Supplement for:

Fast Approximate Inference for Arbitrarily Large Semiparametric Regression Models via Message Passing

BY M.P. WAND[1]

S.1 Exponential Family Theory and Results

The sufficient statistic and log-partition function are linked by the results

$$E\{\boldsymbol{T}(\boldsymbol{x})\} = \mathsf{D}_{\boldsymbol{\eta}} A(\boldsymbol{\eta})^T \quad \text{and} \quad \text{Cov}\{\boldsymbol{T}(\boldsymbol{x})\} = \mathsf{D}_{\boldsymbol{\eta}}\{\mathsf{D}_{\boldsymbol{\eta}} A(\boldsymbol{\eta})^T\} \tag{S.1}$$

where $\text{Cov}\{\boldsymbol{T}(\boldsymbol{x})\}$ is the covariance matrix of $\boldsymbol{T}(\boldsymbol{x})$, and for $\boldsymbol{f}$ a $\mathbb{R}^p$-valued function with argument $\boldsymbol{x} \in \mathbb{R}^d$, $\mathsf{D}_{\boldsymbol{x}} f(\boldsymbol{x})$ is the $p \times d$ matrix whose (i, j) entry is $\partial \boldsymbol{f}(\boldsymbol{x})_i / \partial x_j$. The first expression in (S.1) is particularly important for variational message passing since the messages from factors to stochastic nodes in conjugate factor graphs reduce to sufficient statistic expectations.

The digamma function, denoted by ψ, is

$$\psi(x) \equiv \frac{d}{dx} \log \Gamma(x).$$

Evaluation of ψ is supported in the MATLAB computing environment (The Mathworks Incorporated, 2015) via the function `psi()` and in the R computing environment (R Core Team, 2015) via the function `digamma()`.

The exponential integral function is

$$\text{Ei}(x) \equiv -\int_{-x}^{\infty} \frac{\exp(-t)}{t}\, dt, \quad x \in \mathbb{R}\backslash\{0\}. \tag{S.2}$$

Evaluation of Ei is supported in the MATLAB via the function `expint()`, which returns values of $-\text{Ei}(-x)$ for an input x, and in R via the function `expint_Ei()` within the package gsl (Hankin, 2007).

S.1.1 Bernoulli Distribution

The probability mass function of the Bernoulli distribution with probability of success $\wp \in (0, 1)$ is

$$p(x) = \wp^x (1 - \wp)^{1-x}, \quad x \in \{0, 1\}.$$

The sufficient statistic and base measure are

$$T(x) = x \quad \text{and} \quad h(x) = I(x \in \{0, 1\}).$$

The natural parameter vector and its inverse mapping are

$$\eta = \log\{\wp/(1 - \wp)\} \quad \text{and} \quad \wp = e^\eta/(1 + e^\eta)$$

and the log-partition function is

$$A(\eta) = \log(1 + e^\eta).$$

[1]M.P. Wand is Distinguished Professor, School of Mathematical and Physical Sciences, University of Technology Sydney, P.O. Box 123, Broadway 2007, Australia, and Chief Investigator, Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers.

S.1.2 Univariate Normal Distribution

The density function of the Univariate Normal distribution with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$ is

$$p(x) = (2\pi\sigma^2)^{-1/2} \exp\{-(x-\mu)^2/(2\sigma^2)\}, \quad x \in \mathbb{R}.$$

The sufficient statistic and base measure are

$$\boldsymbol{T}(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix} \quad \text{and} \quad h(x) = (2\pi)^{-1/2}.$$

The natural parameter vector and its inverse mapping are

$$\boldsymbol{\eta} = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} \mu/\sigma^2 \\ -1/(2\sigma^2) \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix} = \begin{bmatrix} -\eta_1/(2\eta_2) \\ -1/(2\eta_2) \end{bmatrix}$$

and the log-partition function is

$$A(\boldsymbol{\eta}) = -\tfrac{1}{4}\,(\eta_1^2/\eta_2) - \tfrac{1}{2}\log(-2\eta_2).$$

S.1.3 Inverse Chi-Squared and Inverse Gamma Distributions

The random variable x has an *Inverse Chi-Squared* distribution with shape parameter $\kappa > 0$ and scale parameter $\lambda > 0$, written $x \sim \text{Inverse-}\chi^2(\kappa, \lambda)$, if the density function of x is

$$p(x) = \{(\lambda/2)^{\kappa/2}/\Gamma(\kappa/2)\}\, x^{-(\kappa/2)-1} \exp\{-(\lambda/2)/x\}, \quad x > 0.$$

The random variable x has an *Inverse Gamma* distribution with shape parameter $\widetilde{\kappa} > 0$ and scale parameter $\widetilde{\lambda} > 0$, written $x \sim \text{Inverse-Gamma}(\widetilde{\kappa}, \widetilde{\lambda})$ if the density function of x is

$$p(x) = \{\widetilde{\lambda}^{\widetilde{\kappa}}/\Gamma(\widetilde{\kappa})\}\, x^{-\widetilde{\kappa}-1} \exp(-\widetilde{\lambda}/x), \quad x > 0.$$

The Inverse Chi-Squared and Inverse Gamma distributions are simple reparametrizations of each other in that

$$x \sim \text{Inverse-}\chi^2(\kappa, \lambda) \quad \text{if and only if} \quad x \sim \text{Inverse-Gamma}(\kappa/2, \lambda/2).$$

As explained in Section S.1.7, the Inverse Wishart distribution for random matrices reduces to the Inverse Chi-Squared distribution in the 1×1 case.

The sufficient statistic and base measure are

$$\boldsymbol{T}(x) = \begin{bmatrix} \log(x) \\ 1/x \end{bmatrix} \quad \text{and} \quad h(x) = I(x > 0).$$

The natural parameter vector and its inverse mappings are

$$\begin{aligned} \boldsymbol{\eta} = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} &= \begin{bmatrix} -\frac{1}{2}(\kappa+2) \\ -\frac{1}{2}\lambda \end{bmatrix} = \begin{bmatrix} -(\widetilde{\kappa}+1) \\ -\widetilde{\lambda} \end{bmatrix}, \\ \begin{bmatrix} \kappa \\ \lambda \end{bmatrix} = \begin{bmatrix} -2-2\eta_1 \\ -2\eta_2 \end{bmatrix} \quad &\text{and} \quad \begin{bmatrix} \widetilde{\kappa} \\ \widetilde{\lambda} \end{bmatrix} = \begin{bmatrix} -1-\eta_1 \\ -\eta_2 \end{bmatrix} \end{aligned} \tag{S.3}$$

and the log-partition function is

$$A(\boldsymbol{\eta}) = (\eta_1 + 1)\log(-\eta_2) + \log\Gamma(-\eta_1 - 1).$$

S.1.4 Beta Distribution

The density function of the Beta distribution with shape parameters $\alpha > 0$ and $\beta > 0$ is

$$p(x) = \frac{\Gamma(\alpha+\beta)\, x^{\alpha-1}(1-x)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)}, \quad 0 < x < 1.$$

The sufficient statistic and base measure are

$$\boldsymbol{T}(x) = \begin{bmatrix} \log(x) \\ \log(1-x) \end{bmatrix} \quad \text{and} \quad h(x) = I(0 < x < 1).$$

The natural parameter vector and its inverse mapping are

$$\boldsymbol{\eta} = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} \alpha - 1 \\ \beta - 1 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} \eta_1 + 1 \\ \eta_2 + 1 \end{bmatrix}$$

and the log-partition function is

$$A(\boldsymbol{\eta}) = \log\Gamma(\eta_1 + 1) + \log\Gamma(\eta_2 + 1) - \log\Gamma(\eta_1 + \eta_2 + 2).$$

S.1.5 Inverse Gaussian Distribution

The random variable x has an Inverse Gaussian distribution with parameters $\mu > 0$ and $\lambda > 0$, written $x \sim$ Inverse-Gaussian(μ, λ), if the density function of x is

$$p(x) = \lambda^{1/2}(2\pi\, x^3)^{-1/2} \exp\left\{-\frac{\lambda(x-\mu)^2}{2\mu^2\, x}\right\}, \quad x > 0.$$

The sufficient statistic and base measure are

$$\boldsymbol{T}(x) = \begin{bmatrix} x \\ 1/x \end{bmatrix} \quad \text{and} \quad h(x) = (2\pi\, x^3)^{-1/2}\, I(x > 0).$$

The natural parameter vector and its inverse mapping are

$$\boldsymbol{\eta} = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} -\lambda/(2\mu^2) \\ -\lambda/2 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \mu \\ \lambda \end{bmatrix} = \begin{bmatrix} (\eta_2/\eta_1)^{1/2} \\ -2\eta_2 \end{bmatrix}$$

and the log-partition function is

$$A(\boldsymbol{\eta}) = -2(\eta_1\eta_2)^{1/2} - \tfrac{1}{2}\log(-2\eta_2).$$

The Inverse Gaussian distribution is the only exponential family distribution in Section S.1 with a non-constant base measure. This implies that the entropy contribution from h, $E\{-\log h(x)\}$ where $x \sim$ Inverse-Gaussian(μ, λ), is not trivial and so we list it here. Using, for example, Lemma 1 of Gopal *et al.* (2012) we obtain

$$E\{-\log h(x)\} = \tfrac{1}{4}\log(4\pi^2\eta_2^3/\eta_1^3) + \frac{3}{2}\exp\Big(4(\eta_1\eta_2)^{1/2}\Big)\mathrm{Ei}\Big(-4(\eta_1\eta_2)^{1/2}\Big).$$

where the function Ei is defined in (S.2).

S.1.6 Multivariate Normal Distribution

The $d \times 1$ random vector $\boldsymbol{x}$ has a *Multivariate Normal* distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, a symmetric positive definite $d \times d$ matrix, written $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, if the density function of $\boldsymbol{x}$ is

$$p(\boldsymbol{x}) = (2\pi)^{-d/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\{-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\}, \quad \boldsymbol{x} \in \mathbb{R}^d.$$

The sufficient statistic and base measure are

$$\boldsymbol{T}(\boldsymbol{x}) = \begin{bmatrix} \boldsymbol{x} \\ \text{vec}(\boldsymbol{x}\boldsymbol{x}^T) \end{bmatrix} \quad \text{and} \quad h(\boldsymbol{x}) = (2\pi)^{-d/2}.$$

The natural parameter vector and inverse mapping are

$$\boldsymbol{\eta} = \begin{bmatrix} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\ -\frac{1}{2}\text{vec}(\boldsymbol{\Sigma}^{-1}) \end{bmatrix} \quad \text{and} \quad \begin{cases} \boldsymbol{\mu} = -\frac{1}{2}\{\text{vec}^{-1}(\boldsymbol{\eta}_2)\}^{-1}\boldsymbol{\eta}_1 \\ \boldsymbol{\Sigma} = -\frac{1}{2}\{\text{vec}^{-1}(\boldsymbol{\eta}_2)\}^{-1} \end{cases} \tag{S.4}$$

and the log-partition function is

$$A(\boldsymbol{\eta}) = -\tfrac{1}{4}\boldsymbol{\eta}_1^T\{\text{vec}^{-1}(\boldsymbol{\eta}_2)\}^{-1}\boldsymbol{\eta}_1 - \tfrac{1}{2}\log\Big|-2\text{vec}^{-1}(\boldsymbol{\eta}_2)\Big|.$$

S.1.7 Inverse Wishart Distribution

The $d \times d$ random matrix $\boldsymbol{X}$ has an *Inverse Wishart* distribution with shape parameter $\kappa > d-1$ and scale matrix $\boldsymbol{\Lambda}$, a symmetric positive definite $d \times d$ matrix, written $\boldsymbol{X} \sim$ Inverse-Wishart$(\kappa, \boldsymbol{\Lambda})$, if the density function of $\boldsymbol{X}$ is

$$\begin{aligned} p(\boldsymbol{X}) \;=\; & \frac{|\boldsymbol{\Lambda}|^{\kappa/2}}{2^{d\kappa/2}\pi^{d(d-1)/4}\prod_{j=1}^{d}\Gamma(\frac{\kappa+1-j}{2})}\,|\boldsymbol{X}|^{-(\kappa+d+1)/2}\exp\{-\tfrac{1}{2}\text{tr}(\boldsymbol{\Lambda}\boldsymbol{X}^{-1})\} \\ & \times I(\boldsymbol{X} \text{ a symmetric and positive definite } d \times d \text{ matrix}). \end{aligned}$$

