{Q}}\Big\{\boldsymbol{\rho}_{n|r_{\pi(\boldsymbol{x}_n)}}\Big\} \\
&\quad - \frac{1}{2}\mathsf{E}_{\mathcal{Q}}\Big\{\boldsymbol{\rho}_{n|r_{\pi(\boldsymbol{x}_n)}}\Big\}(\hat{m}_{n|r_{\pi(\boldsymbol{x}_n)}} - m_{n|r_{\pi(\boldsymbol{x}_n)}})^{\mathrm{T}}\Lambda_{n|r_{\pi(\boldsymbol{x}_n)}}(\hat{m}_{n|r_{\pi(\boldsymbol{x}_n)}} - m_{n|r_{\pi(\boldsymbol{x}_n)}})
\end{aligned}$$

$$\begin{aligned}
&\mathsf{E}_{\mathcal{Q}}\Big\{\log q(\boldsymbol{w}_{n|r_{\pi(\boldsymbol{x}_n)}},\boldsymbol{\rho}_{n|r_{\pi(\boldsymbol{x}_n)}})\Big\} = \\
&\quad \hat{a}_{n|r_{\pi(\boldsymbol{x}_n)}}\log\hat{b}_{n|r_{\pi(\boldsymbol{x}_n)}} - \log\Gamma(\hat{a}_{n|r_{\pi(\boldsymbol{x}_n)}}) + \frac{1}{2}\log\det(\hat{\Lambda}_{n|r_{\pi(\boldsymbol{x}_n)}}) - \frac{M}{2} \\
&\quad - \frac{M}{2}\log 2\pi + \Big(\hat{a}_{n|r_{\pi(\boldsymbol{x}_n)}} + \frac{M}{2} - 1\Big)\mathsf{E}_{\mathcal{Q}}\Big\{\log\boldsymbol{\rho}_{n|r_{\pi(\boldsymbol{x}_n)}}\Big\} - \hat{b}_{n|r_{\pi(\boldsymbol{x}_n)}}\mathsf{E}_{\mathcal{Q}}\Big\{\boldsymbol{\rho}_{n|r_{\pi(\boldsymbol{x}_n)}}\Big\}
\end{aligned}$$

Notice that each individual term is expressed in terms of the expected sufficient statistics. Once the expected sufficient statistics are estimated, we are ready to evaluate a closed form expression for the evidence lower bound.

4.2.2. A Dual Expectation-Maximization Algorithm

Expectation–maximization (EM) algorithm is an iterative optimization technique which is typically used to find the maximum a posteriori (MAP) estimations of the model parameters in latent variable models [24]. Rather than maximizing the posterior distribution directly, EM algorithm is based on maximizing the lower bound of posterior distribution similar to variational methods. This lower bound is devised by decomposing the posterior distribution as follows

$$\begin{aligned}\log p(\theta_{1:K}, \rho_{1:N}, w_{1:N} \mid x_{1:N}^{1:T}) &= \mathsf{E}_{\mathcal{Q}(\boldsymbol{r}_{1:K}^{1:T})}\left\{\log \frac{p(\boldsymbol{r}_{1:K}^{1:T}, \theta_{1:K}, \rho_{1:N}, w_{1:N} \mid x_{1:N}^{1:T})}{\mathcal{Q}(\boldsymbol{r}_{1:K}^{1:T})}\right\} \\ &\quad + \mathsf{E}_{\mathcal{Q}(\boldsymbol{r}_{1:K}^{1:T})}\left\{\log \frac{\mathcal{Q}(\boldsymbol{r}_{1:K}^{1:T})}{p(\boldsymbol{r}_{1:K}^{1:T} \mid \theta_{1:K}, \rho_{1:N}, w_{1:N}, x_{1:N}^{1:T})}\right\} \\ &\equiv \mathcal{B}[\mathcal{Q}, \theta_{1:K}, \rho_{1:N}, w_{1:N}] + \mathrm{KL}\big(\mathcal{Q} \| p(\boldsymbol{r}_{1:K}^{1:T} \mid \theta_{1:K}, \rho_{1:N}, w_{1:N}, x_{1:N}^{1:T})\big)\end{aligned}$$

where $\mathcal{B}[\mathcal{Q}, \theta_{1:K}, \rho_{1:N}, w_{1:N}]$ is the aforementioned lower bound and $\mathcal{Q}$ is an arbitrary distribution on latent variables $\mathbf{r}_{1:K}^{1:T}$. To find the mode of the lower bound, a coordinate ascent algorithm is utilized, in which $\mathcal{B}$ is alternatingly optimized with respect to distribution $\mathcal{Q}$ and parameters $\theta_{1:K}$, $\rho_{1:N}$, $w_{1:N}$, i.e. in each iteration of the algorithm

$$\mathcal{Q}^{(j)} \leftarrow \arg\max_{\mathcal{Q}} \mathcal{B}[\mathcal{Q}, \theta_{1:K}^{(j-1)}, \rho_{1:N}^{(j-1)}, w_{1:N}^{(j-1)}] \tag{4.37}$$

$$\theta_{1:K}^{(j)}, \rho_{1:N}^{(j)}, w_{1:N}^{(j)} \leftarrow \arg\max_{\theta_{1:K}, \rho_{1:N}, w_{1:N}} \mathcal{B}[\mathcal{Q}^{(j)}, \theta_{1:K}, \rho_{1:N}, w_{1:N}] \tag{4.38}$$

Since posterior of the parameters does not depend on $\mathcal{Q}$, maximizing the lower bound w.r.t. $\mathcal{Q}$ is equivalent to minimizing the $\mathrm{KL}\big(\mathcal{Q} \| p(\boldsymbol{r}_{1:K}^{1:T} \mid \theta_{1:K}, \rho_{1:N}, w_{1:N}, x_{1:N}^{1:T})\big)$, and the minimum value of KL divergence is attained only if its arguments are equal. So, the first step of the algorithm reduces to setting $\mathcal{Q}$ to the posterior $p(\boldsymbol{r}_{1:K}^{1:T} \mid \theta_{1:K}, \rho_{1:N}, w_{1:N}, x_{1:N}^{1:T})$. Equivalently, only the expected sufficient statistics of $\mathbf{r}_{1:K}^{1:T}$ can be calculated, because these expectations are the only relevant information of $\mathcal{Q}$ that are required for solving the maximization problem in the second step. Based on this, these consecutive steps are called the *expectation* and *maximization* steps, respectively.

Require: $\mathcal{G}$,$x_{1:N}^{1:T}$

Initialize $\gamma_{1:K}^*$, $m_{1:N}^*$, $\Lambda_{1:N}^*$, $a_{1:N}^*$, $b_{1:N}^*$

repeat

Update expected sufficient statistics

- $\mathsf{E}_{\mathcal{Q}^*}\left\{\log \boldsymbol{\rho}_{n|r_{\pi(\boldsymbol{x}_n)}}\right\} \leftarrow \psi(a^*_{n|r_{\pi(\boldsymbol{x}_n)}}) - \log b^*_{n|r_{\pi(\boldsymbol{x}_n)}}$
- $\mathsf{E}_{\mathcal{Q}^*}\left\{\boldsymbol{\rho}_{n|r_{\pi(\boldsymbol{x}_n)}}\right\} \leftarrow \frac{a^*_{n|r_{\pi(\boldsymbol{x}_n)}}}{b^*_{n|r_{\pi(\boldsymbol{x}_n)}}}$
- $\mathsf{E}_{\mathcal{Q}^*}\left\{\log \theta_{k|r_{\pi(\boldsymbol{r}_k)}}(r_k)\right\} \leftarrow \psi\big(\gamma^*_{k|r_{\pi(\boldsymbol{r}_k)}}(r_k)\big) - \psi\big(\sum_{r'_k} \gamma^*_{k|r_{\pi(\boldsymbol{r}_k)}}(r'_k)\big)$

for $t = 1, \ldots, T$ **do**

$\log \hat{\theta}^t(r^t_{1:K}) \leftarrow \mathsf{E}_{\mathcal{Q}^*}\left\{\log p(r^t_{1:K}, x^t_{1:N} \mid \boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N})\right\}$

$r^{t,*}_{1:K} \leftarrow \arg\max_{r^t_{1:K}} \hat{\theta}^t(r^t_{1:K})$

end for

Update $\gamma^*_{1:K}$, $m^*_{1:N}$, $\Lambda^*_{1:N}$, $a^*_{1:N}$, $b^*_{1:N}$ w.r.t. $r^{1:T,*}_{1:K}$.

until $r^{1:T,*}_{1:K}$ converges to a fixed value.

return $r^{1:T,*}_{1:K}$

Figure 4.6. DEM-CN: Dual Expectation-Maximization algorithm for latent variable causal networks.

In the context of causal discovery, we are interested in inferring the latent confounders rather than estimating the parameters of the model. Our primary goal is to find out most likely configuration of the latent variables where the estimates of the parameters have little bearing on. In other words, we are interested in finding the maximum a posteriori $\boldsymbol{r}^{1:T}_{1:K}$, which can be expressed as the following combinatorial optimization problem:

$$r^{1:T,*}_{1:K} = \arg\max_{r^{1:T}_{1:K}} p(r^{1:T}_{1:K} \mid x^{1:T}_{1:N}) \tag{4.39}$$

$$= \arg\max_{r^{1:T}_{1:K}} p(r^{1:T}_{1:K}, x^{1:T}_{1:N}) \tag{4.40}$$

Although we have a closed form expression for $p(r^{1:T}_{1:K}, x^{1:T}_{1:N})$, it is evident from its form that the individual coordinates of $r^{1:T}_{1:K}$ are tightly coupled. Due to this coupling,

it is difficult to utilize an efficient optimization algorithm, considering the fact that the naive maximization algorithm requires the evaluation of $p(r_{1:K}^{1:T}, x_{1:N}^{1:T})$ for exponentially many configurations of $r_{1:K}^{1:T}$. To enable efficient optimization, we introduce a dual formulation for EM where we break this coupling by defining a lower bound on the posterior distribution of the latent variables:

$$\begin{aligned}
\log p(r_{1:K}^{1:T} \mid x_{1:N}^{1:T}) &= \mathsf{E}_{\mathcal{Q}(\boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N})} \left\{ \log \frac{p(r_{1:K}^{1:T}, \boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N} \mid x_{1:N}^{1:T})}{\mathcal{Q}(\boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N})} \right\} \\
&\quad + \mathsf{E}_{\mathcal{Q}(\boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N})} \left\{ \log \frac{\mathcal{Q}(\boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N})}{p(\boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N} \mid r_{1:K}^{1:T}, x_{1:N}^{1:T})} \right\} \\
&\equiv \mathcal{B}[\mathcal{Q}, r_{1:K}^{1:T}] + \mathrm{KL}\big(\mathcal{Q} \| p(\boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N} \mid r_{1:K}^{1:T}, x_{1:N}^{1:T})\big)
\end{aligned}$$

Here, our intention is maximizing the lower bound $\mathcal{B}[\mathcal{Q}, r_{1:K}^{1:T}]$ on the space that is augmented with the distribution $\mathcal{Q}$, instead of maximizing the posterior density directly on $r_{1:K}^{1:T}$. To do that, we again employ a coordinate ascent algorithm:

$$\mathcal{Q}^* \leftarrow \arg\max_{\mathcal{Q}} \mathcal{B}[\mathcal{Q}, r_{1:K}^{1:T,*}] \tag{4.41}$$

$$r_{1:K}^{1:T,*} \leftarrow \arg\max_{r_{1:K}^{1:T}} \mathcal{B}[\mathcal{Q}^*, r_{1:K}^{1:T}] \tag{4.42}$$

We can simplify this algorithm, based on the similar arguments used in standard EM. The first step of the algorithm sets $\mathcal{Q}$ to the posterior $p(\boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N} \mid r_{1:K}^{1:T}, x_{1:N}^{1:T})$, which requires the calculation of the posterior parameters presented in Chapter 3. By using the factorization property of the joint distribution and discarding the constant terms, we construct the following equivalent objective function for the second step:

$$\mathcal{B}[\mathcal{Q}^*, r_{1:K}^{1:T}] \equiv \mathsf{E}_{\mathcal{Q}^*} \left\{ \log \frac{p(r_{1:K}^{1:T}, \boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N} \mid x_{1:N}^{1:T})}{\mathcal{Q}^*(\boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N})} \right\} \tag{4.43}$$

$$=^+ \mathsf{E}_{\mathcal{Q}^*} \left\{ \log p(r_{1:K}^{1:T}, \boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N} \mid x_{1:N}^{1:T}) \right\} \tag{4.44}$$

$$=^+ \mathsf{E}_{\mathcal{Q}^*} \left\{ \log p(r_{1:K}^{1:T}, x_{1:N}^{1:T} \mid \boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N}) \right\} \tag{4.45}$$

$$= \sum_{t=1}^{T} \mathsf{E}_{\mathcal{Q}^*} \left\{ \log p(r_{1:K}^{t}, x_{1:N}^{t} \mid \boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N}) \right\} \tag{4.46}$$

Therefore we obtain a much simpler fixed point iteration algorithm in the form of:

$$\mathcal{Q}^* \leftarrow p(\boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N} \mid r_{1:K,}^{1:T,*}, x_{1:N}^{1:T}) \tag{4.47}$$

$$\forall t: \qquad r_{1:K}^{t,*} \leftarrow \underset{r_{1:K}^t}{\arg\max}\ \mathsf{E}_{\mathcal{Q}^*}\left\{\log p(r_{1:K}^t, x_{1:N}^t \mid \boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N})\right\} \tag{4.48}$$

where the optimization with respect to $r_{1:K}^{1:T}$ is performed through each variable $r_{1:K}^1$, ..., $r_{1:K}^T$ independently. We finalize our recipe by providing a closed form expression for the expectation in Equation 4.48

$$\begin{aligned}
&\mathsf{E}_{\mathcal{Q}^*}\left\{\log p(r_{1:K}^t, x_{1:N}^t \mid \boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N})\right\} \\
&=^+ \frac{1}{2}\sum_{n=1}^{N} \mathsf{E}_{\mathcal{Q}^*}\left\{\log \rho_{n|r_{\pi(\boldsymbol{x}_n)}^t}\right\} - \frac{1}{2}\sum_{n=1}^{N}\left(m_{n|r_{\pi(\boldsymbol{x}_n)}^t}^{*\,\mathrm{T}}\phi(x_{\pi(\boldsymbol{x}_n)}^t) - x_n^t\right)^2 \mathsf{E}_{\mathcal{Q}^*}\left\{\rho_{n|r_{\pi(\boldsymbol{x}_n)}^t}\right\} \\
&- \frac{1}{2}\sum_{n=1}^{N}\phi(x_{\pi(\boldsymbol{x}_n)}^t)^{\mathrm{T}}\Lambda_{n|r_{\pi(\boldsymbol{x}_n)}^t}^{*\,-1}\phi(x_{\pi(\boldsymbol{x}_n)}^t) + \sum_{k=1}^{K}\mathsf{E}_{\mathcal{Q}^*}\left\{\log \theta_{k|r_{\pi(\boldsymbol{r}_k)}^t}(r_k^t)\right\}
\end{aligned}$$

Note that each iteration of the algorithm monotonically increases the objective function $\mathcal{B}[\mathcal{Q}, r_{1:K}^{1:T}]$, since we make use of a coordinate ascent algorithm. More importantly, the lower bound $\mathcal{B}[\mathcal{Q}, r_{1:K}^{1:T}]$ is tight when $\mathcal{Q}$ is equal to posterior distribution $p(\boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N} \mid r_{1:K}^{1:T}, x_{1:N}^{1:T})$, and therefore maximization of it also ensures the maximization of $\log p(r_{1:K}^{1:T} \mid x_{1:N}^{1:T})$. Finally, we provide an outline of our dual EM algorithm in Figure 4.6 without going into minor details.

