PPV equations

Bart van Erp

November 10, 2020

Abstract

This document provides the derivations of variational messages for the first-order auto-regressive filter bank, also known as a probabilistic phase vocoder.

Contents

1 Overview

1.1 Context

Modelling signals is commonly performed by shaping the frequency spectrum or power spectral density. Auto-regressive processes have the property of having a deterministic power spectral density. If frequency components or Fourier coefficients are modelled using auto-regressive processes, the power spectral distribution of the time-domain signal has a particular shape. The power spectral distributions of the auto-regressive processes get upconverted to the frequency corresponding to the Fourier coefficients, shaping the power spectral distribution of the time-domain signal. The auto-regressive coefficients relate to the width of the lobes and the process noise to the corresponding power density of the lobes. In speech processing, this type of network is known as a probabilistic phase vocoder [1], [2].

1.2 Generative model

The generative model can be represented and factored as

$$p(\boldsymbol{y},\ \boldsymbol{x},\ \boldsymbol{\theta},\ \Lambda) = p(\boldsymbol{y} \mid \boldsymbol{x},\ \boldsymbol{\theta},\ \Lambda)\ p(\boldsymbol{x})\ p(\boldsymbol{\theta})\ p(\Lambda), \tag{1}$$

where the individual factors are found to be

$$p(\boldsymbol{y} \mid \boldsymbol{x},\ \boldsymbol{\theta},\ \Lambda) = \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{x} \circ \boldsymbol{\theta},\ \Lambda^{-1}), \tag{2}$$

$$p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_x,\ \Sigma_x), \tag{3}$$

$$p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta} \mid \boldsymbol{\mu}_\theta,\ \Sigma_\theta) \tag{4}$$

and

$$p(\Lambda) = \mathcal{W}(\Lambda \mid V_\Lambda,\ \nu_\Lambda), \tag{5}$$

where $\circ$ represents the Hadamard, or element-wise, product of two vectors or matrices and where $\mathcal{N}(\cdot)$ and $\mathcal{W}(\cdot)$ denote the Normal and Wishart distributions respectively.

Fig. 1 shows a schematic overview of the discussed model.

1.3 Mean-field factorization

In order to remove the dependencies between the different variables and make inference tractable, the generative model is approximated by its mean-field factorization

$$p(\boldsymbol{y},\ \boldsymbol{x},\ \boldsymbol{\theta},\ \Lambda) \approx q(\boldsymbol{y},\ \boldsymbol{x},\ \boldsymbol{\theta},\ \Lambda) = q(\boldsymbol{y})q(\boldsymbol{x})q(\boldsymbol{\theta})q(\Lambda), \tag{6}$$

where the individual factors are modelled as

$$q(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{m}_y,\ V_y), \tag{7}$$

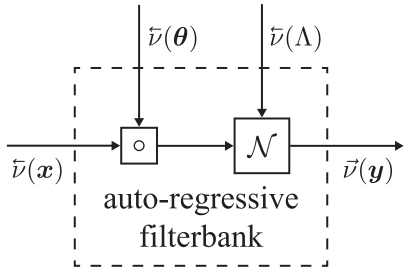

Figure 1: Schematic overview of the auto-regressive filterbank.

$$q(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{m}_x,\ V_x), \tag{8}$$

$$q(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta} \mid \boldsymbol{m}_\theta,\ V_\theta) \tag{9}$$

and

$$q(\Lambda) = \mathcal{W}(\Lambda \mid W_\Lambda, n_\Lambda). \tag{10}$$

2 Preliminaries

Before the derivations of the variational messages, several smaller derivations are discussed here in order to simplify analysis later on.

2.1 Hadamard product

The Hadamard product can be represented in several ways as

$$\begin{aligned}\boldsymbol{x} \circ \boldsymbol{\theta} &= \boldsymbol{\theta} \circ \boldsymbol{x} \\ &= \sum_{i=1}^{d} \boldsymbol{e}_i x_i \theta_i \\ &= A(\boldsymbol{x})\boldsymbol{\theta} \\ &= A(\boldsymbol{\theta})\boldsymbol{x} \\ &= \left(\sum_{i=1}^{d} \boldsymbol{e}_i \boldsymbol{e}_i^\top \underbrace{\boldsymbol{e}_i^\top \boldsymbol{\theta}}_{\text{scalar}} \right) \boldsymbol{x} \\ &= \left(\sum_{i=1}^{d} \boldsymbol{e}_i \boldsymbol{e}_i^\top \underbrace{\boldsymbol{e}_i^\top \boldsymbol{x}}_{\text{scalar}} \right) \boldsymbol{\theta}.\end{aligned} \tag{11}$$

Here $A(\boldsymbol{x})$ denotes a diagonal matrix with diagonal entries corresponding to $\boldsymbol{x}$ as

$$A(\boldsymbol{x}) = \begin{bmatrix} x_1 & 0 & \cdots & 0 \\ 0 & x_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & x_d \end{bmatrix}. \tag{12}$$

As this matrix only has diagonal entries, its transpose is identical to itself as

$$A(\boldsymbol{x})^\top = A(\boldsymbol{x}). \tag{13}$$

2.2 Expectation $A(\boldsymbol{x})$

A common expectation in the derivations below is the expectation with respect to the argument of the $A(\boldsymbol{x})$ matrix. This expectation can be found as

$$\begin{aligned}\mathrm{E}_{\boldsymbol{x}}\left[A(\boldsymbol{x})\right] &= \mathrm{E}_{\boldsymbol{x}}\left[\sum_{i=1}^{d} \boldsymbol{e}_i \boldsymbol{e}_i^\top \boldsymbol{e}_i^\top \boldsymbol{x}\right] \\ &= \sum_{i=1}^{d} \boldsymbol{e}_i \boldsymbol{e}_i^\top \boldsymbol{e}_i^\top \mathrm{E}_{\boldsymbol{x}}\left[\boldsymbol{x}\right] \\ &= A(\mathrm{E}_{\boldsymbol{x}}[\boldsymbol{x}]).\end{aligned} \tag{14}$$