The special case of $d = 1$ coincides with the Inverse Chi-Squared distribution. The sufficient statistic and base measure are

$$\boldsymbol{T}(\boldsymbol{X}) = \begin{bmatrix} \log|\boldsymbol{X}| \\ \text{vec}(\boldsymbol{X}^{-1}) \end{bmatrix} \quad \text{and} \quad h(\boldsymbol{X}) = \frac{I(\boldsymbol{X} \text{ is symmetric and positive definite})}{\pi^{d(d-1)/4}}. \tag{S.5}$$

The natural parameter vector and inverse mapping are

$$\boldsymbol{\eta} = \begin{bmatrix} \eta_1 \\ \boldsymbol{\eta}_2 \end{bmatrix} = \begin{bmatrix} -\frac{1}{2}(\kappa+d+1) \\ -\frac{1}{2}\text{vec}(\boldsymbol{\Lambda}) \end{bmatrix} \quad \text{and} \quad \begin{cases} \kappa = -d-1-2\eta_1 \\ \boldsymbol{\Lambda} = -2\text{vec}^{-1}(\boldsymbol{\eta}_2) \end{cases} \tag{S.6}$$

and the log-partition function is

$$A(\boldsymbol{\eta}) = \{\eta_1 + \tfrac{1}{2}(d+1)\}\log\big|-\text{vec}^{-1}(\boldsymbol{\eta}_2)\big| + \sum_{j=1}^{d}\log\Gamma\{-\eta_1 - \tfrac{1}{2}(d+j)\}.$$

S.1.7.1 Inverse G-Wishart Extension

Now consider the extension of the Inverse Wishart distribution corresponding to the inverse of the $d \times d$ random matrix $\boldsymbol{X}$ having some off-diagonal entries forced to equal zero. Such structure can be represented using undirected graphs and, following the nomenclature of Atay-Kayis & Massam (2005), is referred to as the *Inverse G-Wishart* distribution.

Let G be an undirected graph with d nodes labeled $1, \ldots, d$ and set E consisting of sets of pairs of nodes that are connected by an edge. We say that the $d \times d$ matrix $\boldsymbol{M}$ *respects* G if

$$\boldsymbol{M}_{ij} = 0 \quad \text{for all} \quad \{i, j\} \notin E.$$

Then the $d \times d$ random matrix $\boldsymbol{X}$ has an *Inverse G-Wishart* distribution with d-node undirected graph G, shape parameter $\kappa > d - 1$ and scale matrix $\boldsymbol{\Lambda}$, a symmetric positive definite $d \times d$ matrix that respects G, written $\boldsymbol{X} \sim$ Inverse-G-Wishart$(G, \kappa, \boldsymbol{\Lambda})$, if the density function of $\boldsymbol{X}$ is

$$\begin{aligned} p(\boldsymbol{X}) \;\propto\; & |\boldsymbol{X}|^{-(\kappa+d+1)/2} \exp\{-\tfrac{1}{2}\text{tr}(\boldsymbol{\Lambda}\boldsymbol{X}^{-1})\}\, I(\boldsymbol{X} \text{is symmetric and positive definite}) \\ & \times I(\boldsymbol{X}^{-1} \text{ respects } G). \end{aligned}$$

The normalizing factor follows from the formulae of Uhler *et al.* (2014), although it is quite complicated for general G.

The sufficient statistic $\boldsymbol{T}(\boldsymbol{X})$ and natural parameter vector take the same form as for the ordinary Inverse Wishart distribution, given at (S.5) and (S.6).

The special case of diagonal matrices coincides with G such that $E = \emptyset$, meaning that G is a totally disconnected graph. We denote such G by G_{diag}. Note that

$$\boldsymbol{X} \sim \text{Inverse-G-Wishart}(G_{\text{diag}}, \kappa, \boldsymbol{\Lambda})$$

if and only if

$$p(\boldsymbol{X}) \;=\; \frac{1}{2^{d(\kappa+d-1)/2}\Gamma(\frac{\kappa+d-1}{2})^d} \left\{ \prod_{i=1}^{d} \boldsymbol{\Lambda}_{ii}^{(\kappa+d-1)/2} \boldsymbol{X}_{ii}^{-(\kappa/2)-1} \exp(-\tfrac{1}{2}\boldsymbol{\Lambda}_{ii}/\boldsymbol{X}_{ii}) I(\boldsymbol{X}_{ii} > 0) \right\}$$

and is simply a product of Inverse Chi-Squared density functions.

S.1.8 Table of Sufficient Statistic Expectations

Table S.1 lists the sufficient statistic expectations for each of the exponential family distributions covered in Section S.1. All expressions are in terms of natural parameters.

Distribution	$T(x), \boldsymbol{T}(\boldsymbol{x}), \boldsymbol{T}(\boldsymbol{X})$	$E\{T(x)\}, E\{\boldsymbol{T}(\boldsymbol{x})\}, E\{\boldsymbol{T}(\boldsymbol{X})\}$
Bernoulli	x	$1/(1+e^{-\eta})$
Univariate Normal	$\begin{bmatrix} x \\ x^2 \end{bmatrix}$	$\begin{bmatrix} -\eta_1/(2\eta_2) \\ (\eta_1^2 - 2\eta_2)/(4\eta_2^2) \end{bmatrix}$
Inverse Chi-Squared	$\begin{bmatrix} \log(x) \\ 1/x \end{bmatrix}$	$\begin{bmatrix} \log(-\eta_2) - \psi(-\eta_1 - 1) \\ (\eta_1 + 1)/\eta_2 \end{bmatrix}$
Beta	$\begin{bmatrix} \log(x) \\ \log(1-x) \end{bmatrix}$	$\begin{bmatrix} \psi(\eta_1 + 1) - \psi(\eta_1 + \eta_2 + 2) \\ \psi(\eta_2 + 1) - \psi(\eta_1 + \eta_2 + 2) \end{bmatrix}$
Inverse Gaussian	$\begin{bmatrix} x \\ 1/x \end{bmatrix}$	$\begin{bmatrix} (\eta_2/\eta_1)^{1/2} \\ (\eta_1/\eta_2)^{1/2} - 1/(2\eta_2) \end{bmatrix}$
Multivariate Normal	$\begin{bmatrix} \boldsymbol{x} \\ \text{vec}(\boldsymbol{x}\boldsymbol{x}^T) \end{bmatrix}$	$\begin{bmatrix} -\frac{1}{2}\{\text{vec}^{-1}(\boldsymbol{\eta}_2)\}^{-1}\boldsymbol{\eta}_1 \\ \frac{1}{4}\text{vec}\Big(\{\text{vec}^{-1}(\boldsymbol{\eta}_2)\}^{-1} \times[\boldsymbol{\eta}_1\boldsymbol{\eta}_1^T\{\text{vec}^{-1}(\boldsymbol{\eta}_2)\}^{-1} - 2\,\boldsymbol{I}]\Big) \end{bmatrix}$
Inverse Wishart	$\begin{bmatrix} \log\lvert\boldsymbol{X}\rvert \\ \text{vec}(\boldsymbol{X}^{-1}) \end{bmatrix}$	$\begin{bmatrix} \log\lvert-\text{vec}^{-1}(\boldsymbol{\eta}_2)\rvert \\ -\sum_{j=1}^{d}\psi\{-\eta_1 - \frac{1}{2}(d+j)\} \\ \{\eta_1 + \frac{1}{2}(d+1)\}\text{vec}[\{\text{vec}^{-1}(\boldsymbol{\eta}_2)\}^{-1}] \end{bmatrix}$

Table S.1: *Expressions for sufficient statistics and their expectations in terms of natural parameters for some common exponential family distributions.*

S.2 Derivational Details

Here we provide details on various derivations appearing throughout the article.

S.2.1 Derivation of Message Functional Forms Given by (19)

With 'const' denoting terms that do not depend on the function argument, the logarithms of each of the factors can be expressed as follows:

$$\log\, p(\boldsymbol{\beta}) \;=\; \begin{bmatrix} \boldsymbol{\beta} \\ \text{vec}(\boldsymbol{\beta}\boldsymbol{\beta}^T) \end{bmatrix}^T \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}} \\ -\frac{1}{2}\text{vec}(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}) \end{bmatrix} + \text{const},$$

$$\log\, p(\boldsymbol{y}\,|\,\boldsymbol{\beta},\sigma^2) \;=\; \begin{cases} \begin{bmatrix} \boldsymbol{\beta} \\ \text{vec}(\boldsymbol{\beta}\boldsymbol{\beta}^T) \end{bmatrix}^T \begin{bmatrix} \boldsymbol{X}^T\boldsymbol{y} \\ -\frac{1}{2}\text{vec}(\boldsymbol{X}^T\boldsymbol{X}) \end{bmatrix} \left(\frac{1}{\sigma^2}\right) + \text{const, as a function of } \boldsymbol{\beta}, \\ \begin{bmatrix} \log(\sigma^2) \\ 1/\sigma^2 \end{bmatrix}^T \begin{bmatrix} -\frac{1}{2}\,n \\ -\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|^2 \end{bmatrix} + \text{const, as a function of } \sigma^2, \end{cases}$$

$$\log\, p(\sigma^2\,|\,a) \;=\; \begin{cases} \begin{bmatrix} \log(\sigma^2) \\ 1/\sigma^2 \end{bmatrix}^T \begin{bmatrix} -\frac{3}{2} \\ -1/(2a) \end{bmatrix} + \text{const, as a function of } \sigma^2, \\ \begin{bmatrix} \log(a) \\ 1/a \end{bmatrix}^T \begin{bmatrix} -\frac{1}{2} \\ -1/(2\sigma^2) \end{bmatrix} + \text{const, as a function of } a, \text{ and} \end{cases}$$

$$\log\, p(a) \;=\; \begin{bmatrix} \log(a) \\ 1/a \end{bmatrix}^T \begin{bmatrix} -\frac{3}{2} \\ -1/(2\,A^2) \end{bmatrix} + \text{const}.$$

Then, since the only neighbor of $p(\boldsymbol{\beta})$ in Figure 3 is $\boldsymbol{\beta}$, the expectation in (8) disappears and we immediately get

$$m_{p(\boldsymbol{\beta})\,\to\,\boldsymbol{\beta}}(\boldsymbol{\beta}) \longleftarrow \exp\left\{ \begin{bmatrix} \boldsymbol{\beta} \\ \text{vec}(\boldsymbol{\beta}\boldsymbol{\beta}^T) \end{bmatrix}^T \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}} \\ -\frac{1}{2}\text{vec}(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}) \end{bmatrix} \right\}$$

which confirms the first part of (19). The factor $p(\boldsymbol{y}\,|\,\boldsymbol{\beta},\sigma^2)$ has both stochastic nodes $\boldsymbol{\beta}$ and σ^2 as neighbors so

$$m_{p(\boldsymbol{y}|\boldsymbol{\beta},\,\sigma^2)\,\to\,\boldsymbol{\beta}}(\boldsymbol{\beta}) \longleftarrow \exp\left\{ \begin{bmatrix} \boldsymbol{\beta} \\ \text{vec}(\boldsymbol{\beta}\boldsymbol{\beta}^T) \end{bmatrix}^T \begin{bmatrix} \boldsymbol{X}^T\boldsymbol{y} \\ -\frac{1}{2}\text{vec}(\boldsymbol{X}^T\boldsymbol{X}) \end{bmatrix} E_{p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)\to\boldsymbol{\beta}}\left(\frac{1}{\sigma^2}\right) \right\}$$

and

$$m_{p(\boldsymbol{y}|\boldsymbol{\beta},\,\sigma^2)\,\to\,\sigma^2}(\sigma^2) \longleftarrow \exp\left\{ \begin{bmatrix} \log(\sigma^2) \\ 1/\sigma^2 \end{bmatrix}^T \begin{bmatrix} -\frac{1}{2}\,n \\ -\frac{1}{2}E_{p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)\to\sigma^2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|^2 \end{bmatrix} \right\}$$

which are also of the forms given in (19). Similar arguments show that $m_{p(\sigma^2|\,a)\,\to\,\sigma^2}(\sigma^2)$, $m_{p(\sigma^2|a)\,\to\,a}(a)$ and $m_{p(a)\,\to\,a}(a)$ have the stated Inverse-χ^2 forms after the first iteration of VMP.