5. EXPERIMENTS AND RESULTS

We tested all of our model selection and inference methods with synthetic and real-world data sets in the bivariate scenario. It is a well-known result in Bayesian analysis that *Bayes factors* are consistent, i.e. always indicate the true models as the sample sizes go to infinity [64], but they are highly sensitive to prior parameters [65], especially when the sample sizes are too small. Due to finite sample sizes, one can manipulate an experiment by assuming strong priors and might obtain the desired result without even needing the data. Therefore we avoid specific assumptions regarding the prior parameters throughout the experiments. As we mentioned earlier, our framework relies on inferring the hyperparameters of the prior distributions from labeled data to conduct model selection for causal structure learning.

We divide our experiments section into four parts, serving separate yet complementary purposes. In the first part, our aim is to illustrate the approximation qualities of our SMC and VB algorithms in a toy example where the exact marginal likelihood is computable. This part serves a fundamental mission for the rest of experiments, namely it confirms the reliability of the marginal likelihood estimates before endeavouring to perform model selection with respect to them. The second part includes the experiments conducted on synthetic data, and seeks to test whether the proposed SMC and VB algorithms can correctly identify the true generative model when the true hyperparameters are known. This part serves to demonstrate the reliability of the algorithms in model selection task under various hyperparameter settings.

After demonstrating the accuracy of the marginal likelihood estimates, and confirming that our algorithms can indeed determine the true generative model in each hyperparameter setting for synthetic data, we go on to test how our model applies to real-world causal inference scenarios in which the critical hyperparameters are unknown. This part of the experiments is conducted on *Cause Effect Pairs* (CEP) data set [31], which is frequently used in causal discovery research. CEP data set consists of 100 tuples, the vast majority of which is bivariate. Finally, we conclude our experi-

ments with the illustrations of the hidden representations found by our latent variable inference methods PG and DEM on *Abalone* data set [66].

As described above, since the selection of priors is very important in our construction, we first describe the hyperparameter settings used in the experiments, i.e. which hyperparameters were allowed to vary in which ranges. Then, we present the results of the experiments.

5.1. Hyperparameter Settings

In order to avoid incurring unnecessary computational costs, certain parameters were fixed for all our experiments when it was reasonable to do so. For all experiments the basis functions included 0^{th} and 1^{st} powers of the input variable[3] , i.e. we set $\phi(x) = [1 \; x]^{\text{T}}$. For the bivariate models in Figure 3.1, the cardinality $|\mathcal{R}_1|$ of the latent variable $\boldsymbol{r}_1$ was allowed to range between 1 and 5, in each case the cardinality that leads to the highest marginal likelihood was selected. As the parameters we fixed before inference: for both values of $n \in \{1, 2\}$ and for all values of $r_1 \in \mathcal{R}_1$, $m_{n|r_1}$'s were set to 0, and $\Lambda_{n|r_1}$'s were set to $\frac{1}{10}I$ each; while for all values of $r_1 \in \mathcal{R}_1$, $\gamma_1(r_1)$'s were set to 10.

We next describe the remaining hyperparameters with respect to the causal graph in Figure 3.1(a) in which $\boldsymbol{x}_1$ causes $\boldsymbol{x}_2$. Though their adaptation to other two graphs is straightforward due to symmetry. The hyperparameters of the Gamma distributions, $(a_1, \, b_1, \, a_2, \, b_2)$, from which the precision of the observed variables were drawn, were allowed to take different values with the condition that $a_{n|r_1} \geq b_{n|r_1}$ at all times, but again every element of these vectors corresponding to different values of r_1 assumed to be constant within the vector. This is because the mean of a Gamma distribution $\text{Gamma}(a, b)$ is a/b and its variance is a/b^2, therefore when b is allowed to take a greater value than a, this results in a close to zero precision value for the relevant distribution

[3]Although arbitrary basis functions may be employed in general, here we stick with the linear basis functions to avoid too complex assumptions that may potentially obscure the effect of our key assumptions. Since this constitutes the initial test of our framework, we hope to obtain a solid baseline for actual performance. See Chapter 6 for further discussion.

Table 5.1. All 36 hyperparameter settings that are used in the experiments.

HYPERPARAMETER SETTING #	γ_1	$\Lambda_{n\|r_1}$	$m_{n\|r_1}$	a_1	b_1	a_2	b_2
1	10.0	0.1	0.0	1.0	1.0	1.0	1.0
2	10.0	0.1	0.0	1.0	1.0	10.0	10.0
3	10.0	0.1	0.0	1.0	1.0	10.0	10.0
4	10.0	0.1	0.0	1.0	1.0	100.0	100.0
5	10.0	0.1	0.0	1.0	1.0	100.0	100.0
6	10.0	0.1	0.0	1.0	1.0	100.0	100.0
7	10.0	0.1	0.0	10.0	1.0	1.0	1.0
8	10.0	0.1	0.0	10.0	10.0	1.0	1.0
9	10.0	0.1	0.0	10.0	1.0	10.0	10.0
10	10.0	0.1	0.0	10.0	1.0	10.0	10.0
11	10.0	0.1	0.0	10.0	10.0	10.0	10.0
12	10.0	0.1	0.0	10.0	10.0	10.0	10.0
13	10.0	0.1	0.0	10.0	1.0	100.0	100.0
14	10.0	0.1	0.0	10.0	1.0	100.0	100.0
15	10.0	0.1	0.0	10.0	1.0	100.0	100.0
16	10.0	0.1	0.0	10.0	10.0	100.0	100.0
17	10.0	0.1	0.0	10.0	10.0	100.0	100.0
18	10.0	0.1	0.0	10.0	10.0	100.0	100.0
19	10.0	0.1	0.0	100.0	1.0	1.0	1.0
20	10.0	0.1	0.0	100.0	10.0	1.0	1.0
21	10.0	0.1	0.0	100.0	100.0	1.0	1.0
22	10.0	0.1	0.0	100.0	1.0	10.0	10.0
23	10.0	0.1	0.0	100.0	1.0	10.0	10.0
24	10.0	0.1	0.0	100.0	10.0	10.0	10.0
25	10.0	0.1	0.0	100.0	10.0	10.0	10.0
26	10.0	0.1	0.0	100.0	100.0	10.0	10.0
27	10.0	0.1	0.0	100.0	100.0	10.0	10.0
28	10.0	0.1	0.0	100.0	1.0	100.0	100.0
29	10.0	0.1	0.0	100.0	1.0	100.0	100.0
30	10.0	0.1	0.0	100.0	1.0	100.0	100.0
31	10.0	0.1	0.0	100.0	10.0	100.0	100.0
32	10.0	0.1	0.0	100.0	10.0	100.0	100.0
33	10.0	0.1	0.0	100.0	10.0	100.0	100.0
34	10.0	0.1	0.0	100.0	100.0	100.0	100.0
35	10.0	0.1	0.0	100.0	100.0	100.0	100.0
36	10.0	0.1	0.0	100.0	100.0	100.0	100.0

for the observed variable. Obeying the constraint, the a and b's were allowed to take values among 1, 10, and 100 each. The a parameter was not allowed to be larger than 100 since this leads to an equivalent sample size much larger than the sample size of certain data sets used in experiments, effectively rendering the observations unimportant. The b parameter was not allowed to be smaller than 1 since this again implies extremely imprecise Gaussian distributions for the observed variables to which the Gamma distribution provided the precision variable. The combinations with these constraints lead to a total of 36 sets of hyperparameters. All of these 36 parameters settings used in the experiments are presented in Table 5.1.

On the other hand, while doing model comparison in a hyperparameter setting, we expect several criteria to be satisfied for maintaining consistency. For instance, in the spurious model (Figure 3.1(c)) there is no reason to assign different priors on variables $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. Otherwise, just by permuting the labels of the pairs, we would obtain inconsistent marginal likelihoods. Likewise, when the labels of a pair are permuted, e.g. $\hat{x}_1^{1:T} \equiv x_2^{1:T}$ and $\hat{x}_2^{1:T} \equiv x_1^{1:T}$, we expect the marginal likelihood of the pair $(x_1^{1:T}, x_2^{1:T})$ given the relation $\boldsymbol{x}_1 \to \boldsymbol{x}_2$ to be equal to the marginal likelihood of the permuted pair $(\hat{x}_1^{1:T}, \hat{x}_2^{1:T})$ given the relation $\hat{\boldsymbol{x}}_2 \to \hat{\boldsymbol{x}}_1$. The rule we used to solve inconsistency issues in such situations is the following: the prior parameters of two variables must be identical whenever the parental graphs of them are *homomorphic*. So, if we are calculating the marginal likelihood of the relation $\boldsymbol{x}_1 \to \boldsymbol{x}_2$ with a particular hyperparameter setting, say ($a_1 = 100$, $b_1 = 10$, $a_2 = 10$, $b_2 = 1$), then the corresponding consistent hyperparameter setting for $\boldsymbol{x}_2 \to \boldsymbol{x}_1$ should be ($a_1 = 10$, $b_1 = 1$, $a_2 = 100$, $b_2 = 10$), whereas the corresponding consistent hyperparameters for the spurious relationship should be ($a_1 = 100$, $b_1 = 10$, $a_2 = 100$, $b_2 = 10$). Keeping that in mind, in the rest of this chapter we will only state the hyperparameter settings used for the relation $\boldsymbol{x}_1 \to \boldsymbol{x}_2$, since the necessary conversions for the other two models are straightforward.

5.2. Comparison of Algorithms on Toy Data

Remember that we derived a closed form formula for the marginal distribution of $r_1^{1:T}$, $x_{1:2}^{1:T}$ in Chapter 3. The exact computation of the marginal likelihood $p(x_{1:2}^{1:T})$

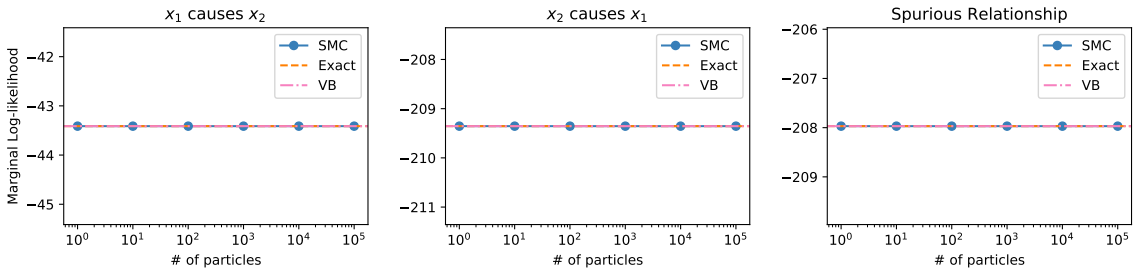
requires the summation of $p(r_1^{1:T}, x_{1:2}^{1:T})$ over all possible values of $r_1^{1:T} \in \mathcal{R}_1^T$, i.e.

$$p(x_{1:2}^{1:T}) = \sum_{r_1^{1:T} \in \mathcal{R}_1^T} p(r_1^{1:T}, x_{1:2}^{1:T}) \tag{5.1}$$

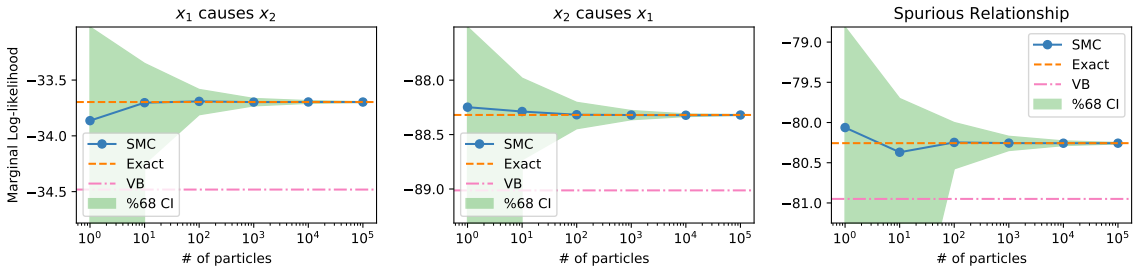
which also requires the evaluation of $p(r_1^{1:T}, x_{1:2}^{1:T})$ for exponentially many times. However, it is feasible to compute this sum in practice for very small values of $|\mathcal{R}_1|$ and T. In order to compare the accuracy of marginal likelihood estimations by VB and SMC, we generate $T = 10$ bivariate observations from the causal model in Figure 3.1(a) with $|\mathcal{R}_1| = 3$, for which it is feasible to calculate the exact marginal likelihood by exhaustive enumeration. We generated this toy data set with hyperparameter setting 20 where we set $a_1 = 100$, $b_1 = 10$, $a_2 = 10$ and $b_2 = 1$.

In Figure 5.1, we illustrate the results of the SMC and VB algorithms along with the exact marginal likelihood value. To be able to analyze the behaviour of the marginal likelihood estimates of SMC and VB algorithms statistically, we ran both of the algorithms 100 times, for each causal hypothesis and for different cardinalities $|\mathcal{R}_1| \in \{1, 2, 3\}$ of $\boldsymbol{r}_1$. Additionally, we varied the number of particles used in SMC over the range $[1, 10^6]$, in order to illustrate the asymptotic convergence visually. In the case of SMC, we report the average of the estimations along with their confidence interval, whereas in VB we report only the maximum of the evidence lower bounds as the final estimation of marginal likelihood.

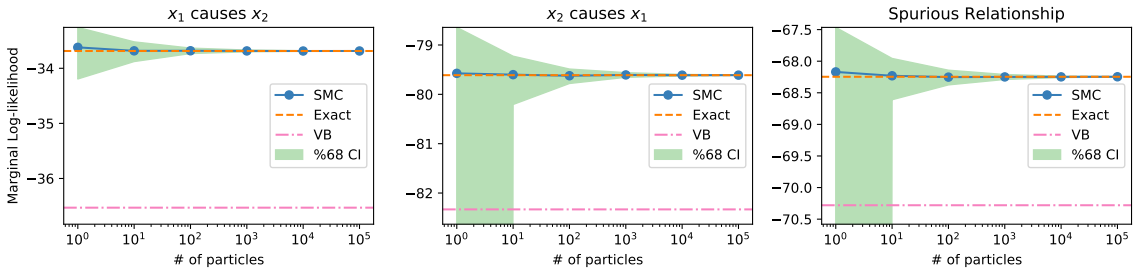
We start our comments with interpreting the results for the trivial case $|\mathcal{R}_1| = 1$. This case is trivial since the assumption of $|\mathcal{R}_1| = 1$ is equivalent to assuming the nonexistence of any latent random variable. Therefore the estimates of SMC and VB should be algebraically equal to the true marginal likelihood, which is confirmed by the Figure 5.1(a). In the remaining cases, we see the approximation qualities of the methods more clearly: For both of the cases and for each causal hypothesis, SMC estimates are unbiased, and become more and more accurate with the increasing number of particles (which can be seen from the shrinking confidence intervals), whereas there is no such asymptotic convergence guarantee for ELBO. Nevertheless, the gap between



(a) $|\mathcal{R}_1| = 1$.



(b) $|\mathcal{R}_1| = 2$.

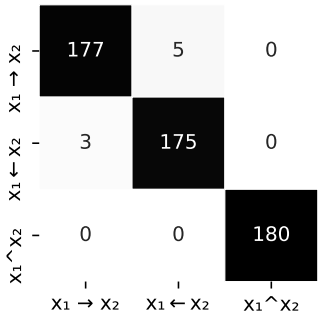


(c) $|\mathcal{R}_1| = 3$.

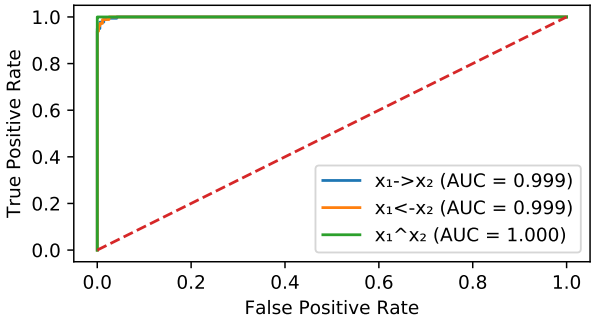
Figure 5.1. Convergence to the true marginal likelihood as the number particles increase, for each of the bivariate models.

true marginal likelihood and ELBO seems to be at a harmless level, and does not seem to be an obstacle for model selection.