2.3 Expectation Hadamard product

Using the previous result, the expectation over both independent variables in the Hadamard product can be evaluated as

$$\begin{aligned}
\mathrm{E}_{\boldsymbol{x}\boldsymbol{\theta}}\left[\boldsymbol{x} \circ \boldsymbol{\theta}\right] &= \mathrm{E}_{\boldsymbol{x}\boldsymbol{\theta}}\left[A(\boldsymbol{x})\boldsymbol{\theta}\right] \\
&= \mathrm{E}_{\boldsymbol{x}}[A(\boldsymbol{x})]\ \mathrm{E}_{\boldsymbol{\theta}}[\boldsymbol{\theta}] \\
&= A(\mathrm{E}_{\boldsymbol{x}}[\boldsymbol{x}])\ \mathrm{E}_{\boldsymbol{\theta}}[\boldsymbol{\theta}] \\
&= \mathrm{E}_{\boldsymbol{x}}[\boldsymbol{x}] \circ \mathrm{E}_{\boldsymbol{\theta}}[\boldsymbol{\theta}].
\end{aligned} \tag{15}$$

2.4 Expected value of a quadratic form

One expected value operation that commonly occurs when dealing with Normal distributions is the expected value of a quadratic form. This expected value can be determined as

$$\mathrm{E}_{\boldsymbol{x}}\left[\boldsymbol{x}^\top \Lambda \boldsymbol{x}\right] = \mathrm{Tr}\left[\Lambda \Sigma_x\right] + \boldsymbol{\mu}_x^\top \Lambda \boldsymbol{\mu}_x. \tag{16}$$

2.5 Trace-Hadamard relationship

Following theorem 7.20b from [3] the following relationship links the trace to the Hadamard product. This relationship reads

$$\boldsymbol{x}^\top (B \circ C)\boldsymbol{y} = \mathrm{Tr}\left[A(\boldsymbol{x})BA(\boldsymbol{x})C^\top\right]. \tag{17}$$

2.6 Trace-Hadamard equality

Here we will prove that

$$\mathrm{Tr}[C(A \circ B^\top)] = \mathrm{Tr}\left[(B \circ C)A\right] \tag{18}$$

holds. Previous works confirming this equality have not yet been found. In this proof we will use bracket notation $\{\cdot\}_{ij}$ to denote the entry at position (i, j) in a matrix.

Proof.

$$\begin{aligned}
\operatorname{Tr}\left[C(A \circ B^\top)\right] &= \sum_i \left\{C(A \circ B^\top)\right\}_{i,i} \\
&= \sum_i \sum_k \{C\}_{i,k} \left\{(A \circ B^\top)\right\}_{k,i} \\
&= \sum_i \sum_k \{C\}_{i,k} \{A\}_{k,i} \left\{B^\top\right\}_{k,i} \\
&= \sum_i \sum_k \{C\}_{i,k} \{A\}_{k,i} \{B\}_{i,k} \\
&= \sum_i \sum_k \{B\}_{i,k} \{C\}_{i,k} \{A\}_{k,i} \\
&= \sum_i \sum_k \{(B \circ C)\}_{i,k} \{A\}_{k,i} \\
&= \sum_i \{(B \circ C)A\}_{i,i} \\
&= \operatorname{Tr}\left[(B \circ C)A\right]
\end{aligned} \tag{19}$$

□

2.7 Expansion $\ln \mathcal{N}(\cdot)$

The multivariate normal distribution can be written as

$$\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_x,\ \Sigma_x) = (2\pi)^{-d/2} |\Sigma_x|^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_x)^\top \Sigma_x^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_x)\right\}. \tag{20}$$

Its natural logarithm can therefore be determined as

$$\ln \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_x,\ \Sigma_x) = -\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln(|\Sigma_x|) - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_x)^\top \Sigma_x^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_x), \tag{21}$$

in which the last term can be expanded as

$$\begin{aligned}
\ln \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_x,\ \Sigma_x) = -\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln(|\Sigma_x|) - \frac{1}{2}\boldsymbol{x}^\top \Sigma_x^{-1} \boldsymbol{x} + \frac{1}{2}\boldsymbol{x}^\top \Sigma_x^{-1} \boldsymbol{\mu}_x \\
+ \frac{1}{2}\boldsymbol{\mu}_x^\top \Sigma_x^{-1} \boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_x^\top \Sigma_x^{-1} \boldsymbol{\mu}_x.
\end{aligned} \tag{22}$$

2.8 Variational message passing

Suppose that we are dealing with the generative model $p(y, z)$ with an intractable posterior distribution $p(z \mid y)$, where y and z denote the observed and latent variables, respectively. The goal of variational inference is to approximate the intractable true posterior with a tractable variational distribution $q(z)$

through the minimization of the variational free energy functional

$$
\begin{aligned}
F[q] &= -\int q(z) \ln \frac{q(z)}{p(y, z)} \, \mathrm{d}z \\
&= \underbrace{-\ln p(y)}_{\text{- log-evidence}} + \underbrace{\int q(z) \ln \frac{q(z)}{p(z \mid y)} \, \mathrm{d}z}_{\text{KL-divergence}},
\end{aligned} \tag{23}
$$

which is also known as the negative evidence lower bound. In practice the optimization of (23) is performed by limiting $q(z)$ to a family of distributions, for which the hyperparameters are optimized. Since the first term of (23) is independent of $q(z)$, the problem reduces to the minimization of the Kullback-Leibler (KL) divergence.