S.2.2 Derivation of the Jaakkola-Jordan Updates

According to (8) and (9), the message passed from the factor $p(\boldsymbol{y}|\boldsymbol{\theta})$, given at (42),is

$$m_{p(\boldsymbol{y}|\boldsymbol{\theta})\,\to\,\boldsymbol{\theta}}(\boldsymbol{\theta}) = p(\boldsymbol{y}|\,\boldsymbol{\theta}) = \exp\Big[\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{\theta} - \mathbf{1}^T\log\{\mathbf{1}+\exp(\boldsymbol{A}\boldsymbol{\theta})\}\Big]. \tag{S.7}$$

This, however, is not conjugate with the Multivariate Normal messages typically passed to $\boldsymbol{\theta}$ from other neighboring factors. The Jaakkola-Jordan device (Jaakkola & Jordan, 2000) is based on the following variational representation of the troublesome function in (S.7):

$$-\log(1+e^x) = \max_{\xi\in\mathbb{R}}\left\{A(\xi)x^2 - \tfrac{1}{2}\,x + C(\xi)\right\} \text{ for all } x\in\mathbb{R} \tag{S.8}$$

where $A(\xi) \equiv -\tanh(\xi/2)/(4\xi)$ and $C(\xi) \equiv \xi/2 - \log(1+e^{\xi}) + \xi\tanh(\xi/2)/4$. Representation (S.8) leads to the following family of variational lower bounds on the logarithm of (42):

$$\log\, p(\boldsymbol{y}|\,\boldsymbol{\theta}) = -\boldsymbol{\theta}^T\boldsymbol{A}^T\text{diag}\left\{\frac{\tanh(\boldsymbol{\xi}/2)}{4\boldsymbol{\xi}}\right\}\boldsymbol{A}\,\boldsymbol{\theta} + (\boldsymbol{y} - \tfrac{1}{2}\mathbf{1})^T\boldsymbol{A}\boldsymbol{\theta} + \mathbf{1}^T C(\boldsymbol{\xi})$$

and corresponding family of conjugate messages

$$\underline{m}_{p(\boldsymbol{y}|\boldsymbol{\theta})\,\rightarrow\,\boldsymbol{\theta}}(\boldsymbol{\theta};\boldsymbol{\xi}) \equiv \exp\left\{\begin{bmatrix}\boldsymbol{\theta}\\ \text{vec}(\boldsymbol{\theta}\boldsymbol{\theta}^T)\end{bmatrix}^T\begin{bmatrix}\boldsymbol{A}^T(\boldsymbol{y}-\frac{1}{2}\,\mathbf{1})\\ -\text{vec}\left(\boldsymbol{A}^T\text{diag}\left\{\frac{\tanh(\boldsymbol{\xi}/2)}{4\boldsymbol{\xi}}\right\}\boldsymbol{A}\right)\end{bmatrix}^T\right\}$$

where $\boldsymbol{\xi}$ is an $n \times 1$ vector of variational parameters.

The updates (43) are driven by the goal of maximizing the following $\boldsymbol{\theta}$-*localized* approximate marginal log-likelihood:

$$\begin{aligned}\log \underline{p}(\boldsymbol{y};q)^{[\boldsymbol{\theta}]} = &\,\text{Entropy}\{q(\boldsymbol{\theta})\} + E_q\{\log\, p(\boldsymbol{y}|\boldsymbol{\theta})\}\\ &+E_q(\text{sum of other log-factors neighboring }\boldsymbol{\theta}).\end{aligned}$$

Rohde & Wand (2016) contains further details on localized approximate marginal log-likelihoods. Application of the Jaakkola-Jordan device leads to the family of approximate marginal log-likelihoods:

$$\begin{aligned}\log \underline{p}(\boldsymbol{y};q,\boldsymbol{\xi})^{[\boldsymbol{\theta}]} = &\,\text{Entropy}\{q(\boldsymbol{\theta};\boldsymbol{\xi})\} - E_{q(\boldsymbol{\theta};\boldsymbol{\xi})}\left[\boldsymbol{\theta}^T\boldsymbol{A}^T\text{diag}\left\{\frac{\tanh(\boldsymbol{\xi}/2)}{4\boldsymbol{\xi}}\right\}\boldsymbol{A}\,\boldsymbol{\theta}\right]\\ &+E_{q(\boldsymbol{\theta};\boldsymbol{\xi})}\{(\boldsymbol{y}-\tfrac{1}{2}\mathbf{1})^T\boldsymbol{A}\boldsymbol{\theta}\} + \mathbf{1}^T C(\boldsymbol{\xi})\\ &+E_{q(\boldsymbol{\theta};\boldsymbol{\xi})}\big(\text{sum of other log-factors neighboring }\boldsymbol{\theta}\big).\end{aligned} \tag{S.9}$$

Courtesy of (10), the current $q(\boldsymbol{\theta};\boldsymbol{\xi})$ density function satisfies

$$q(\boldsymbol{\theta};\boldsymbol{\xi}) \propto \underline{m}_{p(\boldsymbol{y}|\boldsymbol{\theta})\,\rightarrow\,\boldsymbol{\theta}}(\boldsymbol{\theta};\boldsymbol{\xi}) \times (\text{product of messages passed to }\boldsymbol{\theta}\text{ from its other neighbors}). \tag{S.10}$$

Note that update (7) allows us to replace (S.10) by

$$q(\boldsymbol{\theta};\boldsymbol{\xi}) \propto \underline{m}_{p(\boldsymbol{y}|\boldsymbol{\theta})\,\rightarrow\,\boldsymbol{\theta}}(\boldsymbol{\theta};\boldsymbol{\xi})\, m_{\boldsymbol{\theta}\,\rightarrow\,p(\boldsymbol{y}|\,\boldsymbol{\theta})}(\boldsymbol{\theta}). \tag{S.11}$$

As explained in, for example, Section 21.8 of Murphy (2012), a practical approach to optimizing the $\boldsymbol{\xi}$ vector is coordinate ascent applied to (S.9) but with the $\boldsymbol{\xi}$ appearing in $q(\boldsymbol{\theta};\boldsymbol{\xi})$ held fixed. This approach also has an Expectation-Maximization algorithm representation (Jaakkola & Jordan, 2000). Under this strategy

$$\begin{aligned}\log \underline{p}(\boldsymbol{y};q,\boldsymbol{\xi})^{[\boldsymbol{\theta}]} = &\,\mathbf{1}^T C(\boldsymbol{\xi}) - E_{q(\boldsymbol{\theta};\boldsymbol{\xi})}\left[\boldsymbol{\theta}^T\boldsymbol{A}^T\text{diag}\left\{\frac{\tanh(\boldsymbol{\xi}/2)}{4\boldsymbol{\xi}}\right\}\boldsymbol{A}\,\boldsymbol{\theta}\right]\\ &+\text{terms not involving }\boldsymbol{\xi}\text{, excluding } q(\boldsymbol{\theta};\boldsymbol{\xi}).\end{aligned} \tag{S.12}$$

The first line of (S.12) is maximized over $\boldsymbol{\xi}$ by

$$\boldsymbol{\xi} = \sqrt{\text{diagonal}\big\{\boldsymbol{A}\,E_{q(\boldsymbol{\theta};\boldsymbol{\xi})}(\boldsymbol{\theta}\boldsymbol{\theta}^T)\boldsymbol{A}^T\big\}}.$$

(e.g. Murphy, 2012, Section 21.8.3). From (S.11),

$$q(\boldsymbol{\theta};\boldsymbol{\xi}) \propto \exp\left\{\begin{bmatrix}\boldsymbol{\theta}\\ \text{vec}(\boldsymbol{\theta}\boldsymbol{\theta}^T)\end{bmatrix}^T\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta})\,\rightarrow\,\boldsymbol{\theta}} + \boldsymbol{\eta}_{\boldsymbol{\theta}\,\rightarrow\,p(\boldsymbol{y}|\,\boldsymbol{\theta})}\right)\right\}$$

and, so from Table S.1,

$$\begin{aligned}E_{q(\boldsymbol{\theta};\boldsymbol{\xi})}(\boldsymbol{\theta}\boldsymbol{\theta}^T) = &\left\{\text{vec}^{-1}\Big(\big(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta})\,\leftrightarrow\,\boldsymbol{\theta}}\big)_2\Big)\right\}^{-1}\\ &\times\Big[\big(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta})\,\leftrightarrow\,\boldsymbol{\theta}}\big)_1\big(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta})\,\leftrightarrow\,\boldsymbol{\theta}}\big)_1^T\left\{\text{vec}^{-1}\Big(\big(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta})\,\leftrightarrow\,\boldsymbol{\theta}}\big)_2\Big)\right\}^{-1} - 2\boldsymbol{I}\Big]\end{aligned}$$

and the updates (43) follow immediately.

S.2.3 Derivation of the Albert-Chib Updates

The relevant factor graph fragments are displayed in the right panel of Figure 8. The factors are

$$p(\boldsymbol{y}|\,\boldsymbol{a}) = \prod_{i=1}^{n} \{y_i\, I(a_i \geq 0) + (1-y_i)\, I(a_i < 0)\}$$

and

$$p(\boldsymbol{a}|\,\boldsymbol{\theta}) = (2\pi)^{-n/2} \exp\{-\tfrac{1}{2}\|\boldsymbol{a} - \boldsymbol{A}\boldsymbol{\theta}\|^2\}.$$

According to (8) and (9), the messages from $p(\boldsymbol{y}|\,\boldsymbol{a})$ to each a_i are

$$m_{p(\boldsymbol{y}|\,\boldsymbol{a}) \,\to\, a_i}(a_i) = y_i\, I(a_i \geq 0) + (1-y_i)\, I(a_i < 0), \quad 1 \leq i \leq n,$$

and the messages from $p(\boldsymbol{a}|\boldsymbol{\theta})$ to each a_i are

$$m_{p(\boldsymbol{a}|\,\boldsymbol{\theta}) \,\to\, a_i}(a_i) = \exp\left[-\tfrac{1}{2}\{a_i - \left(\boldsymbol{A}\, E_{p(\boldsymbol{a}|\boldsymbol{\theta})\to a_i}(\boldsymbol{\theta})\right)_i\}^2\right], \quad 1 \leq i \leq n$$

where, with the assistance of Table S.1,

$$E_{p(\boldsymbol{a}|\boldsymbol{\theta})\to a_i}(\boldsymbol{\theta}) = -\tfrac{1}{2}\Big\{\text{vec}^{-1}\Big(\big(\boldsymbol{\eta}_{p(\boldsymbol{a}|\,\boldsymbol{\theta}) \,\leftrightarrow\, \boldsymbol{\theta}}\big)_2\Big)\Big\}^{-1}\big(\boldsymbol{\eta}_{p(\boldsymbol{a}|\,\boldsymbol{\theta}) \,\leftrightarrow\, \boldsymbol{\theta}}\big)_1, \quad 1 \leq i \leq n.$$