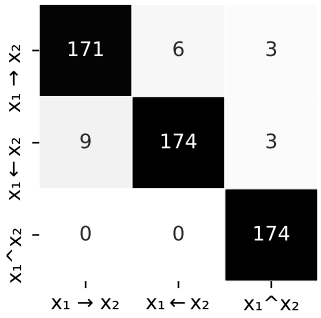
Note that in this section, we were not interested in selecting the true model, since it would be unfair to assess the model selection performance of the algorithms when they are supplied with only a few observations. However, it is apparent from the Figure 5.1, all of the methods are able to successfully detect the true causal relationship regardless of the assumed cardinality.



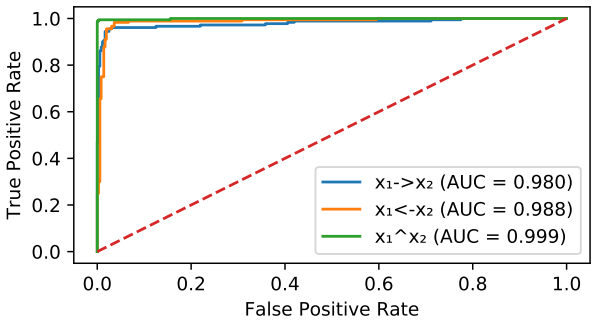
(a) The confusion matrix for SMC.



(b) ROC curves for SMC.



(c) The confusion matrix for VB.



(d) ROC curves for VB.

Figure 5.2. The confusion matrix and ROC curves of the SMC and VB algorithms for the synthetic data experiments.

5.3. Synthetic Data Experiments

In this experiment, for each of 36 hyperparameter combinations, and for each value of $|\mathcal{R}_1| \in \{1, \ldots, 5\}$, a total of three different data pairs (one for each causal hypothesis) with 2000 observations were generated. This amounted to a total of 540 data pairs. For each synthetic data pair, the corresponding hyperparameters were used to compare the three hypotheses demonstrated in Figure 3.1.

The resulting confusion matrix and ROC curves for SMC and VB algorithms can be seen in Figure 5.2. The results are quite impressive: The overall accuracy of SMC is 0.985, while the AUC values were 0.999, 0.999, and 1.0 for the classification of

each individual hypothesis. The performance of the VB is also competitive to SMC: The overall accuracy of VB is 0.961, while the AUC values were 0.980, 0.988, and 0.990. With these results, we verify that both algorithms are able to identify the data generating graph regardless of the given hyperparameter setting. In the light of such a successful performance, we can now safely test our model and algorithms for the real-world scenarios in the subsequent sections.

5.4. Cause Effect Pairs Data Set

To test the model selection performance of our framework on real-world scenarios, we used the version 1.0 of *Cause Effect Pairs* (CEP). As is common [31], the five data sets which included multivariate data were discarded (52-55 and 71), thus only the bivariate data sets were used. In order to be able to compare our methods with other methods numerically (such as those methods that are extensively tested in [31]), while testing our methods on CEP data set we examined bivariate direction detection and spurious relationships separately, since the data is labelled only with regard to the direction. That is, we forced our methods to make a decision with regard to direction of the given pair by only considering the models that include direct causal relationships. Afterwards we also examined the data pairs that were suggested to have a spurious relationship by our approach.

We tested our approach on the CEP data set by using 10×3 cross-validation. That is, we randomly divided the data set into three parts where each part is set apart once as the test set, and the remaining parts are used for training. For each split, we determined the parameter setting (of 36) that obtained the best accuracy on the training set, we then tested the performance of this setting on the test set. We conducted this whole split and test procedure 10 times in total. We report the accuracy and AUC values according to these 10 runs. For the SMC, we obtained a mean accuracy of 0.69 ± 0.08 and AUC score of 0.76 ± 0.13 (the values following the mean values correspond to 68% CI) where the accuracy and AUC calculations are performed by using the weights mentioned in [31]. Similarly for the VB, we obtained a mean accuracy of 0.70 ± 0.10 and AUC score of 0.78 ± 0.12. Mooij *et al.* (2016) compared

Table 5.2. Weighted accuracies of all hyperparameter settings along with the number of pairs classified to each hypothesis (either direction: "→", "←"; spurious: "∧").

(a) SMC

HYPERPARAMETER SETTING #	→	←	∧	ACCURACY
1	37	36	22	0.63
2	31	49	15	0.24
3	29	20	46	0.61
4	7	8	80	0.38
5	23	34	38	0.24
6	16	4	75	0.66
7	55	25	15	0.67
8	37	58	0	0.32
9	51	21	23	0.69
10	41	16	38	0.70
11	32	62	1	0.26
12	38	38	19	0.55
13	5	5	85	0.30
14	11	26	58	0.28
15	27	6	62	0.71
16	12	13	70	0.32
17	31	58	6	0.23
18	27	20	48	0.56
19	67	28	0	0.71
20	62	29	4	0.77
21	36	59	0	0.39
22	71	24	0	0.71
23	67	27	1	0.74
24	56	38	1	0.72
25	53	29	13	0.68
26	33	61	1	0.31
27	41	54	0	0.43
28	55	32	8	0.61
29	68	27	0	0.71
30	65	25	5	0.72
31	8	6	81	0.44
32	48	40	7	0.63
33	49	21	25	0.71
34	15	24	56	0.32
35	32	60	3	0.29
36	57	32	6	0.51

(b) VB

HYPERPARAMETER SETTING #	→	←	∧	ACCURACY
1	35	31	29	0.60
2	28	50	17	0.26
3	28	23	44	0.60
4	7	7	81	0.30
5	15	33	47	0.25
6	14	6	75	0.59
7	51	31	13	0.69
8	36	57	2	0.35
9	42	30	23	0.61
10	38	21	36	0.69
11	34	56	5	0.29
12	31	44	20	0.50
13	8	4	83	0.25
14	10	27	58	0.29
15	30	8	57	0.70
16	13	16	66	0.34
17	26	57	12	0.31
18	25	22	48	0.54
19	67	28	0	0.73
20	56	37	2	0.77
21	36	59	0	0.39
22	60	35	0	0.72
23	66	29	0	0.72
24	56	39	0	0.73
25	50	31	14	0.68
26	36	57	2	0.35
27	39	56	0	0.49
28	44	43	8	0.57
29	59	36	0	0.69
30	65	25	5	0.69
31	11	5	79	0.37
32	37	44	14	0.49
33	41	23	31	0.73
34	16	27	52	0.32
35	35	57	3	0.32
36	50	36	9	0.60

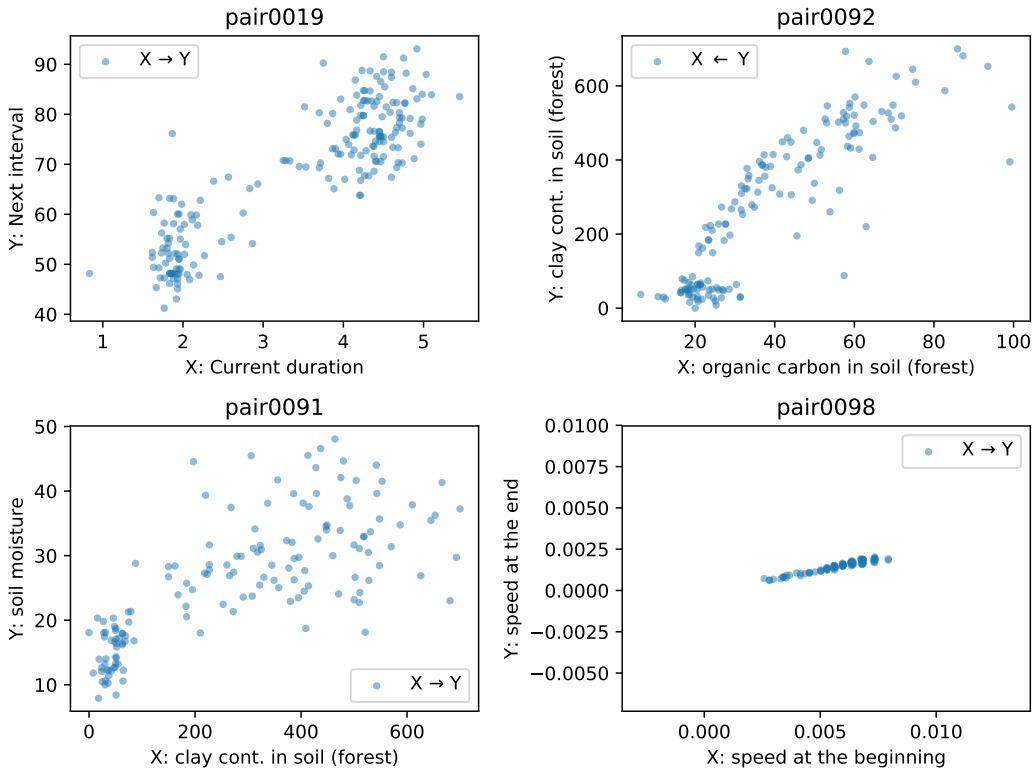


Figure 5.3. Scatter plots of spurious pairs found in Cause Effect Pairs.

most recent methods on their performance on the data set, and our scores can be seen to fall on the higher end of the performances by various methods, as only a few of the methods achieve a mean accuracy slightly higher than 0.70. Indeed, this little difference is not statistically significant due to the size of the data set and the variance of the mean accuracy estimators. Given these results, we conclude that our methods and model perform on par with the best of the currently existing methods in bivariate causality detection. We also present the overall accuracy of each hyperparameter setting for both methods in Table 5.2.

The CEP data set is not labeled as to the spurious relationships, therefore it is not possible to conduct hyperparameter selection with cross-validation. However, we ran the experiments again, this time including the spurious relationship hypothesis in the experiments, for all 36 parameter settings, and recorded the pairs for which the marginal likelihood of the spurious hypothesis was the highest. For both SMC and

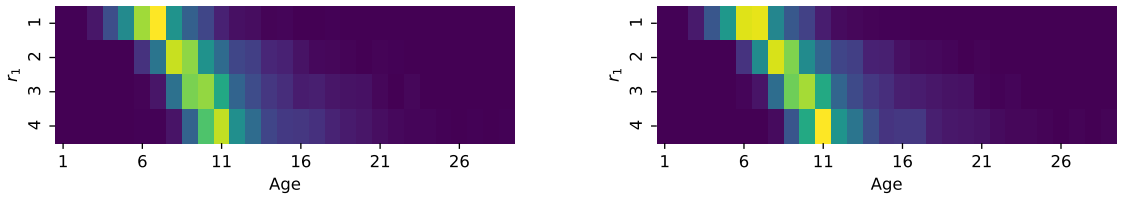
VB methods, we observed that the hyperparameter setting that achieved the highest accuracy in the previous experiment found these four data sets were to be spurious: 19, 91, 92, and 98. The scatter plots of these data sets are presented in Figure 5.3.

Visual examination of the first three pairs reveals that, although each of these pairs are correlated, they can be separated into two clusters in which X and Y axes become independent. In other words, once the confounding variables governing the cluster affiliations are decided, then the variables X and Y generated independently, so their correlation is indeed spurious. As we lack the expertise, we do not know what these confounding variables correspond in reality, but the existence of such variables is evident from the scatter plots. The case of the fourth spurious pair is slightly different than other correlated pairs. The fourth pair consists of the measurements of initial and final speeds of a ball on a ball track where initial speed is thought as the cause of final speed. However, both of our algorithms selected the spurious model with a latent variable having cardinality $|\mathcal{R}_1| = 1$, which actually corresponds to the marginal independence of X and Y. Such an explanation makes sense considering the plot in Figure 5.3, since the changes in the initial speed of the ball do not seem to affect the final speed, and therefore they must be independent. In the light of these results, we can also conclude that our model and algorithms are able to detect spurious relationships, even in the absence of previously labelled spurious examples.

5.5. Abalone Data Set

In this section we will apply our methodology to perform inference about the latent confounders on the popular *Abalone* data set [66]. The Abalone data set is a collection of observations including physical features, sex, and age from the marine animal abalone. All measurements except sex are real valued. Indeed, this data set is included in *CEP* as pairs in which the age variable is labelled as the cause of all the other physical features.

In this experiment, we used only two physical features, namely whole weight and height from Abalone data set, in our opinion which is the most likely spurious pair in



(a) Particle Gibbs.

(b) Dual EM.

Figure 5.4. Concurrence matrices of "$\boldsymbol{r}_1$ vs Age" that are constructed based on the outcomes from particle Gibbs and dual EM algorithms.

the data set to apply our methodology. We modelled them with the causal network in Figure 3.1(c) which states the spurious relationship. For both PG and DEM methods, we conduct our simulations with the 20^{th} hyperparameter setting in Table 5.1. We present the latent representations found by each method in Figure 5.4.

Age of the abalone is a reasonable feature to be the cause of the both weight and height, as it is also included in CEP data set as such. If the latent variables inferred by our methods are be able to function as a common cause, we expect them to be related to age. More clearly, if the latent variables found by our methods make the two physical features independent as the age, then we expect different levels of the latent variable to cluster the instances according to age, even without having the information about age. To see whether this is the case, we examine the age distributions of the abalones that are allocated to different levels of the latent variable. We present such an examination under the latent cardinality $|\mathcal{R}_1| = 4$ in Figure 5.4. As can be seen for both of the methods, the age distributions for each level of the hidden variable is different, potentially corresponding to young, medium, and older aged abalones. We therefore conclude that without observing the actual common cause, our methods are able to establish reliable latent representations like the actual cause.

6. CONCLUSION

We have presented a modeling framework which is potentially applicable to general graph structures including causally insufficient ones, and which can distinguish among Markov equivalent causal networks solely from observational data. We have shown that only by assuming the independence of cause-effect mechanisms, Bayesian model selection is sufficient to learn arbitrary causal structures. While the notion of *mechanism independence* can be made precise in many different ways, we provide a simple Bayesian interpretation in the form of *statistical independence* of parameters, and show that it is indeed a direct consequence of the *manipulation rule* of causal networks.

We believe that our framework is a promising unifying approach for causal structure learning, since it is in the intersection between graphical model based methods and structural equation models. The flexible nature of our framework makes it applicable to various problem contexts. Given that there is nothing in our approach that mandates the use of the specific distributions that we used, our approach can be extended to use other distributions or other variable types. Additionally, we tested our model only in its simplest form, i.e. we used linear basis functions for the conditional distributions. Given that we obtained a performance on the level of the state of the art on the CEP data set [31], the application of our framework to nonlinear basis functions and different conditional distributions is a very promising avenue of research.

On top of our modelling framework, we also developed two complementary approaches that enable us to make inference in causally insufficient contexts. In the experiments, we showed that the results obtained by these two algorithms are consistent with each other. Therefore it is possible to use any of these algorithms depending on the requirements of the problem, or use both of them to make up for each other's shortcomings. Variational algorithms work efficiently and accurately on the *large data* regime [67] but they are likely to fail otherwise; while Sequential Monte Carlo algorithms are more efficient and accurate on *sparse data* regime [50] and they are compu-

tationally intensive in the large data regime. However, given excessive computational power, it is possible to obtain estimations with arbitrary precision via SMC, whereas the precision of the variational approximation has theoretical limits. As well as being a complementary method for our causal framework, we believe that our SMC construction is a novel standalone contribution to inference methodology in general graphical models.

The existence of confounding variables often makes the discovery of truly causal associations difficult. To establish reliable causal links between variables, all these confounders should also be recorded while collecting data, and they should be taken into consideration while doing causal inference. In many empirical studies, however, determining all possible influential variables beforehand is most likely impossible. So, in our framework we account for these unmeasured confounders by explicitly modelling them as latent variables. By doing so, we obtained improved results for the causal direction determination, as well as detecting spurious relationships. Moreover, we developed a particle Gibbs algorithm and a dual EM algorithm to infer these latent variables. Inferring these latent variables allow us to identify how the nature of the underlying relationship (or lack thereof) between the observed variables are introduced, obscured, or changed by the latent variables. Thus in a sense, these algorithms allow us to *deconfound* the observations in the data set by sampling latent representations that are acting as the actual confounders.