The minimization of (23) can be achieved through sum-product message passing or VMP. Here the goal is to iterative update the variational distributions through coordinate descent of (23). The variational message $\nu(x_j)$ flowing out of a generic node $f(x_1, x_2, \ldots, x_n)$ with incoming messages $q(x_{\setminus j})$ can be determined as

$$
\vec{\nu}(x_j) \propto \exp \int \prod_{i \in \{1,\ldots,n\} \setminus j} q(x_i) \ln f(x_1, x_2, \ldots, x_n) \, \mathrm{d}\boldsymbol{x}_{\setminus j}, \tag{24}
$$

as presented in [4]. The updated variational distributions can be calculated through the multiplication of the forward and backward message on that respective edge as

$$
q(x_j) = \vec{\nu}(x_j) \cdot \overleftarrow{\nu}(x_j). \tag{25}
$$

In the current situation the node function $f(\cdot)$ is represented by the likelihood function $p(\boldsymbol{y} \mid \boldsymbol{x},\ \boldsymbol{\theta},\ \Lambda)$, whose natural logarithm has been determined in the previous subsection. As an example we convert the definition of the variational messages into a form that is more easily to be worked with. Given the message

$$
\vec{\nu}(\boldsymbol{y}) \propto \exp \int q(\boldsymbol{x}) q(\boldsymbol{\theta}) q(\Lambda) \ln \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{x} \circ \boldsymbol{\theta},\ \Lambda^{-1}) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{\theta} \, \mathrm{d}\Lambda, \tag{26}
$$

we can simplify this expression by taking the natural logarithm of both sides, consequently converting the proportionality to an equality with additive constant and by noting the expected value operation as

$$
\ln \vec{\nu}(\boldsymbol{y}) = \mathrm{E}_{q(\boldsymbol{x})q(\boldsymbol{\theta})q(\Lambda)} \left[\ln \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{x} \circ \boldsymbol{\theta},\ \Lambda^{-1}) \right] + \mathit{const.} \tag{27}
$$

In an attempt to keep notation uncluttered, the subscript of the expected value operator is simplified as

$$
\mathrm{E}_{\boldsymbol{x}\boldsymbol{\theta}\Lambda}[\cdot] \triangleq \mathrm{E}_{q(\boldsymbol{x})q(\boldsymbol{\theta})q(\Lambda)}[\cdot]. \tag{28}
$$

3 Variational message $\vec{\nu}(\boldsymbol{y})$

The outgoing message $\vec{\nu}(\boldsymbol{y})$ can be derived as

$$
\begin{aligned}
\ln \vec{\nu}(\boldsymbol{y}) &= \mathrm{E}_{\boldsymbol{x}\boldsymbol{\theta}\Lambda}\left[\ln \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{x} \circ \boldsymbol{\theta},\ \Lambda^{-1})\right] + \mathit{const} \\
&= \mathrm{E}_{\boldsymbol{x}\boldsymbol{\theta}\Lambda}\Bigg[-\frac{d}{2}\ln(2\pi) + \frac{1}{2}\ln(|\Lambda|) - \frac{1}{2}\boldsymbol{y}^\top \Lambda \boldsymbol{y} + \frac{1}{2}\boldsymbol{y}^\top \Lambda(\boldsymbol{x} \circ \boldsymbol{\theta}) \\
&\qquad + \frac{1}{2}(\boldsymbol{x} \circ \boldsymbol{\theta})^\top \Lambda \boldsymbol{y} - \frac{1}{2}(\boldsymbol{x} \circ \boldsymbol{\theta})^\top \Lambda (\boldsymbol{x} \circ \boldsymbol{\theta})\Bigg] + \mathit{const.}
\end{aligned}
\tag{29}
$$

By absorbing all terms independent of $\boldsymbol{y}$ in the constant term, we obtain

$$
\ln \vec{\nu}(\boldsymbol{y}) = \mathrm{E}_{\boldsymbol{x}\boldsymbol{\theta}\Lambda}\left[-\frac{1}{2}\boldsymbol{y}^\top \Lambda \boldsymbol{y} + \frac{1}{2}\boldsymbol{y}^\top \Lambda(\boldsymbol{x} \circ \boldsymbol{\theta}) + \frac{1}{2}(\boldsymbol{x} \circ \boldsymbol{\theta})^\top \Lambda \boldsymbol{y}\right] + \mathit{const.} \tag{30}
$$

As all terms depend linearly on the random variables over which we take the expectation, the expected value operator can be applied to all terms individually, resulting into

$$
\begin{aligned}
\ln \vec{\nu}(\boldsymbol{y}) = &-\frac{1}{2}\boldsymbol{y}^\top n_\Lambda W_\Lambda \boldsymbol{y} + \frac{1}{2}\boldsymbol{y}^\top n_\Lambda W_\Lambda(\boldsymbol{m}_x \circ \boldsymbol{m}_\theta) \\
&+ \frac{1}{2}(\boldsymbol{m}_x \circ \boldsymbol{m}_\theta)^\top n_\Lambda W_\Lambda \boldsymbol{y} + \mathit{const.}
\end{aligned}
\tag{31}
$$