Since, from (7), $m_{a_i \,\to\, p(\boldsymbol{a}|\,\boldsymbol{\theta})}(a_i) \longleftarrow m_{p(\boldsymbol{y}|\,\boldsymbol{a}) \,\to\, a_i}(a_i)$ we have

$$m_{p(\boldsymbol{a}|\,\boldsymbol{\theta}) \,\to\, a_i}(a_i) m_{a_i \,\to\, p(\boldsymbol{a}|\,\boldsymbol{\theta})}(a_i) \propto$$
$$\begin{cases} \text{the } N((\boldsymbol{A}\, E_{p(\boldsymbol{a}|\boldsymbol{\theta})\to a_i}(\boldsymbol{\theta}))_i, 1) \text{ density function truncated to } (-\infty, 0) \text{ if } y_i = 0 \\ \text{the } N((\boldsymbol{A}\, E_{p(\boldsymbol{a}|\boldsymbol{\theta})\to a_i}(\boldsymbol{\theta}))_i, 1) \text{ density function truncated to } [0, \infty) \text{ if } y_i = 1. \end{cases}$$

Standard manipulations then lead to the mean of the normalized

$$m_{p(\boldsymbol{a}|\,\boldsymbol{\theta}) \,\to\, a_i}(a_i) m_{a_i \,\to\, p(\boldsymbol{a}|\,\boldsymbol{\theta})}(a_i)$$

equaling

$$\mu_{p(\boldsymbol{a}|\,\boldsymbol{\theta}) \,\leftrightarrow\, a_i} \equiv \nu_i + (2\,y_i - 1)\zeta'\big((2\,y_i - 1)\nu_i\big) \tag{S.13}$$

where

$$\nu_i \equiv (\boldsymbol{A}\, E_{p(\boldsymbol{a}|\boldsymbol{\theta})\to a_i}(\boldsymbol{\theta}))_i = -\tfrac{1}{2}\Big(\boldsymbol{A}\Big\{\text{vec}^{-1}\Big(\big(\boldsymbol{\eta}_{p(\boldsymbol{a}|\,\boldsymbol{\theta}) \,\leftrightarrow\, \boldsymbol{\theta}}\big)_2\Big)\Big\}^{-1}\big(\boldsymbol{\eta}_{p(\boldsymbol{a}|\,\boldsymbol{\theta}) \,\leftrightarrow\, \boldsymbol{\theta}}\big)_1\Big)_i. \tag{S.14}$$

Lastly, the message from $p(\boldsymbol{a}|\boldsymbol{\theta})$ to $\boldsymbol{\theta}$ is

$$m_{p(\boldsymbol{a}|\,\boldsymbol{\theta}) \,\to\, \boldsymbol{\theta}} = \exp\left\{\begin{bmatrix} \boldsymbol{\theta} \\ \text{vec}(\boldsymbol{\theta}\boldsymbol{\theta}^T) \end{bmatrix}^T \boldsymbol{\eta}_{p(\boldsymbol{a}|\,\boldsymbol{\theta}) \,\to\, \boldsymbol{\theta}}\right\}$$

where

$$\boldsymbol{\eta}_{p(\boldsymbol{a}|\,\boldsymbol{\theta}) \,\to\, \boldsymbol{\theta}} \longleftarrow \begin{bmatrix} \boldsymbol{A}^T \mu_{p(\boldsymbol{a}|\,\boldsymbol{\theta}) \,\leftrightarrow\, \boldsymbol{a}} \\ -\tfrac{1}{2}\text{vec}(\boldsymbol{A}^T\boldsymbol{A}) \end{bmatrix} \tag{S.15}$$

and $\mu_{p(\boldsymbol{a}|\,\boldsymbol{\theta}) \,\leftrightarrow\, \boldsymbol{a}}$ is the $n \times 1$ vector containing the $\mu_{p(\boldsymbol{a}|\,\boldsymbol{\theta}) \,\leftrightarrow\, a_i}$. The updates in (46) arise from substitution of (S.13) and (S.14) into (S.15).

S.2.4 Derivation of the Knowles-Minka-Wand Updates

The message passed from $p(\boldsymbol{y}|\boldsymbol{\theta})$ to $\boldsymbol{\theta}$ is

$$m_{p(\boldsymbol{y}|\boldsymbol{\theta}) \to \boldsymbol{\theta}}(\boldsymbol{\theta}) = \exp\{\boldsymbol{y}^T \boldsymbol{A}\boldsymbol{\theta} - \mathbf{1}^T \exp(\boldsymbol{X}\boldsymbol{\beta})\}$$

is not conjugate with Multivariate Normal messages passed to $\boldsymbol{\theta}$ from other factors. A remedy proposed by Knowles & Minka (2011) and dubbed *non-conjugate VMP* involves, in this case, replacement of $m_{p(\boldsymbol{y}|\boldsymbol{\theta}) \to \boldsymbol{\theta}}(\boldsymbol{\theta})$ by

$$\widetilde{m}_{p(\boldsymbol{y}|\,\boldsymbol{\theta}) \to \boldsymbol{\theta}}(\boldsymbol{\theta}) \equiv \exp\left\{ \begin{bmatrix} \boldsymbol{\theta} \\ \text{vec}(\boldsymbol{\theta}\boldsymbol{\theta}^T) \end{bmatrix}^T \boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta}) \to \boldsymbol{\theta}} \right\}$$

to enforce conjugacy with Multivariate Normal messages.

Knowles & Minka (2011) propose that $\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta}) \to \boldsymbol{\theta}}$ be updated according to maximization of a localized Kullback-Leibler divergence criterion, summarized in their Algorithm 1. For the Poisson regression likelihood this criterion can be expressed in closed form. However, expressions in Algorithm 1 of Knowles & Minka (2011) involve inversion of a matrix that is quartic in the length of $\boldsymbol{\theta}$. Wand (2014) derived fully simplified updates for non-conjugate VMP in the special case of Multivariate Normal message approximation.

As with Section S.2.2, the derivation starts with the $\boldsymbol{\theta}$-localized approximate marginal log-likelihood for the Poisson likelihood fragment:

$$\begin{aligned} \log \underline{p}(\boldsymbol{y}; q)^{[\boldsymbol{\theta}]} = \text{Entropy}\{q(\boldsymbol{\theta})\} + E_q\{\log\, p(\boldsymbol{y}|\boldsymbol{\theta})\} \\ + E_q(\text{sum of other log-factors neighboring } \boldsymbol{\theta}) \end{aligned} \tag{S.16}$$

but now the logarithm of the likelihood factor is

$$\log\, p(\boldsymbol{y}|\boldsymbol{\theta}) = \boldsymbol{y}^T \boldsymbol{A}\boldsymbol{\theta} - \mathbf{1}^T \exp(\boldsymbol{A}\boldsymbol{\theta}) - \mathbf{1}^T \log(\boldsymbol{y}!).$$

Using the same argument that led to (S.11), the current $q(\boldsymbol{\theta})$ density function is the Multivariate Normal density function with natural parameter vector $\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta}) \leftrightarrow \boldsymbol{\theta}}$. Let $\boldsymbol{\mu}_{p(\boldsymbol{y}|\,\boldsymbol{\theta}) \leftrightarrow \boldsymbol{\theta}}$ and $\Sigma_{p(\boldsymbol{y}|\,\boldsymbol{\theta}) \leftrightarrow \boldsymbol{\theta}}$ be the corresponding common parameters. Then, because of (S.4), the natural parameters and common parameters are the following functions of one another:

$$\begin{aligned} &\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta}) \leftrightarrow \boldsymbol{\theta}} = \begin{bmatrix} (\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta}) \leftrightarrow \boldsymbol{\theta}})_1 \\ (\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta}) \leftrightarrow \boldsymbol{\theta}})_2 \end{bmatrix} = \begin{bmatrix} \Sigma^{-1}_{p(\boldsymbol{y}|\,\boldsymbol{\theta}) \leftrightarrow \boldsymbol{\theta}} \boldsymbol{\mu}_{p(\boldsymbol{y}|\,\boldsymbol{\theta}) \leftrightarrow \boldsymbol{\theta}} \\ -\frac{1}{2}\text{vec}(\Sigma^{-1}_{p(\boldsymbol{y}|\,\boldsymbol{\theta}) \leftrightarrow \boldsymbol{\theta}}) \end{bmatrix} \\ &\text{and} \quad \begin{cases} \boldsymbol{\mu}_{p(\boldsymbol{y}|\,\boldsymbol{\theta}) \leftrightarrow \boldsymbol{\theta}} = -\frac{1}{2}\{\text{vec}^{-1}\big((\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta}) \leftrightarrow \boldsymbol{\theta}})_2\big)\}^{-1}(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta}) \leftrightarrow \boldsymbol{\theta}})_1 \\ \Sigma_{p(\boldsymbol{y}|\,\boldsymbol{\theta}) \leftrightarrow \boldsymbol{\theta}} = -\frac{1}{2}\{\text{vec}^{-1}\big((\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta}) \leftrightarrow \boldsymbol{\theta}})_2\big)\}^{-1}. \end{cases} \end{aligned} \tag{S.17}$$

For the next part of the derivation with work with the common parameters to make use of a key result in Wand (2014) (see also Rohde & Wand, 2016), and then transform to natural parameter vectors after that. We also adopt the following shorthand:

$$\boldsymbol{\mu}_{q(\boldsymbol{\theta})} \equiv \boldsymbol{\mu}_{p(\boldsymbol{y}|\,\boldsymbol{\theta}) \leftrightarrow \boldsymbol{\theta}} \quad \text{and} \quad \boldsymbol{\Sigma}_{q(\boldsymbol{\theta})} \equiv \Sigma_{p(\boldsymbol{y}|\,\boldsymbol{\theta}) \leftrightarrow \boldsymbol{\theta}}.$$

Under conjugacy, the non-entropy component of (S.16) is

$$\begin{aligned} \text{NonEntropy}(q; \boldsymbol{\mu}_{q(\boldsymbol{\theta})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}) = E_q\{\boldsymbol{y}^T \boldsymbol{A}\boldsymbol{\theta} - \mathbf{1}^T \exp(\boldsymbol{A}\boldsymbol{\theta})\} - \mathbf{1}^T \log(\boldsymbol{y}!) \\ + E_q\left\{ \begin{bmatrix} \boldsymbol{\theta} \\ \text{vec}(\boldsymbol{\theta}\boldsymbol{\theta}^T) \end{bmatrix} \right\}^T \boldsymbol{\eta}^\dagger \end{aligned}$$

where $\boldsymbol{\eta}^{\dagger}$ is the sum of the natural parameters of messages passed to $\boldsymbol{\theta}$ other than the message from $p(\boldsymbol{y}|\boldsymbol{\theta})$. But, because of (7),

$$\boldsymbol{\eta}_{\boldsymbol{\theta} \to p(\boldsymbol{y}|\,\boldsymbol{\theta})} \longleftarrow \boldsymbol{\eta}^{\dagger}$$

and so we get the following explicit form depending only on the messages passed between the nodes of the Poisson likelihood fragment:

$$\begin{aligned}\text{NonEntropy}(q; \boldsymbol{\mu}_{q(\boldsymbol{\theta})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}) = \boldsymbol{y}^T \boldsymbol{A}\boldsymbol{\mu}_{q(\boldsymbol{\theta})} - \mathbf{1}^T \exp\{\boldsymbol{A}\boldsymbol{\mu}_{q(\boldsymbol{\theta})} + \tfrac{1}{2}\text{diagonal}(\boldsymbol{A}\boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}\boldsymbol{A}^T)\} \\ + \boldsymbol{\mu}_{q(\boldsymbol{\theta})}^T(\boldsymbol{\eta}_{\boldsymbol{\theta} \to p(\boldsymbol{y}|\,\boldsymbol{\theta})})_1 + \boldsymbol{\mu}_{q(\boldsymbol{\theta})}^T \text{vec}^{-1}\big((\boldsymbol{\eta}_{\boldsymbol{\theta} \to p(\boldsymbol{y}|\,\boldsymbol{\theta})})_2\big)\boldsymbol{\mu}_{q(\boldsymbol{\theta})} \\ + \text{tr}\Big\{\boldsymbol{\Sigma}_{q(\boldsymbol{\theta})} \text{vec}^{-1}\big((\boldsymbol{\eta}_{\boldsymbol{\theta} \to p(\boldsymbol{y}|\,\boldsymbol{\theta})})_2\big)\Big\} - \mathbf{1}^T \log(\boldsymbol{y}!).\end{aligned}$$