In its current incarnation, our analysis regarding the causal mechanisms in nature involves learning a set of hyperparameters using cross-validation on labeled data. This approach seems valid from a machine learning perspective, but an alternative approach would be to conduct an extensive algebraic analysis to identify the prior settings that are implied by the causal laws of the nature. Such an analysis would also be beneficial for Bayesian model selection literature as well, as the relationship between prior distribution choices and inference is a subject of long-standing debate [68–70].

Although our framework can conduct causal discovery without the assumption of causal sufficiency; selection bias and potential cyclic relations can endanger the infer-

ence conducted by our framework. Future work should focus on adapting our approach for inference in presence of such potentially disturbing factors. Also, here we presented how our approach would work with observational data, but it is important to note that it can be straightforwardly adapted to work with partially or completely experimental data. This adaptability is likely to be crucial for many scientific endeavors in which collecting data through experimentation is either mandatory and/or sufficiently low cost. Another potential development pertains to the performance of the sampling algorithms, parallelization of which would greatly improve the computation time for inference. Overall, we believe that our approach constitutes a general, principled, and extendable Bayesian view to causal structure learning that can contribute substantively to causal discovery research.

REFERENCES

1. Pearl, J. and D. Mackenzie, *The book of why : the new science of cause and effect*, 2018.

2. Pearl, J., *Probabilistic reasoning in intelligent systems: networks of plausible inference*, Elsevier, 2014.

3. Lauritzen, S. L. and D. J. Spiegelhalter, "Local computations with probabilities on graphical structures and their application to expert systems", *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 157–224, 1988.

4. Spirtes, P., C. N. Glymour, R. Scheines, D. Heckerman, C. Meek, G. Cooper and T. Richardson, *Causation, prediction, and search*, MIT press, 1993.

5. Pearl, J., *Causality*, Cambridge university press, 2009.

6. Lauritzen, S. L., A. P. Dawid, B. N. Larsen and H.-G. Leimer, "Independence properties of directed Markov fields", *Networks*, vol. 20, no. 5, pp. 491–505, 1990.

7. Murphy, K. P., "Machine learning - a probabilistic perspective", *Adaptive computation and machine learning series*, 2012.

8. Scheines, R., "An introduction to causal inference", *Causality in crisis*, pp. 185–199, 1997.

9. Janzing, D. and B. Scholkopf, "Causal inference using the algorithmic Markov condition", *IEEE Transactions on Information Theory*, vol. 56, no. 10, pp. 5168–5194, 2010.

10. Daniusis, P., D. Janzing, J. M. Mooij, J. Zscheischler, B. Steudel, K. Zhang and B. Schölkopf, "Inferring deterministic causal relations", *CoRR*, vol. abs/1203.3475, 2012.

11. Schölkopf, B., D. Janzing, J. Peters, E. Sgouritsa, K. Zhang and J. Mooij, "On causal and anticausal learning", *arXiv preprint arXiv:1206.6471*, 2012.

12. Friedman, N., D. Geiger and M. Goldszmidt, "Bayesian network classifiers", *Machine learning*, vol. 29, no. 2-3, pp. 131–163, 1997.

13. Silva, R., R. Scheine, C. Glymour and P. Spirtes, "Learning the structure of linear latent variable models", *Journal of Machine Learning Research*, vol. 7, no. Feb, pp. 191–246, 2006.

14. Janzing, D., J. Peters, J. Mooij and B. Schölkopf, "Identifying confounders using additive noise models", *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 249–257, AUAI Press, 2009.

15. Shimizu, S., P. O. Hoyer and A. Hyvärinen, "Estimation of linear non-Gaussian acyclic models for latent factors", *Neurocomputing*, vol. 72, no. 7-9, pp. 2024–2027, 2009.

16. Chen, Z. and L. Chan, "Causality in linear nongaussian acyclic models in the presence of latent gaussian confounders", *Neural Computation*, vol. 25, no. 6, pp. 1605–1641, 2013.

17. Janzing, D., E. Sgouritsa, O. Stegle, J. Peters and B. Schölkopf, "Detecting low-complexity unobserved causes", *arXiv preprint arXiv:1202.3737*, 2012.

18. Schölkopf, B., D. W. Hogg, D. Wang, D. Foreman-Mackey, D. Janzing, C.-J. Simon-Gabriel and J. Peters, "Modeling confounding by half-sibling regression", *Proceedings of the National Academy of Sciences*, vol. 113, no. 27, pp. 7391–7398, 2016.

19. Zhang, K., J. Zhang and B. Schölkopf, "Distinguishing cause from effect based on exogeneity", *arXiv preprint arXiv:1504.05651*, 2015.

20. Heckerman, D., D. Geiger and D. M. Chickering, "Learning Bayesian networks: The combination of knowledge and statistical data", *Machine learning*, vol. 20, no. 3, pp. 197–243, 1995.

21. Spirtes, P., C. N. Glymour, R. Scheines, D. Heckerman, C. Meek, G. Cooper and T. Richardson, *Causation, prediction, and search*, MIT press, 2000.

22. Hausman, D. M. and J. Woodward, "Independence, invariance and the causal Markov condition", *The British journal for the philosophy of science*, vol. 50, no. 4, pp. 521–583, 1999.

23. Andrieu, C., A. Doucet and R. Holenstein, "Particle markov chain monte carlo methods", *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 72, no. 3, pp. 269–342, 2010.

24. Dempster, A. P., N. M. Laird and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

25. Spirtes, P. and C. Glymour, "An algorithm for fast recovery of sparse causal graphs", *Social science computer review*, vol. 9, no. 1, pp. 62–72, 1991.

26. Shimizu, S., P. O. Hoyer, A. Hyvärinen and A. Kerminen, "A linear non-Gaussian acyclic model for causal discovery", *Journal of Machine Learning Research*, vol. 7, no. Oct, pp. 2003–2030, 2006.

27. Hoyer, P. O., D. Janzing, J. M. Mooij, J. Peters and B. Schölkopf, "Nonlinear causal discovery with additive noise models", *Advances in neural information processing systems*, pp. 689–696, 2009.

28. Zhang, K. and A. Hyvärinen, "Distinguishing causes from effects using nonlinear acyclic causal models", *Proceedings of the 2008th International Conference on Causality: Objectives and Assessment-Volume 6*, pp. 157–164, JMLR. org, 2008.

29. Zhang, K. and A. Hyvärinen, "On the identifiability of the post-nonlinear causal model", *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pp. 647–655, AUAI Press, 2009.

30. Sipser, M. *et al.*, *Introduction to the Theory of Computation*, vol. 2, Thomson Course Technology Boston, 2006.

31. Mooij, J. M., J. Peters, D. Janzing, J. Zscheischler and B. Schölkopf, "Distinguishing cause from effect using observational data: methods and benchmarks", *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1103–1204, 2016.

32. Stegle, O., D. Janzing, K. Zhang, J. M. Mooij and B. Schölkopf, "Probabilistic latent variable models for distinguishing between cause and effect", *Advances in neural information processing systems*, pp. 1687–1695, 2010.

33. Shimizu, S. and K. Bollen, "Bayesian estimation of causal direction in acyclic structural equation models with individual-specific confounder variables and non-Gaussian distributions", *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2629–2652, 2014.

34. Zhang, K., Z. Wang, J. Zhang and B. Schölkopf, "On estimation of functional causal models: general results and application to the post-nonlinear causal model", *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 7, no. 2, p. 13, 2016.

35. Spirtes, P. and K. Zhang, "Causal discovery and inference: concepts and recent methodological advances", *Applied informatics*, vol. 3, p. 3, SpringerOpen, 2016.

36. Maathuis, M., M. Drton, S. Lauritzen and M. Wainwright (Editors), *Handbook of Graphical Models*, CRC Press, Boca Raton, Florida : CRC Press, c2019., 2018, URL `https://www.taylorfrancis.com/books/9780429874246`.

37. Pearl, J., *Probabilistic reasoning in intelligent systems : networks of plausible inference*, Morgan Kaufmann Publishers, 1988.

38. Barber, D., *Bayesian Reasoning and Machine Learning*, Cambridge University Press, 2012.

39. Lauritzen, S. L. and D. J. Spiegelhalter, "Local Computations with Probabilities on Graphical Structures and Their Application to Expert Systems", *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 50, no. 2, pp. 157–224, 1988, URL `http://www.jstor.org/stable/2345762`.

40. Lauritzen, S. L., *Graphical Models*, OUP Oxford, 1996.

41. Cowell, R. G., P. Dawid, S. L. Lauritzen and D. J. Spiegelhalter, *Probabilistic Networks and Expert Systems: Exact Computational Methods for Bayesian Networks*, Springer, 2003.

42. Cowell, R. G., P. Dawid, S. L. Lauritzen and D. J. Spiegelhalter, *Probabilistic networks and expert systems: Exact computational methods for Bayesian networks*, Springer Science & Business Media, 2006.

43. Heckerman, D., C. Meek and G. Cooper, "A Bayesian approach to causal discovery", *Innovations in Machine Learning*, pp. 1–28, Springer, 2006.

44. Kass, R. E. and A. E. Raftery, "Bayes Factors", *Journal of the American Statistical Association*, vol. 90, no. 430, pp. 773–795, 1995.

45. Spiegelhalter, D. J., A. P. Dawid, S. L. Lauritzen and R. G. Cowell, "Bayesian analysis in expert systems", *Statistical science*, pp. 219–247, 1993.

46. Blei, D. M., A. Y. Ng and M. I. Jordan, "Latent dirichlet allocation", *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

47. Geiger, D., D. Heckerman *et al.*, "A characterization of the Dirichlet distribution through global and local parameter independence", *The Annals of Statistics*, vol. 25, no. 3, pp. 1344–1369, 1997.

48. Bell, J. S. and J. S. Bell, *Speakable and unspeakable in quantum mechanics: Collected papers on quantum philosophy*, Cambridge university press, 2004.

49. Dawid, A. P., S. L. Lauritzen *et al.*, "Hyper Markov laws in the statistical analysis of decomposable graphical models", *The Annals of Statistics*, vol. 21, no. 3, pp. 1272–1317, 1993.

50. Doucet, A. and A. M. Johansen, "A tutorial on particle filtering and smoothing: Fifteen years later", *Handbook of nonlinear filtering*, vol. 12, no. 656-704, p. 3, 2009.

51. Doucet, A., N. De Freitas and N. Gordon, "An introduction to sequential Monte Carlo methods", *Sequential Monte Carlo methods in practice*, pp. 3–14, Springer, 2001.

52. Cemgil, A. T., M. B. Kurutmaz, S. Yildirim, M. Barsbey and U. Simsekli, "Bayesian Allocation Model: Inference by Sequential Monte Carlo for Nonneg-

ative Tensor Factorizations and Topic Models using Polya Urns", *arXiv preprint arXiv:1903.04478*, 2019.

53. Dawid, A. P., "Applications of a general propagation algorithm for probabilistic expert systems", *Statistics and computing*, vol. 2, no. 1, pp. 25–36, 1992.

54. Gilks, W. R., S. Richardson and D. Spiegelhalter, *Markov chain Monte Carlo in practice*, Chapman and Hall/CRC, 1995.

55. Liu, J. S., *Monte Carlo strategies in scientific computing*, Springer Science & Business Media, 2008.

56. Geman, S. and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images", *Readings in computer vision*, pp. 564–584, Elsevier, 1987.

57. Blei, D. M., A. Y. Ng and M. I. Jordan, "Latent Dirichlet Allocation", *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003, URL `http://dl.acm.org/citation.cfm?id=944919.944937`.

58. Blei, D. M., "Probabilistic topic models", *Communications of the ACM*, vol. 55, no. 4, p. 77, 2012, URL `http://doi.acm.org/10.1145/2133806.2133826`.

59. Cemgil, A. T., "Bayesian inference for nonnegative matrix factorisation models", *Computational Intelligence and Neuroscience*, vol. 2009, 2009.

60. Beal, M. J., Z. Ghahramani *et al.*, "Variational Bayesian learning of directed graphical models with hidden variables", *Bayesian Analysis*, vol. 1, no. 4, pp. 793–831, 2006.

61. Boyd, S. and L. Vandenberghe, *Convex optimization*, Cambridge university press, 2004.

62. Malham, S., "An introduction to Lagrangian and Hamiltonian mechanics", , 2015.

63. Blei, D. M., A. Kucukelbir and J. D. McAuliffe, "Variational inference: A review for statisticians", *Journal of the American Statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.

64. Chib, S. and T. A. Kuffner, "Bayes factor consistency", *arXiv preprint arXiv:1607.00292*, 2016.

65. Silander, T., P. Kontkanen and P. Myllymäki, "On Sensitivity of the MAP Bayesian Network Structure to the Equivalent Sample Size Parameter", *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, UAI'07, pp. 360–367, AUAI Press, Arlington, Virginia, United States, 2007, URL `http://dl.acm.org/citation.cfm?id=3020488.3020532`.

66. Dheeru, D. and E. Karra Taniskidou, "UCI Machine Learning Repository", , 2017, URL `http://archive.ics.uci.edu/ml`.

67. Wang, Y. and D. M. Blei, "Frequentist consistency of variational Bayes", *Journal of the American Statistical Association*, pp. 1–15, 2018.

68. Steck, H. and T. S. Jaakkola, "On the Dirichlet Prior and Bayesian Regularization", *In Advances in Neural Information Processing Systems 15*, pp. 697–704, MIT Press, 2002.

69. Letac, G. and H. Massam, "Bayes factors and the geometry of discrete hierarchical loglinear models", *Annals of Statistics*, vol. 40, no. 2, pp. 861–890, 2012.

70. Diaconis, P. and G. Wang, "Bayesian goodness of fit tests: a conversation for David Mumford", *arXiv preprint arXiv:1803.11251*, 2018.

APPENDIX A: EXPONENTIAL FAMILY REFRESHER

A.1. Basic Distributions

In this section, we supply the brief descriptions of the basic distributions that we mentioned in the main part of the thesis.

A.1.1. Gamma Distribution

(i) Gamma function:

$$\Gamma(z) \equiv \int_0^\infty x^{z-1} e^{-z} \, \mathrm{d}x$$

which is equal to $(z-1)!$ for nonnegative integer z.

(ii) Gamma density:

$$\mathrm{Gamma}(\rho; a, b) = \exp((a-1)\log\rho - b\rho - \log\Gamma(a) + a\log b)$$

where a is the *shape* and b is the *rate* parameter.

(iii) Expected sufficient statistics:

$$\mathsf{E}\{\rho\} = a/b, \quad \mathsf{E}\{\log\rho\} = \psi(a) - \log(b)$$

(iv) Cross entropy:

$$\begin{aligned}
&\mathsf{E}_{\mathrm{Gamma}(\rho;\hat{a},\hat{b})}\{-\log\mathrm{Gamma}(\rho; a, b)\} \\
&\qquad = -(a-1)\mathsf{E}\{\log\rho\} + b\mathsf{E}\{\rho\} + \log\Gamma(a) - a\log b \\
&\qquad = -(a-1)(\psi(\hat{a}) - \log(\hat{b})) + \frac{\hat{a}b}{\hat{b}} + \log\Gamma(a) - a\log b
\end{aligned}$$

Here, $\psi(x)$ is the *digamma* function which is defined as $\psi(x) = \frac{\mathrm{d}\log\Gamma(x)}{\mathrm{d}x}$.