From this description, the resemblance with a Normal distribution should become apparent. Rewriting the above expression by completing the square gives rise to

$$
\ln \vec{\nu}(\boldsymbol{y}) = -\frac{1}{2}\left[(\boldsymbol{y} - (\boldsymbol{m}_x \circ \boldsymbol{m}_\theta))^\top n_\Lambda W_\Lambda (\boldsymbol{y} - (\boldsymbol{m}_x \circ \boldsymbol{m}_\theta))\right] + \mathit{const}, \tag{32}
$$

from which the corresponding message can be determined as:

$$
\boxed{\vec{\nu}(\boldsymbol{y}) \propto \mathcal{N}\left(\boldsymbol{m}_x \circ \boldsymbol{m}_\theta,\ \frac{1}{n_\Lambda} W_\Lambda^{-1}\right)} \tag{33}
$$

4 Variational message $\overleftarrow{\nu}(\boldsymbol{x})$

The outgoing message $\overleftarrow{\nu}(\boldsymbol{x})$ can be derived as

$$\begin{aligned}\ln \overleftarrow{\nu}(\boldsymbol{x}) &= \mathrm{E}_{\boldsymbol{y}\boldsymbol{\theta}\Lambda}\left[\ln \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{x} \circ \boldsymbol{\theta},\ \Lambda^{-1})\right] + const \\ &= \mathrm{E}_{\boldsymbol{y}\boldsymbol{\theta}\Lambda}\left[-\frac{d}{2}\ln(2\pi) + \frac{1}{2}\ln(|\Lambda|) - \frac{1}{2}\boldsymbol{y}^\top \Lambda \boldsymbol{y} + \frac{1}{2}\boldsymbol{y}^\top \Lambda(\boldsymbol{x} \circ \boldsymbol{\theta}) \right. \\ &\qquad \left. + \frac{1}{2}(\boldsymbol{x} \circ \boldsymbol{\theta})^\top \Lambda \boldsymbol{y} - \frac{1}{2}(\boldsymbol{x} \circ \boldsymbol{\theta})^\top \Lambda (\boldsymbol{x} \circ \boldsymbol{\theta})\right] + const. \end{aligned} \tag{34}$$

By absorbing all terms independent of $\boldsymbol{x}$ in the constant term, we obtain

$$\begin{aligned}\ln \overleftarrow{\nu}(\boldsymbol{x}) = \frac{1}{2}\mathrm{E}_{\boldsymbol{y}\boldsymbol{\theta}\Lambda}\big[&\boldsymbol{y}^\top \Lambda(\boldsymbol{x} \circ \boldsymbol{\theta}) + (\boldsymbol{x} \circ \boldsymbol{\theta})^\top \Lambda \boldsymbol{y} \\ &-(\boldsymbol{x} \circ \boldsymbol{\theta})^\top \Lambda(\boldsymbol{x} \circ \boldsymbol{\theta})\big] + const. \end{aligned} \tag{35}$$

First, the expected value can be determined over Λ as this random variable only appears linearly in the different terms. This gives

$$\begin{aligned}\ln \overleftarrow{\nu}(\boldsymbol{x}) = \frac{1}{2}\mathrm{E}_{\boldsymbol{y}\boldsymbol{\theta}}\big[&\boldsymbol{y}^\top n_\Lambda W_\Lambda(\boldsymbol{x} \circ \boldsymbol{\theta}) + (\boldsymbol{x} \circ \boldsymbol{\theta})^\top n_\Lambda W_\Lambda \boldsymbol{y} \\ &-(\boldsymbol{x} \circ \boldsymbol{\theta})^\top n_\Lambda W_\Lambda(\boldsymbol{x} \circ \boldsymbol{\theta})\big] + const. \end{aligned} \tag{36}$$

Similarly, the expectation over $q(\boldsymbol{y})$ can be taken leading to

$$\begin{aligned}\ln \overleftarrow{\nu}(\boldsymbol{x}) = \frac{1}{2}\mathrm{E}_{\boldsymbol{\theta}}\Bigg[&\underbrace{\boldsymbol{m}_y^\top n_\Lambda W_\Lambda(\boldsymbol{x} \circ \boldsymbol{\theta})}_{1} + \underbrace{(\boldsymbol{x} \circ \boldsymbol{\theta})^\top n_\Lambda W_\Lambda \boldsymbol{m}_y}_{2} \\ &- \underbrace{(\boldsymbol{x} \circ \boldsymbol{\theta})^\top n_\Lambda W_\Lambda(\boldsymbol{x} \circ \boldsymbol{\theta})}_{3}\Bigg] + const. \end{aligned} \tag{37}$$

The expectation with respect to $q(\boldsymbol{\theta})$ for the remaining terms will be derived for each term individually.

$$\begin{aligned}\mathrm{E}_{\boldsymbol{\theta}}\underbrace{\left[\boldsymbol{m}_y^\top n_\Lambda W_\Lambda(\boldsymbol{x} \circ \boldsymbol{\theta})\right]}_{1} &= \mathrm{E}_{\boldsymbol{\theta}}\left[\boldsymbol{m}_y^\top n_\Lambda W_\Lambda A(\boldsymbol{x})\boldsymbol{\theta}\right] \\ &= \boldsymbol{m}_y^\top n_\Lambda W_\Lambda A(\boldsymbol{x}) \mathrm{E}_{\boldsymbol{\theta}}\left[\boldsymbol{\theta}\right] \\ &= \boldsymbol{m}_y^\top n_\Lambda W_\Lambda A(\boldsymbol{x}) \boldsymbol{m}_\theta \\ &= \boldsymbol{m}_y^\top n_\Lambda W_\Lambda A(\boldsymbol{m}_\theta) \boldsymbol{x}. \end{aligned} \tag{38}$$