From equation (7) of Wand (2014) and Result 2 of Rohde & Wand (2016), fixed-point iteration with respect to the natural parameter vector for maximization of (S.16) reduces to

$$\left\{\begin{array}{rcl} \boldsymbol{v}_{q(\boldsymbol{\theta})} & \longleftarrow & \mathsf{D}_{\boldsymbol{\mu}_{q(\boldsymbol{\theta})}}\text{NonEntropy}(q; \boldsymbol{\mu}_{q(\boldsymbol{\theta})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\theta})})^T \\ \boldsymbol{\Sigma}_{q(\boldsymbol{\theta})} & \longleftarrow & -\{\mathsf{H}_{\boldsymbol{\mu}_{q(\boldsymbol{\theta})}}\text{NonEntropy}(q; \boldsymbol{\mu}_{q(\boldsymbol{\theta})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\theta})})\}^{-1} \\ \mu_{q(\boldsymbol{\theta})} & \longleftarrow & \boldsymbol{\mu}_{q(\boldsymbol{\theta})} + \boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}\boldsymbol{v}_{q(\boldsymbol{\theta})} \end{array}\right. \tag{S.18}$$

where $\mathsf{D}_{\boldsymbol{\mu}_{q(\boldsymbol{\theta})}}$ and $\mathsf{H}_{\boldsymbol{\mu}_{q(\boldsymbol{\theta})}}$ denote, respectively, the derivative vector and Hessian matrix with respect to $\boldsymbol{\mu}_{q(\boldsymbol{\theta})}$. Formal definitions are given in Wand (2014). Arguments analogous to those given in Appendix A.4 of Menictas & Wand (2015) lead to the explicit forms for the non-entropy component of (S.16):

$$\begin{aligned}\mathsf{D}_{\boldsymbol{\mu}_{q(\boldsymbol{\theta})}}\text{NonEntropy}(q; \boldsymbol{\mu}_{q(\boldsymbol{\theta})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\theta})})^T = \boldsymbol{A}^T(\boldsymbol{y} - \boldsymbol{\omega}) + (\boldsymbol{\eta}_{\boldsymbol{\theta} \to p(\boldsymbol{y}|\,\boldsymbol{\theta})})_1 \\ + 2\,\text{vec}^{-1}\big((\boldsymbol{\eta}_{\boldsymbol{\theta} \to p(\boldsymbol{y}|\,\boldsymbol{\theta})})_2\big)\boldsymbol{\mu}_{q(\boldsymbol{\theta})}\end{aligned}$$

and

$$\mathsf{H}_{\boldsymbol{\mu}_{q(\boldsymbol{\theta})}}\text{NonEntropy}(q; \boldsymbol{\mu}_{q(\boldsymbol{\theta})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}) = -\boldsymbol{A}^T\text{diag}(\boldsymbol{\omega})\boldsymbol{A} + 2\,\text{vec}^{-1}\big((\boldsymbol{\eta}_{\boldsymbol{\theta} \to p(\boldsymbol{y}|\,\boldsymbol{\theta})})_2\big)$$

where

$$\boldsymbol{\omega} \equiv \exp\{\boldsymbol{A}\boldsymbol{\mu}_{q(\boldsymbol{\theta})} + \tfrac{1}{2}\text{diagonal}(\boldsymbol{A}\boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}\boldsymbol{A}^T)\}.$$

Substitution into (S.18) then gives the updating scheme

$$\left\{\begin{array}{rcl} \boldsymbol{\omega} & \longleftarrow & \exp\{\boldsymbol{A}\boldsymbol{\mu}_{q(\boldsymbol{\theta})} + \tfrac{1}{2}\text{diagonal}(\boldsymbol{A}\boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}\boldsymbol{A}^T)\} \\ \boldsymbol{\Sigma}_{q(\boldsymbol{\theta})} & \longleftarrow & \Big\{\boldsymbol{A}^T\text{diag}(\boldsymbol{\omega})\boldsymbol{A} - 2\,\text{vec}^{-1}\big((\boldsymbol{\eta}_{\boldsymbol{\theta} \to p(\boldsymbol{y}|\,\boldsymbol{\theta})})_2\big)\Big\}^{-1} \\ \mu_{q(\boldsymbol{\theta})} & \longleftarrow & \boldsymbol{\mu}_{q(\boldsymbol{\theta})} + \boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}\big\{\boldsymbol{A}^T(\boldsymbol{y} - \boldsymbol{\omega}) + (\boldsymbol{\eta}_{\boldsymbol{\theta} \to p(\boldsymbol{y}|\,\boldsymbol{\theta})})_1 \\ & & \quad + 2\,\text{vec}^{-1}\big((\boldsymbol{\eta}_{\boldsymbol{\theta} \to p(\boldsymbol{y}|\,\boldsymbol{\theta})})_2\big)\boldsymbol{\mu}_{q(\boldsymbol{\theta})}\big\}. \end{array}\right. \tag{S.19}$$

Using (S.17) the update for $\boldsymbol{\omega}$ can be expressed in terms of the natural parameter vectors as

$$\begin{aligned}\boldsymbol{\omega} \longleftarrow \exp\Bigg(& -\tfrac{1}{2}\boldsymbol{A}\Big\{\text{vec}^{-1}\Big(\big(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta})} \leftrightarrow \boldsymbol{\theta}\big)_2\Big)\Big\}^{-1}\big(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta})} \leftrightarrow \boldsymbol{\theta}\big)_1 \\ & -\tfrac{1}{4}\text{diagonal}\Big[\boldsymbol{A}\Big\{\text{vec}^{-1}\Big(\big(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta})} \leftrightarrow \boldsymbol{\theta}\big)_2\Big)\Big\}^{-1}\boldsymbol{A}^T\Big]\Bigg).\end{aligned}$$

Again using (S.17), the $\boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}$ update can be written as

$$-\tfrac{1}{2}\{\text{vec}^{-1}\big((\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta})} \leftrightarrow \boldsymbol{\theta})_2\big)\}^{-1} \longleftarrow \Big\{\boldsymbol{A}^T\text{diag}(\boldsymbol{\omega})\boldsymbol{A} - 2\,\text{vec}^{-1}\big((\boldsymbol{\eta}_{\boldsymbol{\theta} \to p(\boldsymbol{y}|\,\boldsymbol{\theta})})_2\big)\Big\}^{-1}$$

which is equivalent to

$$(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta})} \to \boldsymbol{\theta})_2 + (\boldsymbol{\eta}_{\boldsymbol{\theta} \to p(\boldsymbol{y}|\,\boldsymbol{\theta})})_2 \longleftarrow\ -\tfrac{1}{2}\text{vec}(\boldsymbol{A}^T\text{diag}(\boldsymbol{\omega})\boldsymbol{A}) + (\boldsymbol{\eta}_{\boldsymbol{\theta} \to p(\boldsymbol{y}|\,\boldsymbol{\theta})})_2$$

which, in turn, is equivalent to the second component of $\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta})\,\to\,\boldsymbol{\theta}}$ being updated according to

$$(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta})\,\to\,\boldsymbol{\theta}})_2 \longleftarrow\ -\tfrac{1}{2}\text{vec}(\boldsymbol{A}^T\text{diag}(\boldsymbol{\omega})\boldsymbol{A}). \tag{S.20}$$

For the update of the first component of $\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta})\,\to\,\boldsymbol{\theta}}$ we note that the last update of (S.19) is equivalent to

$$\begin{aligned}\boldsymbol{\Sigma}^{-1}_{q(\boldsymbol{\theta})}\mu_{q(\boldsymbol{\theta})} \ \longleftarrow\ \ &\boldsymbol{\Sigma}^{-1}_{q(\boldsymbol{\theta})}\boldsymbol{\mu}_{q(\boldsymbol{\theta})} + \boldsymbol{A}^T(\boldsymbol{y}-\boldsymbol{\omega}) + (\boldsymbol{\eta}_{\boldsymbol{\theta} \to p(\boldsymbol{y}|\,\boldsymbol{\theta})})_1 \\ &+2\,\text{vec}^{-1}\big((\boldsymbol{\eta}_{\boldsymbol{\theta} \to p(\boldsymbol{y}|\,\boldsymbol{\theta})})_2\big)\boldsymbol{\mu}_{q(\boldsymbol{\theta})}\end{aligned} \tag{S.21}$$

where, on the right-hand side,

$$\boldsymbol{\Sigma}^{-1}_{q(\boldsymbol{\theta})} = \boldsymbol{A}^T\text{diag}(\boldsymbol{\omega})\boldsymbol{A} - 2\,\text{vec}^{-1}\big((\boldsymbol{\eta}_{\boldsymbol{\theta} \to p(\boldsymbol{y}|\,\boldsymbol{\theta})})_2\big) \tag{S.22}$$

according to its updated value and

$$\mu_{q(\boldsymbol{\theta})} = -\tfrac{1}{2}\{\text{vec}^{-1}\big((\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta})} \leftrightarrow \boldsymbol{\theta})_2\big)\}^{-1}(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta})} \leftrightarrow \boldsymbol{\theta})_1 \tag{S.23}$$

is the terms of the natural parameter from the previous iteration before (S.20) has taken place. Substitution of (S.22) and (S.23) into (S.21) we get

$$\begin{aligned}&(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta})\,\to\,\boldsymbol{\theta}})_1 + (\boldsymbol{\eta}_{\boldsymbol{\theta} \to p(\boldsymbol{y}|\,\boldsymbol{\theta})})_1 \longleftarrow \\ &\Big\{-\tfrac{1}{2}\boldsymbol{A}^T\text{diag}(\boldsymbol{\omega})\boldsymbol{A} + \text{vec}^{-1}\big((\boldsymbol{\eta}_{\boldsymbol{\theta} \to p(\boldsymbol{y}|\,\boldsymbol{\theta})})_2\big)\Big\}\{\text{vec}^{-1}\big((\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta})} \leftrightarrow \boldsymbol{\theta})_2\big)\}^{-1}(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta})} \leftrightarrow \boldsymbol{\theta})_1 \\ &\quad +\boldsymbol{A}^T(\boldsymbol{y}-\boldsymbol{\omega}) + (\boldsymbol{\eta}_{\boldsymbol{\theta} \to p(\boldsymbol{y}|\,\boldsymbol{\theta})})_1 \\ &\quad -\text{vec}^{-1}\big((\boldsymbol{\eta}_{\boldsymbol{\theta} \to p(\boldsymbol{y}|\,\boldsymbol{\theta})})_2\big)\{\text{vec}^{-1}\big((\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta})} \leftrightarrow \boldsymbol{\theta})_2\big)\}^{-1}(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta})} \leftrightarrow \boldsymbol{\theta})_1\end{aligned}$$

which is equivalent to

$$(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta})\,\to\,\boldsymbol{\theta}})_1 \longleftarrow -\tfrac{1}{2}\boldsymbol{A}^T\text{diag}(\boldsymbol{\omega})\boldsymbol{A}\{\text{vec}^{-1}\big((\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta})} \leftrightarrow \boldsymbol{\theta})_2\big)\}^{-1}(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta})} \leftrightarrow \boldsymbol{\theta})_1 + \boldsymbol{A}^T(\boldsymbol{y}-\boldsymbol{\omega}).$$

Scheme (47) follows immediately.

S.2.5 Streamlined Derivation of the Approximate Marginal Log-Likelihood

When performing MFVB-based inference the variational lower bound on the marginal log-likelihood, given by (11), is commonly used to assess convergence. However, the algebra required to obtain the lower bound expression is demanding for large models. The VMP approach offers efficiencies for its calculation, which we now summarize.