A.1.2. Dirichlet Distribution

(i) Multivariate Beta function:

$$B(\gamma) = \frac{\prod_r \Gamma(\gamma_r)}{\Gamma(\sum_r \gamma_r)}$$

(ii) Dirichlet density:

$$\text{Dirichlet}(\theta;\gamma) = \frac{1}{B(\gamma)} \exp\big(\sum_r (\gamma_r - 1)\log\theta_r\big)$$

(iii) Expected sufficient statistics:

$$\mathsf{E}\left\{\theta_r\right\} = \frac{\gamma_r}{\sum_m \gamma_m}, \quad \mathsf{E}\left\{\log\theta_r\right\} = \psi(\gamma_r) - \psi(\sum_m \gamma_m)$$

(iv) Cross entropy:

$$\begin{aligned}\mathsf{E}_{\text{Dirichlet}(\theta;\hat{\gamma})}\left\{-\log\text{Dirichlet}(\theta;\gamma)\right\} &= \log B(\gamma) - \sum_r (\gamma_r - 1)\mathsf{E}\left\{\log\theta_r\right\} \\ &= \log B(\gamma) - \sum_r (\gamma_r - 1)\big(\psi(\gamma_r) - \psi(\sum_m \gamma_m)\big)\end{aligned}$$

A.1.3. Categorical Distribution

(i) Categorical density:

$$\text{Categorical}(r;\theta) = \prod_{k=1}^{K} \theta_k^{\mathbb{1}_{\{r=k\}}}$$

(ii) Expected sufficient statistics:

$$\mathsf{E}\left\{\mathbb{1}_{\{r=k\}}\right\} = \theta_k$$

(iii) Cross entropy:

$$\mathsf{E}_{\text{Categorical}(r;\hat{\theta})}\left\{-\log \text{Categorical}(r;\theta)\right\} = -\sum_k \hat{\theta}_k \log \theta_k$$

A.1.4. Normal Distribution

(i) Normal density:

$$x \sim \mathcal{N}(\mu, \rho^{-1}) \;\; = \;\; \frac{1}{\sqrt{2\pi}} \; \exp\Big(\frac{1}{2}\log\rho - \frac{1}{2}\rho(x-\mu)^2\Big)$$

where μ is the *mean* parameter and ρ is the *precision* parameter, i.e. ρ^{-1} is the variance.

(ii) Expected sufficient statistics

$$\mathsf{E}\left\{x\right\} = \mu \qquad\qquad \mathsf{E}\left\{x^2\right\} = \mu^2 + \rho^{-1}$$

A.1.5. Multivariate Normal Distribution

(i) Multivariate Normal density:

$$x \sim \mathcal{N}(\mu, \Lambda^{-1}) \;\; = \;\; \frac{1}{(2\pi)^{K/2}} \; \exp\Big(\frac{1}{2}\log\det(\Lambda) - \frac{1}{2}(x-\mu)^T\Lambda(x-\mu)\Big)$$

where μ is the *mean vector* and Λ is the *precision matrix*, i.e. Λ^{-1} is the covariance matrix.

(ii) Expected sufficient statistics:

$$\mathsf{E}\left\{x\right\} = \mu \qquad\qquad \mathsf{E}\left\{x^{\mathrm{T}}Ax\right\} = \mu^{\mathrm{T}}A\mu + \mathrm{tr}(\Lambda^{-1}A)$$

for any symmetric matrix A.

A.1.6. Normal-Gamma Distribution

(i) Normal-Gamma density:

$$\mu, \rho \sim \mathcal{NG}(m, \lambda, a, b) \;\; = \;\; \frac{b^a\sqrt{\lambda}}{\Gamma(a)\sqrt{2\pi}} \; \exp\Big(\big(a - \frac{1}{2}\big)\log\rho - b\rho - \frac{\lambda}{2}\rho(\mu - m)^2\Big)$$

which can be equivalently decomposed into a marginal Gamma distribution and a conditional Normal distribution:

$$\rho \sim \text{Gamma}(a, b) \qquad\qquad x \mid \rho \sim \mathcal{N}(m, (\lambda\rho)^{-1})$$

(ii) Expected sufficient statistics:

$$\mathsf{E}\left\{\log\rho\right\} = \psi(a) - \log b \quad \mathsf{E}\left\{\rho\right\} = \frac{a}{b} \quad \mathsf{E}\left\{\rho\mu\right\} = m\frac{a}{b} \quad \mathsf{E}\left\{\rho\mu^2\right\} = \frac{1}{\lambda} + m^2\frac{a}{b}$$

(iii) Cross entropy:

$$\begin{aligned}
&\mathsf{E}_{\mathcal{NG}(\hat{m},\hat{\lambda},\hat{a},\hat{b})}\left\{-\log\mathcal{NG}(\mu, \rho; m, \lambda, a, b)\right\} = \\
&\quad -a\log b + \log\Gamma(a) - \frac{1}{2}\log\lambda + \frac{\lambda}{2\hat{\lambda}} + \frac{1}{2}\log 2\pi \\
&\quad -\Big(a - \frac{1}{2}\Big)(\psi(\hat{a}) - \log\hat{b}) + \frac{\hat{a}b}{\hat{b}} + \frac{\hat{a}}{2\hat{b}}\lambda(\hat{m} - m)^2
\end{aligned}$$

A.1.7. Multivariate Normal-Gamma Distribution

(i) Multivariate Normal-Gamma density:

$$\begin{aligned}
w, \rho \;\; &\sim \;\; \mathcal{NG}(m, \Lambda, a, b) \\
&= \;\; \frac{b^a\sqrt{\det(\Lambda)}}{(2\pi)^{M/2}\Gamma(a)} \; \exp\left(\Big(a + \frac{M}{2} - 1\Big)\log\rho - b\rho - \frac{1}{2}\rho(w - m)^{\mathrm{T}}\Lambda(w - m)\right)
\end{aligned}$$

which can be equivalently decomposed into a marginal Gamma distribution and a conditional Multivariate Normal distribution:

$$\rho \sim \text{Gamma}(a, b) \qquad\qquad x \mid \rho \sim \mathcal{N}(m, (\rho\Lambda)^{-1})$$

(ii) Expected sufficient statistics:

$$\mathsf{E}\left\{\log \rho\right\} = \psi(a) - \log b \qquad \mathsf{E}\left\{\rho\right\} = \frac{a}{b} \qquad \mathsf{E}\left\{\rho w\right\} = \frac{a}{b} m$$
$$\mathsf{E}\left\{\rho\ w^{\mathrm{T}} A w\right\} = \mathrm{tr}(\Lambda^{-1} A) + \frac{a}{b} m^{\mathrm{T}} A m$$

for any symmetric matrix A.

(iii) Cross entropy:

$$\mathsf{E}_{\mathcal{NG}(\hat{m},\hat{\Lambda},\hat{a},\hat{b})}\left\{-\log \mathcal{NG}(w, \rho; m, \Lambda, a, b)\right\} =$$
$$-a \log b + \log \Gamma(a) - \frac{1}{2} \log \det(\Lambda) + \frac{1}{2} \mathrm{tr}(\hat{\Lambda}^{-1}\Lambda) + \frac{M}{2} \log 2\pi$$
$$-\Big(a + \frac{M}{2} - 1\Big)(\psi(\hat{a}) - \log \hat{b}) + \frac{\hat{a}b}{\hat{b}} + \frac{\hat{a}}{2\hat{b}}(\hat{m} - m)^{\mathrm{T}} \Lambda (\hat{m} - m)$$

A.2. Basic Conjugate Models

In this section we summarize the basic conjugate models that are closely related to our model in Chapter 3.

A.2.1. Dirichlet-Categorical Model

(i) Generative model:

$$\begin{aligned} \theta &\sim \text{Dirichlet}(\gamma) \\ r^1, \ldots, r^T &\sim \text{Categorical}(\theta) \end{aligned}$$

(ii) Posterior of θ:

$$\theta \mid r^1, \ldots, r^T \quad \sim \quad \text{Dirichlet}(\gamma^*)$$

where $\gamma_r^* = \gamma_r + \sum_{t=1}^{T} \mathbb{1}_{\{r=r^t\}}$

A.2.2. Normal-Gamma-Normal Model

(i) Generative model:

$$\begin{aligned} \mu, \rho &\sim \mathcal{NG}(m, \lambda, a, b) \\ x^1, \ldots, x^T &\sim \mathcal{N}(\mu, \rho^{-1}) \end{aligned}$$

(ii) Posterior of μ and ρ:

$$\mu, \rho \mid x^1, \ldots, x^T \quad \sim \quad \mathcal{NG}(m^*, \lambda^*, a^*, b^*)$$

where

$$\begin{aligned} \lambda^* &\equiv \lambda + T & m^* &\equiv \frac{\lambda m + \sum_t x^t}{\lambda^*} \\ a^* &\equiv a + \frac{T}{2} & b^* &\equiv b + \frac{1}{2}\big(\lambda m^2 - \lambda^* m^{*2} + \sum_t (x^t)^2\big) \end{aligned}$$

A.2.3. Bayesian Linear Regression

(i) Generative model:

$$y^t = w^{\mathrm{T}} x^t + \rho^{-1/2} \epsilon^t \qquad\qquad \epsilon^t \sim \mathcal{N}(0, 1)$$

An equivalent description with Normal-Gamma priors is

$$\begin{aligned} w, \rho &\sim \mathcal{NG}(m, \Lambda, a, b) \\ y^t \mid x^t &\sim \mathcal{N}(w^{\mathrm{T}} x^t, \rho^{-1}) \end{aligned}$$

(ii) Posterior of w and ρ:

$$w, \rho \mid (x^1, y^1), \ldots, (x^T, y^T) \sim \mathcal{NG}(m^*, \Lambda^*, a^*, b^*)$$

where

$$\begin{aligned} \Lambda^* &\equiv \Lambda + \sum\nolimits_t x^t x^{tT} & m^* &\equiv \Lambda^{*-1}\big(\Lambda m + \sum\nolimits_t y^t x^t\big) \\ a^* &\equiv a + \frac{T}{2} & b^* &\equiv b + \frac{1}{2}\Big(m^{\mathrm{T}}\Lambda m - m^{*\mathrm{T}}\Lambda^* m^* + \sum_t (x^t)^2\Big) \end{aligned}$$

APPENDIX B: MODEL DERIVATIONS

B.1. Posterior Distribution

In this section, we sketch the necessary derivation steps to calculate the posterior distribution for the model introduced in Chapter 3. Our goal is to show the posteriors of the parameters are again from the same family of the prior distributions:

$$
\begin{aligned}
&p(\theta_{1:K}, \rho_{1:N}, w_{1:N} \mid r_{1:K}^{1:T}, x_{1:N}^{1:T}) \propto p(\theta_{1:K}, \rho_{1:N}, w_{1:N})\ p(r_{1:K}^{1:T}, x_{1:N}^{1:T} \mid \theta_{1:K}, \rho_{1:N}, w_{1:N}) \\
&\quad = \Big(\prod_{k=1}^{K} \prod_{r_{\pi(\boldsymbol{r}_k)}} p(\theta_{k|r_{\pi(\boldsymbol{r}_k)}})\Big)\Big(\prod_{n=1}^{N} \prod_{r_{\pi(\boldsymbol{x}_n)}} p(w_{n|r_{\pi(\boldsymbol{x}_n)}}, \rho_{n|r_{\pi(\boldsymbol{x}_n)}})\Big) \\
&\quad \prod_{t=1}^{T}\Big(\prod_{k=1}^{K} p(r_k^t \mid \theta_{k|r_{\pi(\boldsymbol{r}_k)}^t})\Big)\Big(\prod_{n=1}^{N} p(x_n^t \mid x_{\pi(\boldsymbol{x}_n)}^t, w_{n|r_{\pi(\boldsymbol{x}_n)}^t}, \rho_{n|r_{\pi(\boldsymbol{x}_n)}^t})\Big) \\
&\quad = \Big(\prod_{k=1}^{K} \prod_{r_{\pi(\boldsymbol{r}_k)}} \text{Dirichlet}(\theta_{k|r_{\pi(\boldsymbol{r}_k)}}; \gamma_{k|r_{\pi(\boldsymbol{r}_k)}})\Big)\Big(\prod_{t=1}^{T}\prod_{k=1}^{K} \text{Categorical}(r_k^t; \theta_{k|r_{\pi(\boldsymbol{r}_k)}^t})\Big) \\
&\quad \Big(\prod_{n=1}^{N} \prod_{r_{\pi(\boldsymbol{x}_n)}} \mathcal{NG}(w_{n|r_{\pi(\boldsymbol{x}_n)}}, \rho_{n|r_{\pi(\boldsymbol{x}_n)}}; m_{n|r_{\pi(\boldsymbol{x}_n)}}, \Lambda_{n|r_{\pi(\boldsymbol{x}_n)}}, a_{n|r_{\pi(\boldsymbol{x}_n)}}, b_{n|r_{\pi(\boldsymbol{x}_n)}})\Big) \\
&\quad \Big(\prod_{t=1}^{T}\prod_{n=1}^{N} \mathcal{N}(x_n^t; {w_{n|r_{\pi(\boldsymbol{x}_n)}^t}}^{\mathrm{T}} \phi(x_{\pi(\boldsymbol{x}_n)}^t), {\rho_{n|r_{\pi(\boldsymbol{x}_n)}^t}}^{-1})\Big) \\
&\quad = \Big(\prod_{k=1}^{K} \prod_{r_{\pi(\boldsymbol{r}_k)}} \text{Dirichlet}(\theta_{k|r_{\pi(\boldsymbol{r}_k)}}; \gamma_{k|r_{\pi(\boldsymbol{r}_k)}}) \prod_{t=1}^{T} \text{Categorical}(r_k^t; \theta_{k|r_{\pi(\boldsymbol{r}_k)}})^{\mathbb{1}\left\{r_{\pi(\boldsymbol{r}_k)} = r_{\pi(\boldsymbol{r}_k)}^t\right\}}\Big) \\
&\quad \Big(\prod_{n=1}^{N} \prod_{r_{\pi(\boldsymbol{x}_n)}} \mathcal{NG}(w_{n|r_{\pi(\boldsymbol{x}_n)}}, \rho_{n|r_{\pi(\boldsymbol{x}_n)}}; m_{n|r_{\pi(\boldsymbol{x}_n)}}, \Lambda_{n|r_{\pi(\boldsymbol{x}_n)}}, a_{n|r_{\pi(\boldsymbol{x}_n)}}, b_{n|r_{\pi(\boldsymbol{x}_n)}}) \\
&\quad \prod_{t=1}^{T} \mathcal{N}(x_n^t; {w_{n|r_{\pi(\boldsymbol{x}_n)}}}^{\mathrm{T}} \phi(x_{\pi(\boldsymbol{x}_n)}^t), {\rho_{n|r_{\pi(\boldsymbol{x}_n)}}}^{-1})^{\mathbb{1}\left\{r_{\pi(\boldsymbol{x}_n)} = r_{\pi(\boldsymbol{x}_n)}^t\right\}}\Big) \\
&\quad \propto \Big(\prod_{k=1}^{K} \prod_{r_{\pi(\boldsymbol{r}_k)}} \text{Dirichlet}(\theta_{k|r_{\pi(\boldsymbol{r}_k)}}; \gamma_{k|r_{\pi(\boldsymbol{r}_k)}}^*)\Big) \\
&\quad \Big(\prod_{n=1}^{N} \prod_{r_{\pi(\boldsymbol{x}_n)}} \mathcal{NG}(w_{n|r_{\pi(\boldsymbol{x}_n)}}, \rho_{n|r_{\pi(\boldsymbol{x}_n)}}; m_{n|r_{\pi(\boldsymbol{x}_n)}}^*, \Lambda_{n|r_{\pi(\boldsymbol{x}_n)}}^*, a_{n|r_{\pi(\boldsymbol{x}_n)}}^*, b_{n|r_{\pi(\boldsymbol{x}_n)}}^*)\Big)
\end{aligned}
$$