$$
\begin{aligned}
\mathrm{E}_{\boldsymbol{\theta}} \underbrace{\left[(\boldsymbol{x} \circ \boldsymbol{\theta})^\top n_\Lambda W_\Lambda \boldsymbol{m}_y\right]}_{②} &= \mathrm{E}_{\boldsymbol{\theta}} \left[(A(\boldsymbol{x})\boldsymbol{\theta})^\top n_\Lambda W_\Lambda \boldsymbol{m}_y\right] \\
&= \mathrm{E}_{\boldsymbol{\theta}} \left[\boldsymbol{\theta}^\top A(\boldsymbol{x})^\top n_\Lambda W_\Lambda \boldsymbol{m}_y\right] \\
&= \mathrm{E}_{\boldsymbol{\theta}} \left[\boldsymbol{\theta}^\top\right] A(\boldsymbol{x}) n_\Lambda W_\Lambda \boldsymbol{m}_y \\
&= \boldsymbol{m}_\theta^\top A(\boldsymbol{x}) n_\Lambda W_\Lambda \boldsymbol{m}_y \\
&= \boldsymbol{x}^\top A(\boldsymbol{m}_\theta) n_\Lambda W_\Lambda \boldsymbol{m}_y \\
&= \boldsymbol{x}^\top (\boldsymbol{m}_y^\top n_\Lambda W_\Lambda^\top A(\boldsymbol{m}_\theta))^\top .
\end{aligned} \tag{39}
$$

$$
\begin{aligned}
\mathrm{E}_{\boldsymbol{\theta}} \underbrace{\left[(\boldsymbol{x} \circ \boldsymbol{\theta})^\top n_\Lambda W_\Lambda (\boldsymbol{x} \circ \boldsymbol{\theta})\right]}_{③} &= \mathrm{E}_{\boldsymbol{\theta}} \left[(A(\boldsymbol{x})\boldsymbol{\theta})^\top n_\Lambda W_\Lambda (A(\boldsymbol{x})\boldsymbol{\theta})\right] \\
&= \mathrm{E}_{\boldsymbol{\theta}} \left[\boldsymbol{\theta}^\top \left(A(\boldsymbol{x}) n_\Lambda W_\Lambda A(\boldsymbol{x})\right) \boldsymbol{\theta}\right] \\
&= \mathrm{Tr}\left[\left(A(\boldsymbol{x}) n_\Lambda W_\Lambda A(\boldsymbol{x})\right) V_\theta\right] + \boldsymbol{m}_\theta^\top \left(A(\boldsymbol{x}) n_\Lambda W_\Lambda A(\boldsymbol{x})\right) \boldsymbol{m}_\theta \\
&= \mathrm{Tr}\left[A(\boldsymbol{x}) n_\Lambda W_\Lambda A(\boldsymbol{x}) V_\theta\right] + \boldsymbol{m}_\theta^\top A(\boldsymbol{x}) n_\Lambda W_\Lambda A(\boldsymbol{x}) \boldsymbol{m}_\theta \\
&= \boldsymbol{x}^\top (n_\Lambda W_\Lambda \circ V_\theta^\top) \boldsymbol{x} + \boldsymbol{x}^\top A(\boldsymbol{m}_\theta) n_\Lambda W_\Lambda A(\boldsymbol{m}_\theta) \boldsymbol{x} \\
&= \boldsymbol{x}^\top \left((n_\Lambda W_\Lambda \circ V_\theta^\top) + A(\boldsymbol{m}_\theta) n_\Lambda W_\Lambda A(\boldsymbol{m}_\theta)\right) \boldsymbol{x}.
\end{aligned} \tag{40}
$$

From the last equation, the covariance matrix of the Normal message can be determined as

$$
\Sigma_{\bar{\nu}(\boldsymbol{x})} = \left((n_\Lambda W_\Lambda \circ V_\theta^\top) + A(\boldsymbol{m}_\theta) n_\Lambda W_\Lambda A(\boldsymbol{m}_\theta)\right)^{-1} \tag{41}
$$

and consequently the mean can be found as

$$
\boldsymbol{\mu}_{\bar{\nu}(\boldsymbol{x})} = \left((n_\Lambda W_\Lambda \circ V_\theta^\top) + A(\boldsymbol{m}_\theta) n_\Lambda W_\Lambda A(\boldsymbol{m}_\theta)\right)^{-1} (\boldsymbol{m}_y^\top n_\Lambda W_\Lambda A(\boldsymbol{m}_\theta))^\top . \tag{42}
$$

$$
\boxed{
\begin{aligned}
\bar{\nu}(\boldsymbol{x}) \propto \mathcal{N}\Big(&\left((n_\Lambda W_\Lambda \circ V_\theta^\top) + A(\boldsymbol{m}_\theta) n_\Lambda W_\Lambda A(\boldsymbol{m}_\theta)\right)^{-1} (\boldsymbol{m}_y^\top n_\Lambda W_\Lambda A(\boldsymbol{m}_\theta))^\top , \\
&\left((n_\Lambda W_\Lambda \circ V_\theta^\top) + A(\boldsymbol{m}_\theta) n_\Lambda W_\Lambda A(\boldsymbol{m}_\theta)\right)^{-1} \Big)
\end{aligned}
} \tag{43}
$$

5 Variational message $\overleftarrow{\nu}(\boldsymbol{\theta})$

Through the argument of symmetry with the variational message $\overleftarrow{\nu}(\boldsymbol{x})$, the variational message $\overleftarrow{\nu}(\boldsymbol{\theta})$ can be found as

$$\boxed{\begin{aligned}\overleftarrow{\nu}(\boldsymbol{\theta}) \propto \mathcal{N}\Big(&\big((n_\Lambda W_\Lambda \circ V_x^\top) + A(\boldsymbol{m}_x) n_\Lambda W_\Lambda A(\boldsymbol{m}_x)\big)^{-1} (\boldsymbol{m}_y^\top n_\Lambda W_\Lambda A(\boldsymbol{m}_x))^\top, \\ &\big((n_\Lambda W_\Lambda \circ V_x^\top) + A(\boldsymbol{m}_x) n_\Lambda W_\Lambda A(\boldsymbol{m}_x)\big)^{-1}\Big)\end{aligned}} \tag{44}$$