In Section 2.5 we described VMP for a general statistical model with observed data $\boldsymbol{D}$ in terms of factors f_j, $1 \le j \le M$, such that $p(\boldsymbol{\theta}, \boldsymbol{D}) = \prod_{j=1}^{N} f_j$ where each f_j is a function of a sub-vector of $\boldsymbol{\theta}$. The mean field approximation to the posterior density function takes the form

$$p(\boldsymbol{\theta}|\boldsymbol{D}) \approx \prod_{i=1}^{M} q(\boldsymbol{\theta}_i)$$

for some partition $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\}$ of $\boldsymbol{\theta}$. The expressions in Winn & Bishop (2005) and Minka & Winn (2008) give rise to

$$\log \underline{p}(q; \boldsymbol{D}) = \sum_{i=1}^{M} \text{Entropy}\{q(\boldsymbol{\theta}_i)\} + \sum_{j=1}^{N} E_q\{\log(f_j)\} \tag{S.24}$$

where

$$\text{Entropy}\{q(\boldsymbol{\theta}_i)\} \equiv E_{q(\boldsymbol{\theta}_i)}\{-\log\, q(\boldsymbol{\theta}_i)\}$$

is the *entropy* (also known as the *differential entropy*) of q.

For models such that the optimal $q(\boldsymbol{\theta}_i)$ are exponential density functions, which includes each of the models treated in Sections 4 and 5, the value of $\text{Entropy}\{q(\boldsymbol{\theta}_i)\}$ can be looked up in a table. Table S.2 lists the entropies for each of the exponential family distributions covered in Section S.1. All expressions are in terms of natural parameters.

Distribution	Entropy
Bernoulli	$\log(1+e^{\eta}) - \eta\, e^{\eta}/(1+e^{\eta})$
Univariate Normal	$\frac{1}{2}\{1+\log(2\pi)\} + \frac{1}{2}\log\left(\frac{-1}{2\eta_2}\right)$
Inverse Chi-Squared	$\log\Gamma(-\eta_1-1) + \eta_1\psi(-\eta_1-1) + \log(-\eta_2) - \eta_1 - 1$
Beta	$\log\Gamma(\eta_1+1) + \log\Gamma(\eta_2+1) - \log\Gamma(\eta_1+\eta_2+2)$ $-\eta_1\psi(\eta_1+1) - \eta_2\psi(\eta_2+1) + (\eta_1+\eta_2)\psi(\eta_1+\eta_2+2)$
Inverse Gaussian	$\frac{1}{2} + \frac{1}{4}\log(\pi^2\eta_2/\eta_1^3) + \frac{3}{2}\exp\left(4(\eta_1\eta_2)^{1/2}\right)\text{Ei}\left(-4(\eta_1\eta_2)^{1/2}\right)$
Multivariate Normal	$\frac{d}{2}\{1+\log(2\pi)\} + \frac{1}{2}\log\left\lvert -\frac{1}{2}\{\text{vec}^{-1}(\boldsymbol{\eta}_2)\}^{-1}\right\rvert$
Inverse Wishart	$\sum_{j=1}^{d}\left[\log\Gamma\{-\eta_1 - \frac{1}{2}(d+j)\} + \eta_1\psi\{-\eta_1 - \frac{1}{2}(d+j)\}\right]$ $+\frac{1}{2}(d+1)\log\left\lvert -\text{vec}^{-1}(\boldsymbol{\eta}_2)\right\rvert - d\eta_1 - \frac{1}{2}\,d(d+1) + \frac{1}{4}d(d-1)\log(\pi)$

Table S.2: *Expressions for entropies in terms of natural parameters for some common exponential family distributions.*

As an example, consider VMP fitting of the linear regression model described in Section 3 and the updates of the stochastic node natural parameters given by (25). From Table S.2, the entropy contributions to $\log \underline{p}(q; \boldsymbol{y})$ are

$$\begin{aligned}
\text{Entropy}\{q(\boldsymbol{\beta})\} &= \tfrac{d}{2}\{1+\log(2\pi)\} + \tfrac{1}{2}\log\left\lvert -\tfrac{1}{2}\left\{\text{vec}^{-1}\left(\left(\boldsymbol{\eta}_{q(\boldsymbol{\beta})}\right)_2\right)\right\}^{-1}\right\rvert, \\
\text{Entropy}\{q(\sigma^2)\} &= \log\Gamma\left(-\left(\eta_{q(\sigma^2)}\right)_1 - 1\right) + \left(\eta_{q(\sigma^2)}\right)_1\psi\left(-\left(\eta_{q(\sigma^2)}\right)_1 - 1\right) \\
&\quad + \log\left(-\left(\eta_{q(\sigma^2)}\right)_2\right) - \left(\eta_{q(\sigma^2)}\right)_1 - 1, \\
\text{and}\quad \text{Entropy}\{q(a)\} &= \log\Gamma\left(-\left(\eta_{q(a)}\right)_1 - 1\right) + \left(\eta_{q(a)}\right)_1\psi\left(-\left(\eta_{q(a)}\right)_1 - 1\right) \\
&\quad + \log\left(-\left(\eta_{q(a)}\right)_2\right) - \left(\eta_{q(a)}\right)_1 - 1.
\end{aligned} \tag{S.25}$$

For conjugate models with exponential family stochastic nodes, the factor contributions reduce to linear combinations of expected values of sufficient statistics. Their formulae in terms of natural parameters can be looked up in tables such as Table S.1 in Section S.1.8. For the linear regression model of Section 3 the q-density expectation of the logarithm of the likelihood factor is

$$
\begin{aligned}
&E_{q(\boldsymbol{\beta},\sigma^2)}\{\log\, p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)\} = \\
&\quad E_{q(\sigma^2)}(1/\sigma^2)\left\{\begin{bmatrix} E_{q(\boldsymbol{\beta})}(\boldsymbol{\beta}) \\ E_{q(\boldsymbol{\beta})}\{\mathrm{vec}(\boldsymbol{\beta}\boldsymbol{\beta}^T)\}\end{bmatrix}^T \begin{bmatrix} \boldsymbol{X}^T\boldsymbol{y} \\ -\frac{1}{2}\mathrm{vec}(\boldsymbol{X}^T\boldsymbol{X})\end{bmatrix} - \tfrac{1}{2}\boldsymbol{y}^T\boldsymbol{y}\right\} \\
&\qquad -\frac{n}{2}E_{q(\sigma^2)}\{\log(\sigma^2)\} - \frac{n}{2}\log(2\pi)
\end{aligned}
$$

$$
\begin{aligned}
&= \left\{\frac{\left(\boldsymbol{\eta}_{q(\sigma^2)}\right)_1 + 1}{\left(\boldsymbol{\eta}_{q(\sigma^2)}\right)_2}\right\} \\
&\times \left\{\begin{bmatrix} -\frac{1}{2}\{\mathrm{vec}^{-1}\Big(\left(\boldsymbol{\eta}_{q(\boldsymbol{\beta})}\right)_2\Big)\}^{-1}\left(\boldsymbol{\eta}_{q(\boldsymbol{\beta})}\right)_1 \\ \frac{1}{4}\mathrm{vec}\Bigg(\Big\{\mathrm{vec}^{-1}\Big(\left(\boldsymbol{\eta}_{q(\boldsymbol{\beta})}\right)_2\Big)\Big\}^{-1} \\ \times\Big[\left(\boldsymbol{\eta}_{q(\boldsymbol{\beta})}\right)_1\left(\boldsymbol{\eta}_{q(\boldsymbol{\beta})}\right)_1^T\Big\{\mathrm{vec}^{-1}\Big(\left(\boldsymbol{\eta}_{q(\boldsymbol{\beta})}\right)_2\Big)\Big\}^{-1} - 2\,\boldsymbol{I}\Big]\Bigg)\end{bmatrix}^T \begin{bmatrix} \boldsymbol{X}^T\boldsymbol{y} \\ -\frac{1}{2}\mathrm{vec}(\boldsymbol{X}^T\boldsymbol{X})\end{bmatrix} - \tfrac{1}{2}\boldsymbol{y}^T\boldsymbol{y}\right\} \\
&\qquad -\frac{n}{2}\left\{\log\Big(-\left(\boldsymbol{\eta}_{q(\sigma^2)}\right)_2\Big) - \psi\Big(-\left(\boldsymbol{\eta}_{q(\sigma^2)}\right)_1 - 1\Big)\right\} - \frac{n}{2}\log(2\pi).
\end{aligned}
$$

The contributions from the remaining three factors in Figure 3 can be handled using similar algebra. These expressions can then be added to the $E_{q(\boldsymbol{\beta},\sigma^2)}\{\log\, p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)\}$ expression and the entropy expressions given in (S.25) to give the full $\log \underline{p}(q;\boldsymbol{y})$ expression.

For the classes of semiparametric regression models treated in Sections 4 and 5 the $E_q\{\log(f_j)\}$ terms in (S.24) can be handled efficiently via fragment categorization. The marginal log-likelihood lower bound contributions of each of the fragments identified in Sections 4 and 5 only need to be worked out once and can be tabulated and looked up.

Next we derive the $E_q\{\log(f_j)\}$-type contributions from each of the Section 4.1 fragment factors. Illustration is then provided for the penalized spline regression model introduced in Section 3.2.1. Other fragments, such as the generalized response fragments of Section 5, can be handled similarly.

S.2.5.1 Contribution from an Gaussian Prior Fragment Factor

In the notation of Section 4.1.1 the logarithm of the factor in the Gaussian prior fragment is

$$
\log\, p(\boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{\theta} \\ \mathrm{vec}(\boldsymbol{\theta}\boldsymbol{\theta}^T)\end{bmatrix}^T \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\theta}} \\ -\frac{1}{2}\mathrm{vec}(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1})\end{bmatrix} - \tfrac{1}{2}d^{\theta}\log(2\pi) - \tfrac{1}{2}\log|\boldsymbol{\Sigma}_{\boldsymbol{\theta}}|.
$$

Hence, using Table S.1,

$$E_q\{\log\, p(\boldsymbol{\theta})\} = \begin{bmatrix} -\frac{1}{2}\{\mathrm{vec}^{-1}\Big(\big(\boldsymbol{\eta}_{q(\boldsymbol{\theta})}\big)_2\Big)\}^{-1}\big(\boldsymbol{\eta}_{q(\boldsymbol{\theta})}\big)_1 \\ \\ \frac{1}{4}\mathrm{vec}\Bigg(\Big\{\mathrm{vec}^{-1}\Big(\big(\boldsymbol{\eta}_{q(\boldsymbol{\theta})}\big)_2\Big)\Big\}^{-1} \\ \times\Big[\big(\boldsymbol{\eta}_{q(\boldsymbol{\theta})}\big)_1\big(\boldsymbol{\eta}_{q(\boldsymbol{\theta})}\big)_1^T\Big\{\mathrm{vec}^{-1}\Big(\big(\boldsymbol{\eta}_{q(\boldsymbol{\theta})}\big)_2\Big)\Big\}^{-1} - 2\,\boldsymbol{I}\Big]\Bigg) \end{bmatrix}^T \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\theta}} \\ -\frac{1}{2}\mathrm{vec}(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}) \end{bmatrix}$$

$$-\tfrac{1}{2}d^{\,\theta}\log(2\pi) - \tfrac{1}{2}\log|\boldsymbol{\Sigma}_{\boldsymbol{\theta}}|.$$