The last step of the equation follows from the conjugate models we introduced in Appendix A.2.1, A.2.2 and A.2.3. The posterior parameters in the last equation is defined in terms of sufficient statistics of the data:

$$
\begin{aligned}
\gamma^*_{k|r_{\pi(\boldsymbol{r}_k)}}(r_k) &\equiv \gamma_{k|r_{\pi(\boldsymbol{r}_k)}} + \sum_{t=1}^{T} \mathbb{1}_{\left\{r^t_{\pi(\boldsymbol{r}_k)} = r_{\pi(\boldsymbol{r}_k)}\right\}} \mathbb{1}_{\left\{r^t_k = r_k\right\}} \\
\Lambda^*_{n|r_{\pi(\boldsymbol{x}n)}} &\equiv \Lambda_{n|r_{\pi(\boldsymbol{x}_n)}} + \sum_{t=1}^{T} \mathbb{1}_{\left\{r^t_{\pi(\boldsymbol{x}_n)} = r_{\pi(\boldsymbol{x}_n)}\right\}} \phi(x^t_{\pi(\boldsymbol{x}_n)}) \phi(x^t_{\pi(\boldsymbol{x}_n)})^{\mathrm{T}} \\
m^*_{n|r_{\pi(\boldsymbol{x}n)}} &\equiv {\Lambda^*_{n|r_{\pi(\boldsymbol{x}_n)}}}^{-1} \big(\Lambda_{n|r_{\pi(\boldsymbol{x}_n)}} m_{n|r_{\pi(\boldsymbol{x}_n)}} + \sum_{t=1}^{T} \mathbb{1}_{\left\{r^t_{\pi(\boldsymbol{x}_n)} = r_{\pi(\boldsymbol{x}_n)}\right\}} x^t_n \phi(x^t_{\pi(\boldsymbol{x}_n)}) \big) \\
a^*_{n|r_{\pi(\boldsymbol{x}n)}} &\equiv a_{n|r_{\pi(\boldsymbol{x}_n)}} + \frac{1}{2} \sum_{t=1}^{T} \mathbb{1}_{\left\{r^t_{\pi(\boldsymbol{x}_n)} = r_{\pi(\boldsymbol{x}_n)}\right\}} \\
b^*_{n|r_{\pi(\boldsymbol{x}n)}} &\equiv b_{n|r_{\pi(\boldsymbol{x}_n)}} + \frac{1}{2} \Big({m_{n|r_{\pi(\boldsymbol{x}_n)}}}^{\mathrm{T}} \Lambda_{n|r_{\pi(\boldsymbol{x}_n)}} m_{n|r_{\pi(\boldsymbol{x}_n)}} \\
&\quad - {m^*_{n|r_{\pi(\boldsymbol{x}_n)}}}^{\mathrm{T}} \Lambda^*_{n|r_{\pi(\boldsymbol{x}_n)}} m^*_{n|r_{\pi(\boldsymbol{x}_n)}} + \sum_{t=1}^{T} \mathbb{1}_{\left\{r^t_{\pi(\boldsymbol{x}_n)} = r_{\pi(\boldsymbol{x}_n)}\right\}} (x^t_n)^2 \Big)
\end{aligned}
$$

B.2. Marginal Distribution

To be able to utilize the *SIS-CN* algorithm, we also derive a closed form formula for the marginal distribution $p(r^{1:T}_{1:K}, x^{1:T}_{1:N})$ by simply using Bayes rule:

$$
\begin{aligned}
p(r^{1:T}_{1:K}, x^{1:T}_{1:N}) &= \frac{p(\theta_{1:K})\; p(w_{1:N}, \rho_{1:N})\; p(r^{1:T}_{1:K}, x^{1:T}_{1:N} \mid \theta_{1:K}, \rho_{1:N}, w_{1:N})}{p(\theta_{1:K}, \rho_{1:N}, w_{1:N} \mid r^{1:T}_{1:K}, x^{1:T}_{1:N})} \\
&= \Big(\prod_{k=1}^{K} \prod_{r_{\pi(\boldsymbol{r}_k)}} \frac{p(\theta_{k|r_{\pi(\boldsymbol{r}_k)}})}{p(\theta_{k|r_{\pi(\boldsymbol{r}_k)}} \mid r^{1:T}_{1:K})} \Big) \Big(\prod_{n=1}^{N} \prod_{r_{\pi(\boldsymbol{x}_n)}} \frac{p(w_{n|r_{\pi(\boldsymbol{x}_n)}}, \rho_{n|r_{\pi(\boldsymbol{x}_n)}})}{p(w_{n|r_{\pi(\boldsymbol{x}_n)}}, \rho_{n|r_{\pi(\boldsymbol{x}_n)}} \mid r^{1:T}_{1:K}, x^{1:T}_{1:N})} \Big) \\
&\quad \prod_{t=1}^{T} \Big(\prod_{k=1}^{K} p(r^t_k \mid \theta_{k|r^t_{\pi(\boldsymbol{r}_k)}}) \Big) \Big(\prod_{n=1}^{N} p(x^t_n \mid x^t_{\pi(\boldsymbol{x}_n)}, w_{n|r^t_{\pi(\boldsymbol{x}_n)}}, \rho_{n|r^t_{\pi(\boldsymbol{x}_n)}}) \Big) \\
&= \frac{1}{(2\pi)^{NT/2}} \left(\prod_{k=1}^{K} \prod_{r_{\pi(\boldsymbol{r}_k)}} \frac{\Gamma\big(\sum_{r_k} \gamma_{k|r_{\pi(\boldsymbol{r}_k)}}(r_k)\big)}{\Gamma\big(\sum_{r_k} \gamma^*_{k|r_{\pi(\boldsymbol{r}_k)}}(r_k)\big)} \prod_{r_k} \frac{\Gamma\big(\gamma^*_{k|r_{\pi(\boldsymbol{r}_k)}}(r_k)\big)}{\Gamma\big(\gamma_{k|r_{\pi(\boldsymbol{r}_k)}}(r_k)\big)} \right) \\
&\quad \left(\prod_{n=1}^{N} \prod_{r_{\pi(\boldsymbol{x}_n)}} \frac{{b_{n|r_{\pi(\boldsymbol{x}_n)}}}^{a_{n|r_{\pi(\boldsymbol{x}_n)}}}}{{b^*_{n|r_{\pi(\boldsymbol{x}_n)}}}^{a^*_{n|r_{\pi(\boldsymbol{x}_n)}}}} \frac{\Gamma(a^*_{n|r_{\pi(\boldsymbol{x}_n)}})}{\Gamma(a_{n|r_{\pi(\boldsymbol{x}_n)}})} \sqrt{\frac{\det(\Lambda_{n|r_{\pi(\boldsymbol{x}_n)}})}{\det(\Lambda^*_{n|r_{\pi(\boldsymbol{x}_n)}})}} \right)
\end{aligned}
$$

B.3. Derivations for Bivariate Models

In this section we present the necessary derivation steps of posterior and marginal distributions for the models (in Figure 3.1) that we used in the experiments. Although we have presented the results for a general graph, which directly implies the results of the bivariate case; for pedagogical purposes, we will explicitly state all the results related to bivariate graphs.

B.3.1. Causal Relationships

We will first derive the marginal distribution of the case that $\boldsymbol{x_1}$ is the cause of $\boldsymbol{x}_2$ (graphical model in Figure 3.1(a)), then the derivations for the opposite case (graphical model in Figure 3.1(b)) are straightforward due to symmetry.

(i) Generative Model:

$$\begin{aligned} \theta_1 &\sim \text{Dirichlet}(\gamma_1) & r_1^t &\sim \text{Categorical}(\theta_1) \\ w_{1|r_1}, \rho_{1|r_1} &\sim \mathcal{NG}(m_{1|r_1}, \Lambda_{1|r_1}, a_{1|r_1}, b_{1|r_1}) & x_1^t \mid r_1^t &\sim \mathcal{N}(w_{1|r_1^t}, \rho_{1|r_1^t}^{-1}) \\ w_{2|r_1}, \rho_{2|r_1} &\sim \mathcal{NG}(m_{2|r_1}, \Lambda_{2|r_1}, a_{2|r_1}, b_{2|r_1}) & x_2^t \mid x_1^t, r_1^t &\sim \mathcal{N}({w_{2|r_1^t}}^{\mathrm{T}}\phi(x_1^t), \rho_{2|r_1^t}^{-1}) \end{aligned}$$

(ii) Posterior Distribution of $\rho_{1:2}$, $w_{1:2}$ and θ_1

$$\begin{aligned} &p(\theta_1, \rho_{1:2}, w_{1:2} \mid r_1^{1:T}, x_{1:2}^{1:T}) \\ &\quad = \text{Dirichlet}(\theta_1; \gamma_1^*) \prod_{n=1}^{2} \prod_{r_1 \in \mathcal{R}_1} \mathcal{NG}(w_{n|r_1}, \rho_{n|r_1}; m_{n|r_1}^*, \Lambda_{n|r}^*, a_{n|r_1}^*, b_{n|r_1}^*) \end{aligned}$$

where the posterior parameters of ρ_1, w_1 and θ_1 are

$$\gamma_1^*(r_1) \equiv \gamma_1(r_1) + \sum_{t=1}^{T} \mathbb{1}_{\{r_1^t = r_1\}} \qquad \Lambda_{1|r_1}^* \equiv \Lambda_{1|r_1} + \sum_{t=1}^{T} \mathbb{1}_{\{r^t = r_1\}}$$

$$m_{1|r_1}^* \equiv \frac{\Lambda_{1|r_1} m_{1|r_1} + \sum_{t=1}^{T} \mathbb{1}_{\{r^t = r_1\}} x_1^t}{\Lambda_{1|r_1}^*} \qquad a_{1|r_1}^* \equiv a + \frac{1}{2} \sum_{t=1}^{T} \mathbb{1}_{\{r^t = r_1\}}$$

$$b_{1|r_1}^* \equiv b_{1|r_1} + \frac{1}{2}\big(\Lambda_{1|r_1} m_{1|r_1}^2 - \Lambda_{1|r_1}^* (m_{1|r_1}^*)^2 + \sum_{t=1}^{T} \mathbb{1}_{\{r^t = r_1\}} (x_1^t)^2\big)$$

whereas the posterior parameters of ρ_2 and w_2 are defined as

$$\begin{aligned}
\Lambda_{2|r_1}^* &\equiv \Lambda_{2|r_1} + \sum_{t=1}^{T} \mathbb{1}_{\{r_1^t = r_1\}} \phi(x_1^t)\phi(x_1^t)^T \\
m_{2|r_1}^* &\equiv {\Lambda_{2|r_1}^*}^{-1}\big(\Lambda_{2|r_1} m_{2|r_1} + \sum_{t=1}^{T} \mathbb{1}_{\{r_1^t = r\}} x_2^t \phi(x_1^t)\big) \\
a_{2|r_1}^* &\equiv a_{2|r_1} + \frac{1}{2} \sum_{t=1}^{T} \mathbb{1}_{\{r_1^t = r_1\}} \\
b_{2|r_1}^* &\equiv b_{2|r_1} + \frac{1}{2}\Big(m_{2|r_1}^{\mathrm{T}} \Lambda_{2|r_1} m_{2|r} - {m_{2|r}^*}^{\mathrm{T}} \Lambda_{2|r}^* m_{2|r}^* + \sum_{t=1}^{T} \mathbb{1}_{\{r_1^t = r_1\}} (x_2^t)^2\Big)
\end{aligned}$$

(iii) Marginal distribution of $r^{1:T}$, $x_{1:2}^{1:T}$

$$\begin{aligned}
p(r_1^{1:T}, x_{1:2}^{1:T}) \;=\; & \frac{1}{(2\pi)^T} \frac{\Gamma(\sum_{r_1 \in \mathcal{R}_1} \gamma_1(r_1))}{\Gamma(\sum_{r_1 \in \mathcal{R}_1} \gamma_1^*(r_1))} \prod_{r_1 \in \mathcal{R}_1} \frac{\Gamma(\gamma_1^*(r_1))}{\Gamma(\gamma_1(r_1))} \frac{b_{1|r_1}^{a_{1|r_1}}}{(b_{1|r_1}^*)^{a_{1|r_1}^*}} \frac{b_{2|r_1}^{a_{2|r_1}}}{(b_{2|r_1}^*)^{a_{2|r_1}^*}} \\
& \frac{\Gamma(a_{1|r_1}^*)}{\Gamma(a_{1|r_1})} \frac{\Gamma(a_{2|r_1}^*)}{\Gamma(a_{2|r_1})} \sqrt{\frac{\Lambda_{1|r_1}}{\Lambda_{1|r_1}^*}} \sqrt{\frac{\det(\Lambda_{2|r_1})}{\det(\Lambda_{2|r_1}^*)}}
\end{aligned}$$

B.3.2. Spurious Relationship

(i) Generative Model:

$$\begin{aligned}
\theta_1 &\sim \text{Dirichlet}(\gamma_1) & r_1^t &\sim \text{Categorical}(\theta_1) \\
w_{1|r_1}, \rho_{1|r_1} &\sim \mathcal{NG}(m_{1|r_1}, \Lambda_{1|r_1}, a_{1|r_1}, b_{1|r_1}) & x_1^t \mid r_1^t &\sim \mathcal{N}(w_{1|r_1^t}, \rho_{1|r_1^t}^{-1}) \\
w_{2|r_1}, \rho_{2|r_1} &\sim \mathcal{NG}(m_{2|r_1}, \Lambda_{2|r_1}, a_{2|r_1}, b_{2|r_1}) & x_2^t \mid r_1^t &\sim \mathcal{N}(w_{2|r_1^t}, \rho_{2|r_1^t}^{-1})
\end{aligned}$$

(ii) Posterior Distribution of $\rho_{1:2}$, $w_{1:2}$ and θ_1:

$$p(\theta_1, \rho_{1:2}, w_{1:2} \mid r_1^{1:T}, x_{1:2}^{1:T}) \\ = \text{Dirichlet}(\theta_1; \gamma_1^*) \prod_{n=1}^{2} \prod_{r_1 \in \mathcal{R}_1} \mathcal{NG}(w_{n|r_1}, \rho_{n|r_1}; m_{n|r_1}^*, \Lambda_{n|r}^*, a_{n|r_1}^*, b_{n|r_1}^*)$$

where the parameters of the posterior are defined in terms of sufficient statistics of the data:

$$\gamma_1^*(r_1) \equiv \gamma_1(r_1) + \sum_{t=1}^{T} \mathbb{1}_{\{r_1^t = r_1\}} \qquad \Lambda_{n|r_1}^* \equiv \Lambda_{n|r_1} + \sum_{t=1}^{T} \mathbb{1}_{\{r^t = r_1\}}$$

$$m_{n|r_1}^* \equiv \frac{\Lambda_{n|r_1} m_{n|r_1} + \sum_{t=1}^{T} \mathbb{1}_{\{r^t = r_1\}} x_n^t}{\Lambda_{n|r_1}^*} \qquad a_{n|r_1}^* \equiv a + \frac{1}{2} \sum_{t=1}^{T} \mathbb{1}_{\{r^t = r_1\}}$$

$$b_{n|r_1}^* \equiv b_{n|r_1} + \frac{1}{2}\big(\Lambda_{n|r_1} m_{n|r_1}^2 - \Lambda_{n|r_1}^* (m_{n|r_1}^*)^2 + \sum_{t=1}^{T} \mathbb{1}_{\{r^t = r_1\}} (x_n^t)^2\big)$$

(iii) Marginal distribution of $r_1^{1:T}$, $x_{1:2}^{1:T}$

$$p(r_1^{1:T}, x_{1:2}^{1:T}) = \\ \frac{1}{(2\pi)^T} \frac{\Gamma(\sum_{r_1 \in \mathcal{R}_1} \gamma_1(r_1))}{\Gamma(\sum_{r_1 \in \mathcal{R}_1} \gamma_1^*(r_1))} \prod_{r_1 \in \mathcal{R}_1} \frac{\Gamma(\gamma_1^*(r_1))}{\Gamma(\gamma_1(r_1))} \prod_{n=1}^{2} \frac{b_{n|r_1}^{a_{n|r_1}}}{(b_{n|r_1}^*)^{a_{n|r_1}^*}} \frac{\Gamma(a_{n|r_1}^*)}{\Gamma(a_{n|r_1})} \sqrt{\frac{\Lambda_{n|r_1}}{\Lambda_{n|r_1}^*}}$$