6 Variational message $\overleftarrow{\nu}(\Lambda)$

The outgoing message $\overleftarrow{\nu}(\Lambda)$ can be derived as

$$\begin{aligned}\ln \overleftarrow{\nu}(\Lambda) &= \mathrm{E}_{\boldsymbol{y}\boldsymbol{x}\boldsymbol{\theta}}\left[\ln \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{x} \circ \boldsymbol{\theta},\ \Lambda^{-1})\right] + const\\ &= \mathrm{E}_{\boldsymbol{y}\boldsymbol{x}\boldsymbol{\theta}}\left[-\frac{d}{2}\ln(2\pi) + \frac{1}{2}\ln(|\Lambda|) - \frac{1}{2}\boldsymbol{y}^\top\Lambda\boldsymbol{y} + \frac{1}{2}\boldsymbol{y}^\top\Lambda(\boldsymbol{x}\circ\boldsymbol{\theta})\right.\\ &\qquad\left. + \frac{1}{2}(\boldsymbol{x}\circ\boldsymbol{\theta})^\top\Lambda\boldsymbol{y} - \frac{1}{2}(\boldsymbol{x}\circ\boldsymbol{\theta})^\top\Lambda(\boldsymbol{x}\circ\boldsymbol{\theta})\right] + const.\end{aligned} \tag{45}$$

By shifting the terms independent of Λ to the constant, we obtain

$$\begin{aligned}\ln \overleftarrow{\nu}(\Lambda) = \mathrm{E}_{\boldsymbol{y}\boldsymbol{x}\boldsymbol{\theta}}&\left[\frac{1}{2}\ln(|\Lambda|) - \frac{1}{2}\boldsymbol{y}^\top\Lambda\boldsymbol{y} + \frac{1}{2}\boldsymbol{y}^\top\Lambda(\boldsymbol{x}\circ\boldsymbol{\theta})\right.\\ &\left. + \frac{1}{2}(\boldsymbol{x}\circ\boldsymbol{\theta})^\top\Lambda\boldsymbol{y} - \frac{1}{2}(\boldsymbol{x}\circ\boldsymbol{\theta})^\top\Lambda(\boldsymbol{x}\circ\boldsymbol{\theta})\right] + const.\end{aligned} \tag{46}$$

Let us start off with the expectation over $q(\boldsymbol{y})$. The terms in which y only appears as a linear factor, the expectation operator can be shifted to that respective term, e.g.

$$\mathrm{E}_{\boldsymbol{y}}\left[\frac{1}{2}\boldsymbol{y}^\top\Lambda(\boldsymbol{x}\circ\boldsymbol{\theta})\right] = \frac{1}{2}\mathrm{E}_{\boldsymbol{y}}[\boldsymbol{y}^\top]\Lambda(\boldsymbol{x}\circ\boldsymbol{\theta}). \tag{47}$$

The expectation over the quadratic second term $\boldsymbol{y}^\top\Lambda\boldsymbol{y}$ can be determined as

$$\mathrm{E}_{\boldsymbol{y}}[\boldsymbol{y}^\top\Lambda\boldsymbol{y}] = \mathrm{Tr}[\Lambda V_y] + \boldsymbol{m}_y^\top\Lambda\boldsymbol{m}_y. \tag{48}$$

Therefore the expectation over $q(\boldsymbol{y})$ can be determined as

$$\begin{aligned}\ln \overleftarrow{\nu}(\Lambda) = \frac{1}{2}\mathrm{E}_{\boldsymbol{x}\boldsymbol{\theta}}&\left[\ln(|\Lambda|) - \mathrm{Tr}[\Lambda V_y] - \boldsymbol{m}_y^\top\Lambda\boldsymbol{m}_y. + \boldsymbol{m}_y^\top\Lambda(\boldsymbol{x}\circ\boldsymbol{\theta})\right.\\ &\left. + (\boldsymbol{x}\circ\boldsymbol{\theta})^\top\Lambda\boldsymbol{m}_y - (\boldsymbol{x}\circ\boldsymbol{\theta})^\top\Lambda(\boldsymbol{x}\circ\boldsymbol{\theta})\right] + const.\end{aligned} \tag{49}$$

Next up let us calculate the expectation over $q(\boldsymbol{x})$. Before doing so, the equation is rewritten as

$$\begin{aligned}\ln \overleftarrow{\nu}(\Lambda) &= \frac{1}{2}\mathrm{E}_{\boldsymbol{x}\boldsymbol{\theta}}\left[\ln(|\Lambda|) - \mathrm{Tr}[\Lambda V_y] - \boldsymbol{m}_y^\top\Lambda\boldsymbol{m}_y + \boldsymbol{m}_y^\top\Lambda A(\boldsymbol{\theta})\boldsymbol{x}\right.\\ &\qquad\left. + (A(\boldsymbol{\theta})\boldsymbol{x})^\top\Lambda\boldsymbol{m}_y - (A(\boldsymbol{\theta})\boldsymbol{x})^\top\Lambda(A(\boldsymbol{\theta})\boldsymbol{x})\right] + const\\ &= \frac{1}{2}\mathrm{E}_{\boldsymbol{x}\boldsymbol{\theta}}\left[\ln(|\Lambda|) - \mathrm{Tr}[\Lambda V_y] - \boldsymbol{m}_y^\top\Lambda\boldsymbol{m}_y. + \boldsymbol{m}_y^\top\Lambda A(\boldsymbol{\theta})\boldsymbol{x}\right.\\ &\qquad\left. + \boldsymbol{x}^\top A(\boldsymbol{\theta})\Lambda\boldsymbol{m}_y - \boldsymbol{x}^\top A(\boldsymbol{\theta})\Lambda A(\boldsymbol{\theta})\boldsymbol{x}\right] + const.\end{aligned} \tag{50}$$