S.2.5.2 Contribution from an Inverse Wishart Prior Fragment Factor

In the notation of Section 4.1.2 the logarithm of the factor in the Inverse Wishart prior fragment is

$$\log\, p(\boldsymbol{\Theta}) = \begin{bmatrix} \log|\boldsymbol{\Theta}| \\ \mathrm{vec}(\boldsymbol{\Theta}^{-1}) \end{bmatrix}^T \begin{bmatrix} -(\kappa_{\boldsymbol{\Theta}} + d^{\Theta} + 1)/2 \\ -\frac{1}{2}\mathrm{vec}(\boldsymbol{\Lambda}_{\boldsymbol{\Theta}}) \end{bmatrix} - \log(\mathcal{C}_{d^{\Theta},\kappa_{\boldsymbol{\Theta}}}) + \tfrac{1}{2}\kappa_{\boldsymbol{\Theta}}\log|\boldsymbol{\Lambda}_{\boldsymbol{\Theta}}|.$$

Table S.1 then gives

$$E_q\{\log\, p(\boldsymbol{\Theta})\} = \begin{bmatrix} \log\Big|-\mathrm{vec}^{-1}\big((\boldsymbol{\eta}_{q(\boldsymbol{\Theta})})_2\big)\Big| \\ \quad -\sum_{j=1}^{d^{\Theta}}\psi\{-(\eta_{q(\boldsymbol{\Theta})})_1 - \tfrac{1}{2}(d^{\Theta}+j)\} \\ \\ \{(\eta_{q(\boldsymbol{\Theta})})_1 + \tfrac{1}{2}(d^{\Theta}+1)\}\mathrm{vec}[\{\mathrm{vec}^{-1}\big((\boldsymbol{\eta}_{q(\boldsymbol{\Theta})})_2\big)\}^{-1}] \end{bmatrix}^T \begin{bmatrix} -\frac{1}{2}(\kappa_{\boldsymbol{\Theta}} + d^{\Theta} + 1) \\ -\frac{1}{2}\mathrm{vec}(\boldsymbol{\Lambda}_{\boldsymbol{\Theta}}) \end{bmatrix} \quad \text{(S.26)}$$

$$-\log(\mathcal{C}_{d^{\Theta},\kappa_{\boldsymbol{\Theta}}}) + \tfrac{1}{2}\kappa_{\boldsymbol{\Theta}}\log|\boldsymbol{\Lambda}_{\boldsymbol{\Theta}}|.$$

S.2.5.3 Contribution from an Iterated Inverse G-Wishart Fragment Factor

As in Section 4.1.3 we first treat the scalar case before dealing with the more delicate matrix case.

The Case of $d^{\Theta} = 1$

When $d^{\Theta} = 1$ the covariance matrices $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$ reduce to scalars $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ and the logarithm of the fragment factor is

$$\log\, p(\theta_1|\theta_2) = \begin{bmatrix} \log(\theta_1) \\ 1/\theta_1 \end{bmatrix}^T \begin{bmatrix} -\frac{1}{2}(\kappa+2) \\ -\frac{1}{2}(1/\theta_2) \end{bmatrix} - \tfrac{1}{2}\kappa\log(\theta_1) - \tfrac{1}{2}\kappa\log(2) - \log\Gamma(\tfrac{1}{2}\kappa)$$

so using Table S.1 we get

$$E_q\{\log\, p(\theta_1|\theta_2)\} = \begin{bmatrix} \log\big(-(\eta_{q(\theta_1)})_2\big) - \psi\big(-(\eta_{q(\theta_1)})_1 - 1\big) \\ \big((\eta_{q(\theta_1)})_1 + 1\big)/(\eta_{q(\theta_1)})_2 \end{bmatrix}^T \begin{bmatrix} -\frac{1}{2}(\kappa+2) \\ -\frac{1}{2}\big((\eta_{q(\theta_2)})_1 + 1\big)/(\eta_{q(\theta_2)})_2 \end{bmatrix} \quad \text{(S.27)}$$

$$-\tfrac{1}{2}\kappa\Big\{\log\big(-(\eta_{q(\theta_2)})_2\big) - \psi\big(-(\eta_{q(\theta_2)})_1 - 1\big)\Big\} - \tfrac{1}{2}\kappa\log(2) - \log\Gamma(\tfrac{1}{2}\kappa).$$

The Case of $d^{\Theta} > 1$ and G Totally Connected or Totally Disconnected

If $\boldsymbol{\Theta}_1|\boldsymbol{\Theta}_2 \sim$ Inverse-G-Wishart$(G, \kappa, \boldsymbol{\Theta}_2^{-1})$ where G is totally connected then

$$\log\, p(\boldsymbol{\Theta}_1|\boldsymbol{\Theta}_2) = \begin{bmatrix} \log|\boldsymbol{\Theta}_1| \\ \text{vec}(\boldsymbol{\Theta}_1^{-1}) \end{bmatrix}^T \begin{bmatrix} -\frac{1}{2}(\kappa + d^{\Theta} + 1) \\ -\frac{1}{2}\text{vec}(\boldsymbol{\Theta}_2^{-1}) \end{bmatrix} - \tfrac{1}{2}\kappa\, \log|\boldsymbol{\Theta}_2| - \log(\mathcal{C}_{d^{\Theta},\kappa}).$$

Table S.1 immediately gives

$$E_q\{\log\, p(\boldsymbol{\Theta}_1|\boldsymbol{\Theta}_2)\} =$$

$$\begin{bmatrix} \log|-\text{vec}^{-1}\big((\boldsymbol{\eta}_{q(\boldsymbol{\Theta}_1)})_2\big)| \\ \quad -\sum_{j=1}^{d}\psi\{-(\boldsymbol{\eta}_{q(\boldsymbol{\Theta}_1)})_1 - \frac{1}{2}(d^{\Theta}+j)\} \\ \\ \{(\boldsymbol{\eta}_{q(\boldsymbol{\Theta}_1)})_1 + \frac{1}{2}(d^{\Theta}+1)\}\text{vec}[\{\text{vec}^{-1}\big((\boldsymbol{\eta}_{q(\boldsymbol{\Theta}_1)})_2\big)\}^{-1}] \end{bmatrix}^T \begin{bmatrix} -\frac{1}{2}(\kappa + d^{\Theta} + 1) \\ -\frac{1}{2}\text{vec}(E_{q(\boldsymbol{\Theta}_2)}(\boldsymbol{\Theta}_2^{-1})) \end{bmatrix}$$

$$-\tfrac{1}{2}\kappa\, E_{q(\boldsymbol{\Theta}_2)}\{\log|\boldsymbol{\Theta}_2|\} - \log(\mathcal{C}_{d^{\Theta},\kappa}).$$

If $\boldsymbol{\Theta}_2$ has a totally disconnected Inverse G-Wishart distribution then

$$E_{q(\boldsymbol{\Theta}_2)}\{\log|\boldsymbol{\Theta}_2|\} = \log\Big|-\text{vec}^{-1}\big(\boldsymbol{\eta}_{q(\boldsymbol{\Theta}_2)}\big)_2\Big| - \sum_{j=1}^{d^{\Theta}}\psi\big\{-\big(\boldsymbol{\eta}_{q(\boldsymbol{\Theta}_2)}\big)_1 - \tfrac{1}{2}(d^{\Theta}+j)\big\}$$

and

$$E_{q(\boldsymbol{\Theta}_2)}(\boldsymbol{\Theta}_2^{-1}) = \Big\{\big(\boldsymbol{\eta}_{q(\boldsymbol{\Theta}_2)}\big)_1 + \tfrac{1}{2}(d^{\Theta}+1)\Big\}\Big\{\text{vec}^{-1}\big(\boldsymbol{\eta}_{q(\boldsymbol{\Theta}_2)}\big)_2\Big\}^{-1}.$$

If $\boldsymbol{\Theta}_2$ has an totally disconnected Inverse G-Wishart distribution, which is the case for the auxiliary variable representation of the covariance matrix prior of Huang & Wand (2013), then

$$E_{q(\boldsymbol{\Theta}_2)}\{\log|\boldsymbol{\Theta}_2|\} = \sum_{j=1}^{d^{\Theta}}\Big\{\log\Big(-\big(\boldsymbol{\eta}_{q((\boldsymbol{\Theta}_2)_{jj})}\big)_2\Big) - \psi\Big(-\big(\boldsymbol{\eta}_{q((\boldsymbol{\Theta}_2)_{jj})}\big)_1 - 1\Big)\Big\}$$

and

$$E_{q(\boldsymbol{\Theta}_2)}(\boldsymbol{\Theta}_2^{-1}) = \underset{1\le j\le d^{\Theta}}{\text{diag}}\left(\frac{\big(\boldsymbol{\eta}_{q((\boldsymbol{\Theta}_2)_{jj})}\big)_1 + 1}{\big(\boldsymbol{\eta}_{q((\boldsymbol{\Theta}_2)_{jj})}\big)_2}\right).$$

Other Cases

The other cases such as $\boldsymbol{\Theta}_1$ having a Inverse G-Wishart distribution with G partially connected or totally disconnected are not common in Bayesian semiparametric regression analysis and are left aside here.

S.2.5.4 Contribution from a Gaussian Penalization Factor

For this fragment, the logarithm of the factor is

$$\log\, p(\boldsymbol{\theta}_0, \ldots, \boldsymbol{\theta}_L|\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_L) = \begin{bmatrix} \boldsymbol{\theta}_0 \\ \text{vec}(\boldsymbol{\theta}_0\boldsymbol{\theta}_0^T) \end{bmatrix}^T \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\theta}_0}^{-1}\boldsymbol{\mu}_{\boldsymbol{\theta}_0} \\ -\frac{1}{2}\text{vec}(\boldsymbol{\Sigma}_{\boldsymbol{\theta}_0}^{-1}) \end{bmatrix} - \tfrac{1}{2}d_0^{\theta}\log(2\pi) - \tfrac{1}{2}\log|\boldsymbol{\Sigma}_{\boldsymbol{\theta}_0}|$$

$$+\sum_{\ell=1}^{L}\left\{\begin{bmatrix} \boldsymbol{\theta}_\ell \\ \text{vec}(\boldsymbol{\theta}_\ell\boldsymbol{\theta}_\ell^T) \end{bmatrix}^T \begin{bmatrix} \mathbf{0} \\ -\frac{1}{2}\text{vec}\big(\boldsymbol{I}_{m_\ell}\otimes\boldsymbol{\Theta}_\ell^{-1}\big) \end{bmatrix} - \tfrac{1}{2}m_\ell d_\ell^{\Theta}\log(2\pi) - \tfrac{1}{2}m_\ell\log|\boldsymbol{\Theta}_\ell|\right\}.$$