APPENDIX C: VARIATIONAL BAYES

C.1. Mean-Field Approximation

Let $\mathcal{P}(\boldsymbol{x}, \boldsymbol{y})$ be an arbitrary distribution on the random vectors $\boldsymbol{x}$ and $\boldsymbol{y}$. Our goal is to approximate this distribution with an instrumental distribution $\mathcal{Q}(\boldsymbol{x}, \boldsymbol{y})$ which factorizes into two independent parts:

$$\mathcal{Q}(\boldsymbol{x}, \boldsymbol{y}) \equiv q_x(\boldsymbol{x})\ q_y(\boldsymbol{y}) \tag{C.1}$$

We then express the mean-field approximation as a variational optimization problem:

$$\begin{aligned} \underset{q_x, q_y}{\text{minimize}} \quad & \mathrm{KL}(\mathcal{Q} \| \mathcal{P}) = \int_{y \in \mathcal{Y}} \int_{x \in \mathcal{X}} \mathcal{Q}(x, y) \log \frac{\mathcal{Q}(x, y)}{\mathcal{P}(x, y)}\ \mathrm{d}x\ \mathrm{d}y \\ \text{subject to} \quad & \mathcal{Q}(x, y) = q_x(x)\ q_y(y) \\ & \int_{x \in \mathcal{X}} q_x(x)\ \mathrm{d}x = 1 \\ & \int_{y \in \mathcal{Y}} q_y(y)\ \mathrm{d}y = 1 \end{aligned} \tag{C.2}$$

In order to solve this variational minimization problem, we construct a *Lagrangian functional* $\mathcal{L}$ by extending the domain of the original loss function with the scalar *Lagrange multipliers* λ_x and λ_y.

$$\begin{aligned} \mathcal{L}(q_x, q_y, \lambda_x, \lambda_y) \quad \equiv \quad & \int_{y \in \mathcal{Y}} \int_{x \in \mathcal{X}} q_x(x)\ q_y(y)\ \log \frac{q_x(x)\ q_y(y)}{\mathcal{P}(x, y)}\ \mathrm{d}x\ \mathrm{d}y \\ + \quad & \lambda_x \Big(1 - \int_{x \in \mathcal{X}} q_x(x)\ \mathrm{d}x \Big) + \lambda_y \Big(1 - \int_{y \in \mathcal{Y}} q_y(y)\ \mathrm{d}y \Big) \end{aligned}$$

Now, without loss of generality, we assume q_y is fixed and solve the corresponding *Euler-Lagrange* equations

$$\frac{\partial}{\partial q_x}\left(\int_{y\in\mathcal{Y}} q_x(x)\ q_y(y)\ \log\frac{q_x(x)\ q_y(y)}{\mathcal{P}(x,y)}\mathrm{d}y - \lambda_x q_x(x)\right) = 0 \tag{C.3}$$

$$\int_{x\in\mathcal{X}} q_x(x)\ \mathrm{d}x = 1 \tag{C.4}$$

where the evaluation of the partial derivative yields

$$\begin{aligned}
0 &= \frac{\partial}{\partial q_x}\left(\int_{y\in\mathcal{Y}} q_x(x)\ q_y(y)\ \log\frac{q_x(x)\ q_y(y)}{\mathcal{P}(x,y)}\mathrm{d}y - \lambda_x q_x(x)\right) \\
&= \int_{y\in\mathcal{Y}} q_y(y)\ \mathrm{d}y + \int_{y\in\mathcal{Y}} q(y)\log\frac{q_x(x)\ q_y(y)}{\mathcal{P}(x,y)}\ \mathrm{d}y - \lambda_x \\
&= 1 - \lambda_x + \mathsf{E}_{q_y}\left\{\log q_y(y)\right\} + \mathsf{E}_{q_y}\left\{\log q_x(x)\right\} - +\mathsf{E}_{q_y}\left\{\log\mathcal{P}(x,y)\right\} \\
&= 1 - \lambda_x + \mathsf{E}_{q_y}\left\{\log q_y(y)\right\} + \log q_x(x) - \mathsf{E}_{q_y}\left\{\log\mathcal{P}(x,y)\right\}
\end{aligned}$$

Here the terms $1-\lambda_x$ and $\mathsf{E}_{q_y}\left\{\log q_y(y)\right\}$ are constant. Therefore, at the extremum, the form of $\log q_x(x)$ up to an additive constant is found to be

$$\begin{aligned}
\log q_x(x) &=^{+} \mathsf{E}_{q_y}\left\{\log\mathcal{P}(x,y)\right\} \\
q_x(x) &\propto \exp\big(\mathsf{E}_{q_y}\left\{\log\mathcal{P}(x,y)\right\}\big)
\end{aligned}$$

By making use of the constraint in Equation C.4, we find the final form of q_x as the following:

$$q_x(x) = \frac{\exp\big(\mathsf{E}_{q_y}\left\{\log\mathcal{P}(x,y)\right\}\big)}{\int_{x\in\mathcal{X}}\exp\big(\mathsf{E}_{q_y}\left\{\log\mathcal{P}(x,y)\right\}\big)} \tag{C.5}$$

Due to symmetry, q_y also results in a similar form:

$$q_y(y) = \frac{\exp\big(\mathsf{E}_{q_x}\left\{\log\mathcal{P}(x,y)\right\}\big)}{\int_{y\in\mathcal{Y}}\exp\big(\mathsf{E}_{q_x}\left\{\log\mathcal{P}(x,y)\right\}\big)} \tag{C.6}$$

where both terms exhibit a circular dependency, and solving Equation C.5 and Equation C.6 together may not be practical. Therefore, in practice, these two equations are solved via fixed point iteration algorithms.

C.2. Variational Posterior

We have shown that minimization of $\mathrm{KL}(\mathcal{Q}\|\mathcal{P})$ ends up with the following marginal variational distributions:

$$q(\boldsymbol{r}_{1:K}^{1:T}) \propto \exp\big(\mathsf{E}_{q(\boldsymbol{\theta}_{1:K},\boldsymbol{\rho}_{1:N},\boldsymbol{w}_{1:N})}\left\{\log p(\boldsymbol{r}_{1:K}^{1:T}, x_{1:N}^{1:T}, \boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N})\right\}\big) \quad \text{(C.7)}$$

$$q(\boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N}) \propto \exp\big(\mathsf{E}_{q(\boldsymbol{r}_{1:K}^{1:T})}\left\{\log p(\boldsymbol{r}_{1:K}^{1:T}, x_{1:N}^{1:T}, \boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N})\right\}\big) \quad \text{(C.8)}$$

In this section, we will explicitly evaluate these equations to derive closed form expressions for the variational posteriors:

(i) We first simplify the Equation C.7 via factorization property of the joint distribution and removing the multiplicative constants

$$\begin{aligned}
q(\boldsymbol{r}_{1:K}^{1:T}) &\propto \exp\big(\mathsf{E}_{q(\boldsymbol{\theta}_{1:K},\boldsymbol{\rho}_{1:N},\boldsymbol{w}_{1:N})}\left\{\log p(\boldsymbol{r}_{1:K}^{1:T}, x_{1:N}^{1:T}, \boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N})\right\}\big) \\
&\propto \exp\big(\mathsf{E}_{q(\boldsymbol{\theta}_{1:K},\boldsymbol{\rho}_{1:N},\boldsymbol{w}_{1:N})}\left\{\log p(\boldsymbol{r}_{1:K}^{1:T}, x_{1:N}^{1:T} \mid \boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N})\right\}\big) \\
&= \prod_{t=1}^{T} \exp\big(\mathsf{E}_{q(\boldsymbol{\theta}_{1:K},\boldsymbol{\rho}_{1:N},\boldsymbol{w}_{1:N})}\left\{\log p(\boldsymbol{r}_{1:K}^{t}, x_{1:N}^{t} \mid \boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N})\right\}\big) \\
&\propto \prod_{t=1}^{T} q(\boldsymbol{r}_{1:K}^{t})
\end{aligned}$$

In order to keep the notation uncluttered, from now on we will omit the implicit subscripts in expectation operators. So each individual factor $q(\boldsymbol{r}_{1:K}^{t})$ above is equal to

$$
\begin{aligned}
q(r^t_{1:K}) &\propto \exp\big(\mathsf{E}\left\{\log p(r^t_{1:K}, x^t_{1:N} \mid \boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N})\right\}\big) \\
&= \exp\big(\sum_{k=1}^{K} \mathsf{E}\left\{\log p(r^t_k \mid r^t_{\pi(\boldsymbol{r}_k)}, \boldsymbol{\theta}_k)\right\} + \sum_{n=1}^{N} \mathsf{E}\left\{\log p(x^t_n \mid r^t_{\pi(\boldsymbol{x}_n)}, x^t_{\pi(\boldsymbol{x}_n)}, \boldsymbol{w_n}, \boldsymbol{\rho_n})\right\}\big) \\
&\propto \exp\Big(\sum_{k=1}^{K} \mathsf{E}\left\{\log \boldsymbol{\theta}_{k|r^t_{\pi(\boldsymbol{r}_k)}}(r^t_k)\right\} + \frac{1}{2}\sum_{n=1}^{N} \mathsf{E}\left\{\log \boldsymbol{\rho}_{n|r^t_{\pi(\boldsymbol{x}_n)}}\right\} \\
&- \frac{1}{2}\sum_{n=1}^{N} \mathsf{E}\left\{\boldsymbol{\rho}_{n|r^t_{\pi(\boldsymbol{x}_n)}} \big(\boldsymbol{w}_{n|r^t_{\pi(\boldsymbol{x}_n)}}{}^{\mathrm{T}} \phi(x^t_{\pi(\boldsymbol{x}_n)}) - x^t_n\big)^2\right\} \\
&\propto \text{Categorical}(r^t_{1:K}; \hat{\theta}^t)
\end{aligned}
$$

(ii) We now pursue the same strategy for the expression in Equation C.8

$$
\begin{aligned}
q(\theta_{1:K}, \rho_{1:N}, w_{1:N}) &\propto \exp\big(\mathsf{E}_{q(\boldsymbol{r}^{1:T}_{1:K})}\left\{\log p(\theta_{1:K}, \rho_{1:N}, w_{1:N} \mid \boldsymbol{r}^{1:T}_{1:K}, x^{1:T}_{1:N})\right\}\big) \\
&= \Big(\prod_{k=1}^{K}\prod_{r_{\pi(\boldsymbol{r}_k)}} \exp\big(\mathsf{E}\left\{\log p(\theta_{k|r_{\pi(\boldsymbol{r}_k)}} \mid \boldsymbol{r}^{1:T}_{1:K})\right\}\big)\Big) \\
&\Big(\prod_{n=1}^{N}\prod_{r_{\pi(\boldsymbol{x}_n)}} \exp\big(\mathsf{E}\left\{\log p(w_{n|r_{\pi(\boldsymbol{x}_n)}}, \rho_{n|r_{\pi(\boldsymbol{x}_n)}} \mid \boldsymbol{r}^{1:T}_{1:K}, x^{1:T}_{1:N})\right\}\big)\Big) \\
&\propto \Big(\prod_{k=1}^{K}\prod_{r_{\pi(\boldsymbol{r}_k)}} q(\theta_{k|r_{\pi(\boldsymbol{r}_k)}})\Big)\Big(\prod_{n=1}^{N}\prod_{r_{\pi(\boldsymbol{x}_n)}} q(w_{n|r_{\pi(\boldsymbol{x}_n)}}, \rho_{n|r_{\pi(\boldsymbol{x}_n)}})\Big)
\end{aligned}
$$

where each individual factor turns out to be

$$
\begin{aligned}
q(\theta_{k|r_{\pi(\boldsymbol{r}_k)}}) &\propto \exp\Big(\mathsf{E}\left\{\log p(\theta_{k|r_{\pi(\boldsymbol{r}_k)}} \mid \boldsymbol{r}^{1:T}_{1:K})\right\}\Big) \\
&\propto \exp\left(\sum_{r_k}\left(\gamma_{k|r_{\pi(\boldsymbol{r}_k)}} + \sum_{t=1}^{T} \mathsf{E}\left\{\mathbb{1}_{\{r^t_k = r_k\}}\mathbb{1}_{\{r^t_{\pi(\boldsymbol{r}_k)} = r_{\pi(\boldsymbol{r}_k)}\}}\right\} - 1\right)\log \theta_{k|r_{\pi(\boldsymbol{r}_k)}}(r_k)\right) \\
&\propto \text{Dirichlet}(\theta_{k|r_{\pi(\boldsymbol{r}_k)}}; \hat{\gamma}_{k|r_{\pi(\boldsymbol{r}_k)}})
\end{aligned}
$$

$$
\begin{aligned}
q(w_{n|r_{\pi(\boldsymbol{x}_n)}}, \rho_{n|r_{\pi(\boldsymbol{x}_n)}}) &\propto \exp\Big(\mathsf{E}\left\{\log p(w_{n|r_{\pi(\boldsymbol{x}_n)}}, \rho_{n|r_{\pi(\boldsymbol{x}_n)}} \mid \boldsymbol{r}_{1:K}^{1:T}, x_{1:N}^{1:T})\right\}\Big) \\
&\propto \exp\Big(\log p(w_{n|r_{\pi(\boldsymbol{x}_n)}}, \rho_{n|r_{\pi(\boldsymbol{x}_n)}}) + \mathsf{E}\left\{\log p(x_{1:N}^{1:T} \mid \boldsymbol{r}_{1:K}^{1:T}, w_{n|r_{\pi(\boldsymbol{x}_n)}}, \rho_{n|r_{\pi(\boldsymbol{x}_n)}})\right\}\Big) \\
&\propto \exp\Big(\log p(w_{n|r_{\pi(\boldsymbol{x}_n)}}, \rho_{n|r_{\pi(\boldsymbol{x}_n)}}) \\
&+ \sum_{t=1}^{T} \mathsf{E}\left\{\mathbb{1}_{\left\{r_{\pi(\boldsymbol{x}_n)}^t = r_{\pi(\boldsymbol{x}_n)}\right\}}\right\} \log p(x_n^t \mid x_{\pi(\boldsymbol{x}_n)}^t, w_{n|r_{\pi(\boldsymbol{x}_n)}}, \rho_{n|r_{\pi(\boldsymbol{x}_n)}})\Big) \\
&\propto \mathcal{NG}(w_{n|r_{\pi(\boldsymbol{x}_n)}}, \rho_{n|r_{\pi(\boldsymbol{x}_n)}}; \hat{m}_{n|r_{\pi(\boldsymbol{x}_n)}}, \hat{\Lambda}_{n|r_{\pi(\boldsymbol{x}_n)}}, \hat{a}_{n|r_{\pi(\boldsymbol{x}_n)}}, \hat{b}_{n|r_{\pi(\boldsymbol{x}_n)}})
\end{aligned}
$$