The expectation over the quadratic second term $\boldsymbol{x}^\top A(\boldsymbol{\theta})\Lambda A(\boldsymbol{\theta})\boldsymbol{x}$ can be determined as

$$\begin{aligned}\mathrm{E}_{\boldsymbol{x}}\left[\boldsymbol{x}^\top A(\boldsymbol{\theta})\Lambda A(\boldsymbol{\theta})\boldsymbol{x}\right] &= \mathrm{Tr}\left[A(\boldsymbol{\theta})\Lambda A(\boldsymbol{\theta})V_x\right] + \boldsymbol{m}_x^\top A(\boldsymbol{\theta})\Lambda A(\boldsymbol{\theta})\boldsymbol{m}_x\\ &= \boldsymbol{\theta}^\top(\Lambda\circ V_x^\top)\boldsymbol{\theta} + \boldsymbol{\theta}^\top A(\boldsymbol{m}_x)\Lambda A(\boldsymbol{m}_x)\boldsymbol{\theta}.\end{aligned} \tag{51}$$

Taking the entire expectation over $q(\boldsymbol{x})$ gives

$$\begin{aligned}\ln \tilde{\nu}(\Lambda) &= \frac{1}{2}\mathrm{E}_{\boldsymbol{\theta}}\left[\ln(|\Lambda|) - \mathrm{Tr}[\Lambda V_y] - \boldsymbol{m}_y^\top \Lambda \boldsymbol{m}_y + \boldsymbol{m}_y^\top \Lambda A(\boldsymbol{\theta})\boldsymbol{m}_x\right.\\ &\quad \left. +\boldsymbol{m}_x^\top A(\boldsymbol{\theta})\Lambda \boldsymbol{m}_y - \boldsymbol{\theta}^\top (\Lambda \circ V_x^\top)\boldsymbol{\theta} - \boldsymbol{\theta}^\top A(\boldsymbol{m}_x)\Lambda A(\boldsymbol{m}_x)\boldsymbol{\theta}\right] + const\\ &= \frac{1}{2}\mathrm{E}_{\boldsymbol{\theta}}\left[\ln(|\Lambda|) - \mathrm{Tr}[\Lambda V_y] - \boldsymbol{m}_y^\top \Lambda \boldsymbol{m}_y + \boldsymbol{m}_y^\top \Lambda A(\boldsymbol{m}_x)\boldsymbol{\theta}\right.\\ &\quad \left. +\boldsymbol{\theta}^\top A(\boldsymbol{m}_x)\Lambda \boldsymbol{m}_y - \boldsymbol{\theta}^\top (\Lambda \circ V_x^\top)\boldsymbol{\theta} - \boldsymbol{\theta}^\top A(\boldsymbol{m}_x)\Lambda A(\boldsymbol{m}_x)\boldsymbol{\theta}\right] + const.\end{aligned} \tag{52}$$

For the final expecation over $q(\boldsymbol{\theta})$ the linear terms can be familiarly determined. Now we are dealing with two quadratic terms, which we will determine first as

$$\mathrm{E}_{\boldsymbol{\theta}}\left[\boldsymbol{\theta}^\top (\Lambda \circ V_x^\top)\boldsymbol{\theta}\right] = \mathrm{Tr}\left[(\Lambda \circ V_x^\top)V_\theta\right] + \boldsymbol{m}_\theta^\top (\Lambda \circ V_x^\top)\boldsymbol{m}_\theta \tag{53}$$

and

$$\begin{aligned}\mathrm{E}_{\boldsymbol{\theta}}\left[\boldsymbol{\theta}^\top A(\boldsymbol{m}_x)\Lambda A(\boldsymbol{m}_x)\boldsymbol{\theta}\right] &= \mathrm{Tr}\left[A(\boldsymbol{m}_x)\Lambda A(\boldsymbol{m}_x)V_\theta\right] + \boldsymbol{m}_\theta^\top A(\boldsymbol{m}_x)\Lambda A(\boldsymbol{m}_x)\boldsymbol{m}_\theta\\ &= \boldsymbol{m}_x^\top (\Lambda \circ V_\theta^\top)\boldsymbol{m}_x + (\boldsymbol{m}_\theta \circ \boldsymbol{m}_x)^\top \Lambda(\boldsymbol{m}_\theta \circ \boldsymbol{m}_x).\end{aligned} \tag{54}$$

Substitution of these results gives

$$\begin{aligned}\ln \tilde{\nu}(\Lambda) = \frac{1}{2}&\left(\ln(|\Lambda|) - \mathrm{Tr}[\Lambda V_y] - \boldsymbol{m}_y^\top \Lambda \boldsymbol{m}_y + \boldsymbol{m}_y^\top \Lambda(\boldsymbol{m}_x \circ \boldsymbol{m}_\theta)\right.\\ &+ (\boldsymbol{m}_\theta \circ \boldsymbol{m}_x)^\top \Lambda \boldsymbol{m}_y - \mathrm{Tr}\left[(\Lambda \circ V_x^\top)V_\theta\right] - \boldsymbol{m}_\theta^\top (\Lambda \circ V_x^\top)\boldsymbol{m}_\theta\\ &\left.-\boldsymbol{m}_x^\top (\Lambda \circ V_\theta^\top)\boldsymbol{m}_x - (\boldsymbol{m}_\theta \circ \boldsymbol{m}_x)^\top \Lambda(\boldsymbol{m}_\theta \circ \boldsymbol{m}_x)\right) + const.\end{aligned} \tag{55}$$