Application of results in Table S.1 then gives

$$
\begin{aligned}
&E_q\{\log\, p(\boldsymbol{\theta}_0,\ldots,\boldsymbol{\theta}_L|\boldsymbol{\Theta}_1,\ldots,\boldsymbol{\Theta}_L)\} \\
&= \begin{bmatrix} -\frac{1}{2}\left\{\mathrm{vec}^{-1}\Big(\big(\boldsymbol{\eta}_{q(\boldsymbol{\theta}_0)}\big)_2\Big)\right\}^{-1}\big(\boldsymbol{\eta}_{q(\boldsymbol{\theta}_0)}\big)_1 \\ \frac{1}{4}\mathrm{vec}\left(\left\{\mathrm{vec}^{-1}\Big(\big(\boldsymbol{\eta}_{q(\boldsymbol{\theta}_0)}\big)_2\Big)\right\}^{-1} \times\Big[\big(\boldsymbol{\eta}_{q(\boldsymbol{\theta}_0)}\big)_1\big(\boldsymbol{\eta}_{q(\boldsymbol{\theta}_0)}\big)_1^T\left\{\mathrm{vec}^{-1}\Big(\big(\boldsymbol{\eta}_{q(\boldsymbol{\theta}_0)}\big)_2\Big)\right\}^{-1}-2\,\boldsymbol{I}\Big]\right) \\ -\frac{1}{2}d_0^{\theta}\log(2\pi)-\frac{1}{2}\log|\boldsymbol{\Sigma}_{\boldsymbol{\theta}_0}| \end{bmatrix}^T \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\theta}_0}^{-1}\boldsymbol{\mu}_{\boldsymbol{\theta}_0} \\ -\frac{1}{2}\mathrm{vec}(\boldsymbol{\Sigma}_{\boldsymbol{\theta}_0}^{-1}) \end{bmatrix} \\
&\quad +\sum_{\ell=1}^{L}\left\{\begin{bmatrix} \frac{1}{4}\mathrm{vec}\left(\left\{\mathrm{vec}^{-1}\Big(\big(\boldsymbol{\eta}_{q(\boldsymbol{\theta}_\ell)}\big)_2\Big)\right\}^{-1} \times\Big[\big(\boldsymbol{\eta}_{q(\boldsymbol{\theta}_\ell)}\big)_1\big(\boldsymbol{\eta}_{q(\boldsymbol{\theta}_\ell)}\big)_1^T\left\{\mathrm{vec}^{-1}\Big(\big(\boldsymbol{\eta}_{q(\boldsymbol{\theta}_\ell)}\big)_2\Big)\right\}^{-1}-2\,\boldsymbol{I}\Big]\right) \end{bmatrix}^T \right. \\
&\quad \times\Big[-\tfrac{1}{2}\mathrm{vec}\Big(\boldsymbol{I}_{m_\ell}\otimes\Big[\{(\eta_{q(\boldsymbol{\Theta}_\ell)})_1+\tfrac{1}{2}(d_\ell^{\Theta}+1)\}\{\mathrm{vec}^{-1}((\boldsymbol{\eta}_{q(\boldsymbol{\Theta}_\ell)})_2)\}^{-1}\Big]\Big)\Big] \\
&\quad -\tfrac{1}{2}m_\ell d_\ell^{\Theta}\log(2\pi) \\
&\quad \left.-\tfrac{1}{2}m_\ell\log\Big|-\mathrm{vec}^{-1}((\boldsymbol{\eta}_{q(\boldsymbol{\Theta}_\ell)})_2)\Big|+\tfrac{1}{2}m_\ell\textstyle\sum_{j=1}^{d_\ell^{\Theta}}\psi\{-(\eta_{q(\boldsymbol{\Theta}_\ell)})_1-\tfrac{1}{2}(d_\ell^{\Theta}+j)\}\right\}.
\end{aligned}
\tag{S.28}
$$

S.2.5.5 Contribution from a Gaussian Likelihood Factor

The logarithm of the factor is

$$
\log\, p(\boldsymbol{y}|\,\boldsymbol{\theta}_1,\theta_2)=\frac{1}{\theta_2}\left\{\begin{bmatrix}\boldsymbol{\theta}_1\\ \mathrm{vec}(\boldsymbol{\theta}_1\boldsymbol{\theta}_1^T)\end{bmatrix}^T\begin{bmatrix}\boldsymbol{A}^T\boldsymbol{y}\\ -\frac{1}{2}\mathrm{vec}(\boldsymbol{A}^T\boldsymbol{A})\end{bmatrix}-\tfrac{1}{2}\boldsymbol{y}^T\boldsymbol{y}\right\}-\frac{n}{2}\log(\theta_2)-\frac{n}{2}\log(2\pi).
$$

Then, from Table S.1 we have

$$
\begin{aligned}
&E_q\{\log\, p(\boldsymbol{y}|\boldsymbol{\theta}_1,\theta_2)\}= \\
&=\left\{\frac{\big(\boldsymbol{\eta}_{q(\theta_2)}\big)_1+1}{\big(\boldsymbol{\eta}_{q(\theta_2)}\big)_2}\right\} \\
&\times\left\{\begin{bmatrix} -\frac{1}{2}\left\{\mathrm{vec}^{-1}\Big(\big(\boldsymbol{\eta}_{q(\boldsymbol{\theta}_1)}\big)_2\Big)\right\}^{-1}\big(\boldsymbol{\eta}_{q(\boldsymbol{\theta}_1)}\big)_1 \\ \frac{1}{4}\mathrm{vec}\left(\left\{\mathrm{vec}^{-1}\Big(\big(\boldsymbol{\eta}_{q(\boldsymbol{\theta}_1)}\big)_2\Big)\right\}^{-1} \times\Big[\big(\boldsymbol{\eta}_{q(\boldsymbol{\theta}_1)}\big)_1\big(\boldsymbol{\eta}_{q(\boldsymbol{\theta}_1)}\big)_1^T\left\{\mathrm{vec}^{-1}\Big(\big(\boldsymbol{\eta}_{q(\boldsymbol{\theta}_1)}\big)_2\Big)\right\}^{-1}-2\,\boldsymbol{I}\Big]\right) \end{bmatrix}^T \begin{bmatrix}\boldsymbol{A}^T\boldsymbol{y}\\ -\frac{1}{2}\mathrm{vec}(\boldsymbol{A}^T\boldsymbol{A})\end{bmatrix}-\tfrac{1}{2}\boldsymbol{y}^T\boldsymbol{y}\right\} \\
&\qquad -\tfrac{n}{2}\left\{\log\Big(-\big(\boldsymbol{\eta}_{q(\theta_2)}\big)_2\Big)-\psi\Big(-\big(\boldsymbol{\eta}_{q(\theta_2)}\big)_1-1\Big)\right\}-\tfrac{n}{2}\log(2\pi).
\end{aligned}
\tag{S.29}
$$

S.2.5.6 Illustration for Penalized Spline Nonparametric Regression

We now illustrate approximate marginal log-likelihood calculation for penalized spline regression, corresponding to the factor graph shown in Figure 5. Using Table S.2, the first

two entropy contributions to $\log \underline{p}(q; \boldsymbol{y})$ are

$$\text{Entropy}\{q(\boldsymbol{\beta}, \boldsymbol{u})\} = \tfrac{2+K}{2}\{1 + \log(2\pi)\} + \tfrac{1}{2}\log\Big| - \tfrac{1}{2}\Big\{\text{vec}^{-1}\Big(\big(\boldsymbol{\eta}_{q(\boldsymbol{\beta},\boldsymbol{u})}\big)_2\Big)\Big\}^{-1}\Big| \tag{S.30}$$

and

$$\begin{aligned}\text{Entropy}\{q(\sigma_\varepsilon^2)\} \;=\; &\log\Gamma\Big(-\big(\eta_{q(\sigma_\varepsilon^2)}\big)_1 - 1\Big) + \big(\eta_{q(\sigma_\varepsilon^2)}\big)_1\psi\Big(-\big(\eta_{q(\sigma_\varepsilon^2)}\big)_1 - 1\Big) \\ &+ \log\Big(-\big(\eta_{q(\sigma_\varepsilon^2)}\big)_2\Big) - \big(\eta_{q(\sigma_\varepsilon^2)}\big)_1 - 1.\end{aligned} \tag{S.31}$$

The entropy contributions

$$\text{Entropy}\{q(\sigma_u^2)\}, \quad \text{Entropy}\{q(a_\varepsilon)\} \quad \text{and} \quad \text{Entropy}\{q(a_u)\} \tag{S.32}$$

take exactly the same form as (S.31) but as functions of the natural parameter vectors $\boldsymbol{\eta}_{q(\sigma_u^2)}$, $\boldsymbol{\eta}_{q(a_\varepsilon)}$ and $\boldsymbol{\eta}_{q(a_u)}$.

The factor contributions are each special cases of (S.26)–(S.29). The contribution from the factor $p(a_\varepsilon)$ is

$$\text{the right-hand side of (S.26) with } \boldsymbol{\Theta} = a_\varepsilon,\ d^\Theta = 1,\ \kappa_{\boldsymbol{\Theta}} = 1 \text{ and } \boldsymbol{\Lambda}_{\boldsymbol{\Theta}} = 1/A_\varepsilon^2. \tag{S.33}$$

The contribution from the factor $p(a_u)$ is

$$\text{the right-hand side of (S.26) with } \boldsymbol{\Theta} = a_u,\ d^\Theta = 1,\ \kappa_{\boldsymbol{\Theta}} = 1 \text{ and } \boldsymbol{\Lambda}_{\boldsymbol{\Theta}} = 1/A_u^2. \tag{S.34}$$

The contribution from the factor $p(\sigma_\varepsilon^2 |\, a_\varepsilon)$ is

$$\text{the right-hand side of (S.27) with } \theta_1 = \sigma_\varepsilon^2,\ \theta_2 = a_\varepsilon \text{ and } \kappa = 1. \tag{S.35}$$

The contribution from the factor $p(\sigma_u^2 |\, a_u)$ is

$$\text{the right-hand side of (S.27) with } \theta_1 = \sigma_u^2,\ \theta_2 = a_u \text{ and } \kappa = 1. \tag{S.36}$$

The contribution from the factor $p(\boldsymbol{\beta}, \boldsymbol{u} |\, \sigma_u^2)$ is

$$\begin{aligned}&\text{the right-hand side of (S.28) with } L = 1,\ d^\Theta = 1,\ m_1 = K,\ \boldsymbol{\theta}_0 = \boldsymbol{\beta}, \\ &\boldsymbol{\theta}_1 = \boldsymbol{u} \text{ and } \boldsymbol{\Theta}_1 = \sigma_u^2.\end{aligned} \tag{S.37}$$

The contribution from the factor $p(\boldsymbol{y} |\, \boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2)$ is

$$\text{the right-hand side of (S.29) with } \boldsymbol{\theta}_1 = (\boldsymbol{\beta}, \boldsymbol{u}),\ \theta_2 = \sigma_\varepsilon^2 \text{ and } \boldsymbol{A} = [\boldsymbol{X}\ \boldsymbol{Z}]. \tag{S.38}$$

During the VMP iterations for fitting (26), the approximate marginal log-likelihood $\log \underline{p}(\boldsymbol{y}; q)$ can be computed by obtaining

$$\begin{aligned}\boldsymbol{\eta}_{q(\boldsymbol{\beta},\boldsymbol{u})} &\longleftarrow \boldsymbol{\eta}_{p(\boldsymbol{\beta},\boldsymbol{u}|\,\sigma_u^2)\,\to\,(\boldsymbol{\beta},\boldsymbol{u})} + \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2)\,\to\,(\boldsymbol{\beta},\boldsymbol{u})}, \\ \boldsymbol{\eta}_{q(\sigma_\varepsilon^2)} &\longleftarrow \boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2)\,\to\,\sigma_\varepsilon^2} + \boldsymbol{\eta}_{p(\sigma_\varepsilon^2|a_\varepsilon)\,\to\,\sigma_\varepsilon^2}, \\ \boldsymbol{\eta}_{q(\sigma_u^2)} &\longleftarrow \boldsymbol{\eta}_{p(\boldsymbol{\beta},\boldsymbol{u}|\,\sigma_u^2)\,\to\,\sigma_u^2} + \boldsymbol{\eta}_{p(\sigma_u^2|a_u)\,\to\,\sigma_u^2}, \\ \boldsymbol{\eta}_{q(a_\varepsilon)} &\longleftarrow \boldsymbol{\eta}_{p(\sigma_\varepsilon^2|\,a_\varepsilon)\,\to\,a_\varepsilon} + \boldsymbol{\eta}_{p(a_\varepsilon)\,\to\,a_\varepsilon} \\ \text{and}\quad \boldsymbol{\eta}_{q(a_u)} &\longleftarrow \boldsymbol{\eta}_{p(\sigma_u^2|\,a_u)\,\to\,a_u} + \boldsymbol{\eta}_{p(a_u)\,\to\,a_u}\end{aligned}$$

and summing up the entropy contributions (S.30–S.32) and the factor contributions (S.33)–(S.38).

Additional References

Murphy, K.P. (2012). *Machine Learning: A Probabilistic Perspective*. Cambridge, Massachusetts: The MIT Press.

Menictas, M. and Wand, M.P. (2015). Variational inference for heteroscedastic semiparametric regression. *Australian and New Zealand Journal of Statistics*, **57**, 119–138.

Rohde, D. & Wand, M.P. (2016). Semiparametric mean field variational Bayes: General principles and numerical issues. *Journal of Machine Learning Research*, **17(172)**, 1–47.