Finally, we match the coefficients of the sufficient statistics in above equations with the natural parameters and find the following variational parameters in terms of the expected sufficient statistics:

$$
\begin{aligned}
\log \hat{\theta}^t(r_{1:K}^t) &=^+ \sum_{k=1}^{K} \mathsf{E}_{\mathcal{Q}}\left\{\log \theta_{k|r_{\pi(\boldsymbol{r}_k)}^t}(r_k^t)\right\} - \frac{1}{2}\sum_{n=1}^{N} \phi(x_{\pi(\boldsymbol{x}_n)}^t)^{\mathrm{T}} \hat{\Lambda}_{n|r_{\pi(\boldsymbol{x}_n)}^t}^{-1} \phi(x_{\pi(\boldsymbol{x}_n)}^t) \\
&\quad + \frac{1}{2}\sum_{n=1}^{N} \mathsf{E}_{\mathcal{Q}}\left\{\log \rho_{n|r_{\pi(\boldsymbol{x}_n)}^t}\right\} - \frac{1}{2}\sum_{n=1}^{N} \big(\hat{m}_{n|r_{\pi(\boldsymbol{x}_n)}^t}^{\mathrm{T}} \phi(x_{\pi(\boldsymbol{x}_n)}^t) - x_n^t\big)^2 \mathsf{E}_{\mathcal{Q}}\left\{\rho_{n|r_{\pi(\boldsymbol{x}_n)}^t}\right\} \\
\hat{\gamma}_{k|r_{\pi(\boldsymbol{r}_k)}}(r_k) &= \gamma_{k|r_{\pi(\boldsymbol{r}_k)}} + \sum_{t=1}^{T} \mathsf{E}_{\mathcal{Q}}\left\{\mathbb{1}_{\left\{r_{\pi(\boldsymbol{r}_k)}^t = r_{\pi(\boldsymbol{r}_k)}\right\}} \mathbb{1}_{\left\{r_k^t = r_k\right\}}\right\} \\
\hat{\Lambda}_{n|r_{\pi(\boldsymbol{x}_n)}} &= \Lambda_{n|r_{\pi(\boldsymbol{x}_n)}} + \sum_{t=1}^{T} \mathsf{E}_{\mathcal{Q}}\left\{\mathbb{1}_{\left\{r_{\pi(\boldsymbol{x}_n)}^t = r_{\pi(\boldsymbol{x}_n)}\right\}}\right\} \phi(x_{\pi(\boldsymbol{x}_n)}^t)\phi(x_{\pi(\boldsymbol{x}_n)}^t)^{\mathrm{T}} \\
\hat{m}_{n|r_{\pi(\boldsymbol{x}_n)}} &= \hat{\Lambda}_{n|r_{\pi(\boldsymbol{x}_n)}}^{-1}\Big(\Lambda_{n|r_{\pi(\boldsymbol{x}_n)}} m_{n|r_{\pi(\boldsymbol{x}_n)}} + \sum_{t=1}^{T} \mathsf{E}_{\mathcal{Q}}\left\{\mathbb{1}_{\left\{r_{\pi(\boldsymbol{x}_n)}^t = r_{\pi(\boldsymbol{x}_n)}\right\}}\right\} x_n^t \phi(x_{\pi(\boldsymbol{x}_n)}^t)\Big) \\
\hat{a}_{n|r_{\pi(\boldsymbol{x}_n)}} &= a_{n|r_{\pi(\boldsymbol{x}_n)}} + \frac{1}{2}\sum_{t=1}^{T} \mathsf{E}_{\mathcal{Q}}\left\{\mathbb{1}_{\left\{r_{\pi(\boldsymbol{x}_n)}^t = r_{\pi(\boldsymbol{x}_n)}\right\}}\right\} \\
\hat{b}_{n|r_{\pi(\boldsymbol{x}_n)}} &= b_{n|r_{\pi(\boldsymbol{x}_n)}} + \frac{1}{2}\Big(m_{n|r_{\pi(\boldsymbol{x}_n)}}{}^{\mathrm{T}} \Lambda_{n|r_{\pi(\boldsymbol{x}_n)}} m_{n|r_{\pi(\boldsymbol{x}_n)}} \\
&\quad - \hat{m}_{n|r_{\pi(\boldsymbol{x}_n)}}^{\mathrm{T}} \hat{\Lambda}_{n|r_{\pi(\boldsymbol{x}_n)}} \hat{m}_{n|r_{\pi(\boldsymbol{x}_n)}} + \sum_{t=1}^{T} \mathsf{E}_{\mathcal{Q}}\left\{\mathbb{1}_{\left\{r_{\pi(\boldsymbol{x}_n)}^t = r_{\pi(\boldsymbol{x}_n)}\right\}}\right\} (x_n^t)^2\Big)
\end{aligned}
$$

C.3. Evidence Lower Bound

Remember that in Section 4.2.1 we expressed ELBO as a sum of expectation terms most of which are in the form of negative cross entropy or negative entropy:

$$
\begin{aligned}
&\mathcal{B}_{\mathcal{P}}[\mathcal{Q}] \equiv \mathsf{E}_{\mathcal{Q}}\left\{\log p(\boldsymbol{r}_{1:K}^{1:T}, x_{1:N}^{1:T}, \boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N}) - \log \mathcal{Q}(\boldsymbol{r}_{1:K}^{1:T}, \boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N})\right\} \\
&= \sum_{t=1}^{T}\sum_{n=1}^{N} \mathsf{E}_{\mathcal{Q}}\left\{\log p(x_n^t \mid x_{\pi(\boldsymbol{x}_n)}^t, \boldsymbol{r}_{\pi(\boldsymbol{x}_n)}^t, \boldsymbol{w}_n, \boldsymbol{\rho}_n)\right\} \\
&+ \sum_{t=1}^{T}\Big(\sum_{k=1}^{K} \mathsf{E}_{\mathcal{Q}}\left\{\log p(\boldsymbol{r}_k^t \mid \boldsymbol{r}_{\pi(\boldsymbol{r}_k)}^t, \boldsymbol{\theta}_k)\right\} - \mathsf{E}_{\mathcal{Q}}\left\{\log q(\boldsymbol{r}_{1:K}^t)\right\}\Big) \\
&+ \sum_{k=1}^{K}\sum_{r_{\pi(\boldsymbol{r}_k)}} \Big(\mathsf{E}_{\mathcal{Q}}\left\{\log p(\theta_{k|r_{\pi(\boldsymbol{r}_k)}})\right\} - \mathsf{E}_{\mathcal{Q}}\left\{\log q(\theta_{k|r_{\pi(\boldsymbol{r}_k)}})\right\}\Big) \\
&+ \sum_{n=1}^{N}\sum_{r_{\pi(\boldsymbol{x}_n)}} \Big(\mathsf{E}_{\mathcal{Q}}\left\{\log p(\boldsymbol{w}_{n|r_{\pi(\boldsymbol{x}_n)}}, \boldsymbol{\rho}_{n|r_{\pi(\boldsymbol{x}_n)}})\right\} - \mathsf{E}_{\mathcal{Q}}\left\{\log q(\boldsymbol{w}_{n|r_{\pi(\boldsymbol{x}_n)}}, \boldsymbol{\rho}_{n|r_{\pi(\boldsymbol{x}_n)}})\right\}\Big)
\end{aligned}
$$

In this section we will evaluate each of those expectations explicitly. We start our derivation with the trickier Gaussian log-likelihood term, then the rest of the expectations will correspond to negative cross entropy values of standard exponential family distributions:

$$
\begin{aligned}
&\mathsf{E}_{\mathcal{Q}}\left\{\log p(x_n^t \mid x_{\pi(\boldsymbol{x}_n)}^t, \boldsymbol{r}_{\pi(\boldsymbol{x}_n)}^t, \boldsymbol{w}_n, \boldsymbol{\rho}_n)\right\} \\
&\quad= \sum_{r_{\pi(\boldsymbol{x}_n)}} \mathsf{E}\left\{\mathbb{1}_{\left\{\boldsymbol{r}_{\pi(\boldsymbol{x}_n)}^t = r_{\pi(\boldsymbol{x}_n)}\right\}}\right\} \mathsf{E}\left\{\log p(x_n^t \mid x_{\pi(\boldsymbol{x}_n)}^t, \boldsymbol{w}_{n|r_{\pi(\boldsymbol{x}_n)}}, \boldsymbol{\rho}_{n|r_{\pi(\boldsymbol{x}_n)}})\right\} \\
&\quad= \frac{1}{2}\sum_{r_{\pi(\boldsymbol{x}_n)}} \mathsf{E}\left\{\mathbb{1}_{\left\{\boldsymbol{r}_{\pi(\boldsymbol{x}_n)}^t = r_{\pi(\boldsymbol{x}_n)}\right\}}\right\} \Big(\mathsf{E}\left\{\log \boldsymbol{\rho}_{n|r_{\pi(\boldsymbol{x}_n)}}\right\} \\
&\quad-\mathsf{E}\left\{\boldsymbol{\rho}_{n|r_{\pi(\boldsymbol{x}_n)}}\big(x_n^t - \boldsymbol{w}_{n|r_{\pi(\boldsymbol{x}_n)}}{}^{\mathrm{T}}\phi(x_{\pi(\boldsymbol{x}_n)}^t)\big)^2\right\} - \log 2\pi\Big) \\
&\quad= \frac{1}{2}\sum_{r_{\pi(\boldsymbol{x}_n)}}\sum_{r_{\neg\pi(\boldsymbol{x}_n)}} \hat{\theta}^t(r_{\pi(\boldsymbol{x}_n)}, r_{\neg\pi(\boldsymbol{x}_n)})\Big(\psi(\hat{a}_{n|r_{\pi(\boldsymbol{x}_n)}}) - \log \hat{b}_{n|r_{\pi(\boldsymbol{x}_n)}} \\
&\quad-\frac{\hat{a}_{n|r_{\pi(\boldsymbol{x}_n)}}}{\hat{b}_{n|r_{\pi(\boldsymbol{x}_n)}}}\big(x_n^t - \hat{m}_{n|r_{\pi(\boldsymbol{x}_n)}}^{\mathrm{T}}\phi(x_{\pi(\boldsymbol{x}_n)}^t)\big)^2 - \phi(x_{\pi(\boldsymbol{x}_n)}^t)^{\mathrm{T}}\hat{\Lambda}_{n|r_{\pi(\boldsymbol{x}_n)}}^{-1}\phi(x_{\pi(\boldsymbol{x}_n)}^t) - \log 2\pi\Big)
\end{aligned}
$$

Variational distribution $\mathcal{Q}$ treats $r^t_{1:K}$ and $\theta_{1:K}$ as independent variables. So, the expectations of the categorical log-likelihood terms admit the following form

$$\begin{aligned}\mathsf{E}_{\mathcal{Q}}\left\{\log p(\boldsymbol{r}^t_k \mid \boldsymbol{r}^t_{\pi(\boldsymbol{r}_k)}, \boldsymbol{\theta}_k)\right\} &= \sum_{r_k}\sum_{r_{\pi(\boldsymbol{r}_k)}} \mathsf{E}\left\{\mathbb{1}_{\{\boldsymbol{r}^t_k = r_k\}}\mathbb{1}_{\{\boldsymbol{r}^t_{\pi(\boldsymbol{r}_k)} = r_{\pi(\boldsymbol{r}_k)}\}}\right\}\mathsf{E}\left\{\log \theta_{k|r_{\pi(\boldsymbol{r}_k)}}(r_k)\right\} \\ &= \sum_{r_{1:K}} \hat{\theta}^t(r_{1:K})\Big(\psi\big(\hat{\gamma}_{k|r_{\pi(\boldsymbol{r}_k)}}(r_k)\big) - \psi\big(\sum_{r'_k}\hat{\gamma}_{k|r_{\pi(\boldsymbol{r}_k)}}(r'_k)\big)\Big)\end{aligned}$$

The rest of the terms are related to cross entropy or entropy of the well-known exponential family distributions, and closed form expressions for them are supplied in Appendix A. So here, we only modify these expressions by changing their parameters with the appropriate variational parameters.

(i) By using the negative cross entropy formulation in Appendix A.1.3 for categorical distributions:

$$\mathsf{E}_{\mathcal{Q}}\left\{\log q(\boldsymbol{r}^t_{1:K})\right\} = \sum_{r_{1:K}} \hat{\theta}^t(r_{1:K}) \log \hat{\theta}^t(r_{1:K})$$

(ii) By using the Dirichlet negative cross entropy formulation in Appendix A.1.2:

$$\begin{aligned}\mathsf{E}_{\mathcal{Q}}\left\{\log p(\boldsymbol{\theta}_{k|r_{\pi(\boldsymbol{r}_k)}})\right\} &= \log\Gamma\big(\sum_{r_k}\gamma_{k|r_{\pi(\boldsymbol{r}_k)}}(r_k)\big) - \sum_{r_k}\log\Gamma\big(\gamma_{k|r_{\pi(\boldsymbol{r}_k)}}(r_k)\big) \\ &\quad + \sum_{r_k}\big(\gamma_{k|r_{\pi(\boldsymbol{r}_k)}}(r_k) - 1\big)\Big(\psi\big(\hat{\gamma}_{k|r_{\pi(\boldsymbol{r}_k)}}(r_k)\big) - \psi\big(\sum_{r'_k}\hat{\gamma}_{k|r_{\pi(\boldsymbol{r}_k)}}(r'_k)\big)\Big)\end{aligned}$$

$$\begin{aligned}\mathsf{E}_{\mathcal{Q}}\left\{\log q(\boldsymbol{\theta}_{k|r_{\pi(\boldsymbol{r}_k)}})\right\} &= \log\Gamma\big(\sum_{r_k}\hat{\gamma}_{k|r_{\pi(\boldsymbol{r}_k)}}(r_k)\big) - \sum_{r_k}\log\Gamma\big(\hat{\gamma}_{k|r_{\pi(\boldsymbol{r}_k)}}(r_k)\big) \\ &\quad + \sum_{r_k}\big(\hat{\gamma}_{k|r_{\pi(\boldsymbol{r}_k)}}(r_k) - 1\big)\Big(\psi\big(\hat{\gamma}_{k|r_{\pi(\boldsymbol{r}_k)}}(r_k)\big) - \psi\big(\sum_{r'_k}\hat{\gamma}_{k|r_{\pi(\boldsymbol{r}_k)}}(r'_k)\big)\Big)\end{aligned}$$

(iii) Finally, by using the Multivariate Normal-Gamma negative cross entropy formulation in Appendix A.1.7:

$$\mathsf{E}_{\mathcal{Q}}\left\{\log p(\boldsymbol{w}_{n|r_{\pi(\boldsymbol{x}n)}}, \boldsymbol{\rho}_{n|r_{\pi(\boldsymbol{x}n)}})\right\} =$$
$$a_{n|r_{\pi(\boldsymbol{x}n)}} \log b_{n|r_{\pi(\boldsymbol{x}n)}} - \log\Gamma(a_{n|r_{\pi(\boldsymbol{x}n)}}) + \frac{1}{2}\log\det(\Lambda_{n|r_{\pi(\boldsymbol{x}n)}}) - \frac{1}{2}\mathrm{tr}(\hat{\Lambda}^{-1}_{n|r_{\pi(\mathbf{x}_n)}}\Lambda_{n|r_{\pi(\mathbf{x}_n)}})$$
$$-\frac{M}{2}\log 2\pi + \left(a_{n|r_{\pi(\boldsymbol{x}n)}} + \frac{M}{2} - 1\right)\left(\psi(\hat{a}_{n|r_{\pi(\boldsymbol{x}n)}}) - \log\hat{b}_{n|r_{\pi(\boldsymbol{x}n)}})\right) - b_{n|r_{\pi(\boldsymbol{x}n)}}\frac{\hat{a}_{n|r_{\pi(\boldsymbol{x}n)}}}{\hat{b}_{n|r_{\pi(\boldsymbol{x}n)}}}$$
$$-\frac{\hat{a}_{n|r_{\pi(\boldsymbol{x}n)}}}{2\hat{b}_{n|r_{\pi(\boldsymbol{x}n)}}}(\hat{m}_{n|r_{\pi(\boldsymbol{x}n)}} - m_{n|r_{\pi(\boldsymbol{x}n)}})^{\mathrm{T}}\Lambda_{n|r_{\pi(\boldsymbol{x}n)}}(\hat{m}_{n|r_{\pi(\boldsymbol{x}n)}} - m_{n|r_{\pi(\boldsymbol{x}n)}})$$

$$\mathsf{E}\left\{\log q(\boldsymbol{w}_{n|r_{\pi(\boldsymbol{x}n)}}, \boldsymbol{\rho}_{n|r_{\pi(\boldsymbol{x}n)}})\right\} =$$
$$\hat{a}_{n|r_{\pi(\boldsymbol{x}n)}} \log \hat{b}_{n|r_{\pi(\boldsymbol{x}n)}} - \log\Gamma(\hat{a}_{n|r_{\pi(\boldsymbol{x}n)}}) + \frac{1}{2}\log\det(\hat{\Lambda}_{n|r_{\pi(\boldsymbol{x}n)}}) - \frac{M}{2}$$
$$-\frac{M}{2}\log 2\pi + \left(\hat{a}_{n|r_{\pi(\boldsymbol{x}n)}} + \frac{M}{2} - 1\right)\left(\psi(\hat{a}_{n|r_{\pi(\boldsymbol{x}n)}}) - \log\hat{b}_{n|r_{\pi(\boldsymbol{x}n)}})\right) - \hat{a}_{n|r_{\pi(\boldsymbol{x}n)}}$$