By noting that the vector-matrix-vector products (e.g. $\boldsymbol{m}_y^\top \Lambda \boldsymbol{m}_y$) resemble scalars and by using the cyclic shift property of the trace operator, we can write

$$\boldsymbol{m}_y^\top \Lambda \boldsymbol{m}_y = \mathrm{Tr}\left[\boldsymbol{m}_y^\top \Lambda \boldsymbol{m}_y\right] = \mathrm{Tr}\left[\boldsymbol{m}_y \boldsymbol{m}_y^\top \Lambda\right]. \tag{56}$$

Rewriting the previous result, with the Hadamard-trace equality in mind, gives

$$\begin{aligned}\ln \tilde{\nu}(\Lambda) &= \frac{1}{2}\ln(|\Lambda|) - \frac{1}{2}\Big(\mathrm{Tr}[V_y\Lambda] + \mathrm{Tr}\left[\boldsymbol{m}_y \boldsymbol{m}_y^\top \Lambda\right] - \mathrm{Tr}\left[(\boldsymbol{m}_x \circ \boldsymbol{m}_\theta)\boldsymbol{m}_y^\top \Lambda\right]\\ &\quad - \mathrm{Tr}\left[\boldsymbol{m}_y(\boldsymbol{m}_\theta \circ \boldsymbol{m}_x)^\top \Lambda\right] + \mathrm{Tr}\left[(V_\theta \circ V_x)\Lambda\right] + \mathrm{Tr}\left[A(\boldsymbol{m}_\theta)V_x A(\boldsymbol{m}_\theta)\Lambda\right]\\ &\quad + \mathrm{Tr}\left[A(\boldsymbol{m}_x)V_\theta A(\boldsymbol{m}_x)\Lambda\right] + \mathrm{Tr}\left[(\boldsymbol{m}_\theta \circ \boldsymbol{m}_x)(\boldsymbol{m}_\theta \circ \boldsymbol{m}_x)^\top \Lambda\right]\Big) + const\\ &= \frac{1}{2}\ln(|\Lambda|) - \frac{1}{2}\Bigg(\mathrm{Tr}\Big[\big(V_y + \boldsymbol{m}_y \boldsymbol{m}_y^\top - (\boldsymbol{m}_x \circ \boldsymbol{m}_\theta)\boldsymbol{m}_y^\top\\ &\quad - \boldsymbol{m}_y(\boldsymbol{m}_\theta \circ \boldsymbol{m}_x)^\top + (V_\theta \circ V_x) + A(\boldsymbol{m}_\theta)V_x A(\boldsymbol{m}_\theta)\\ &\quad + A(\boldsymbol{m}_x)V_\theta A(\boldsymbol{m}_x) + (\boldsymbol{m}_\theta \circ \boldsymbol{m}_x)(\boldsymbol{m}_\theta \circ \boldsymbol{m}_x)^\top\big)\Lambda\Big]\Bigg) + const.\end{aligned} \tag{57}$$

From this results, the resemblance with the Wishart distribution should become clear. The degrees of freedom can be found as

$$\nu_{\bar{\nu}(\Lambda)} = p + 2, \tag{58}$$

where p refers to the dimensionality of $\Lambda \in \mathbb{R}^{p \times p}$. The scale matrix can be found as

$$\begin{aligned} V_{\bar{\nu}(\Lambda)} = \Big(&(\boldsymbol{m}_y + \boldsymbol{m}_\theta \circ \boldsymbol{m}_x)(\boldsymbol{m}_y + \boldsymbol{m}_\theta \circ \boldsymbol{m}_x)^\top + V_y + (V_\theta \circ V_x) \\ &+ A(\boldsymbol{m}_\theta) V_x A(\boldsymbol{m}_\theta) + A(\boldsymbol{m}_x) V_\theta A(\boldsymbol{m}_x) \Big)^{-1}. \end{aligned} \tag{59}$$

The outgoing message can be determined as

$$\boxed{\begin{aligned} \bar{\nu}(\Lambda) \propto \mathcal{W}\Bigg(\Big(&(\boldsymbol{m}_y + \boldsymbol{m}_\theta \circ \boldsymbol{m}_x)(\boldsymbol{m}_y + \boldsymbol{m}_\theta \circ \boldsymbol{m}_x)^\top + (V_\theta \circ V_x) \\ &+ V_y + A(\boldsymbol{m}_\theta) V_x A(\boldsymbol{m}_\theta) + A(\boldsymbol{m}_x) V_\theta A(\boldsymbol{m}_x) \Big)^{-1},\ p + 2 \Bigg) \end{aligned}} \tag{60}$$

References

[1] A. T. Cemgil and S. J. Godsill, "Probabilistic phase vocoder and its application to interpolation of missing values in audio signals," in *2005 13th European Signal Processing Conference*, pp. 1–4, Sept. 2005. ISSN: null.

[2] R. E. Turner and M. Sahani, "Time-Frequency Analysis as Probabilistic Inference," *IEEE Transactions on Signal Processing*, vol. 62, pp. 6171–6183, Dec. 2014.

[3] J. R. Schott, *Matrix Analysis for Statistics.* John Wiley & Sons, June 2016. Google-Books-ID: Y2PpCgAAQBAJ.

[4] J. Dauwels, "On Variational Message Passing on Factor Graphs," in *2007 IEEE International Symposium on Information Theory*, pp. 2546–2550, June 2007. ISSN: 2157-8117.