Expectation Propagation on the Maximum of Correlated Normal Variables

Philipp Hennig
Cavendish Laboratory
University of Cambridge
CB3 0HE Cambridge, UK
ph347@cam.ac.uk

July 2009

Abstract

Many inference problems involving questions of optimality ask for the maximum or the minimum of a finite set of unknown quantities. This technical report derives the first two posterior moments of the maximum of two correlated Gaussian variables and the first two posterior moments of the two generating variables (corresponding to Gaussian approximations minimizing relative entropy). It is shown how this can be used to build a heuristic approximation to the maximum relationship over a finite set of Gaussian variables, allowing approximate inference by Expectation Propagation on such quantities.

1 Introduction

Many optimization problems involve inference on the maximum or minimum of a set of variables. This very broad class includes shortest path problems [Burton and Toint, 1992], Reinforcement Learning [Dearden et al., 1998], and scientific inference in Seismology [Neumann-Denzau and Behrens, 1984], to name but a few. Often, there is a corresponding inverse optimization problem [Ahuja and Orlin, 2001, Heuberger, 2004], where the optimal solution is known with some uncertainty and the question is about the quantities generating this optimum. Most contemporary algorithms for this case aim to provide a point estimate (typically the least-squares solution), but have trouble offering an error estimate on this estimate as well.

This work derives (Section 2) mean and variance of the posterior of the maximum of two correlated Gaussian variables (for forward optimization problems), and the mean and variance on the posterior of the Gaussian variables generating the maximum (for inverse optimization problems). These two moments correspond to the approximation within the exponential family of Gaussian distributions minimizing the Kullback-Leibler Divergence (relative entropy) to the true posterior. It will be shown how these results can be used to build a heuristic approximation to the max of a finite set of normal variables (Section 3). Together, this provides the necessary results for Expectation Propagation [Minka, 2001] on graphs involving the "max" relationship. Because maximum and minimum obey the simple relationship $\max(\{x_i\}) = -\min(\{-x_i\})$, this also allows inference on the minimum where necessary. Limitations of this approximation are examined in Section 4.

The moments of the normalized likelihood function of the maximum of two normal variables have previously been derived by Clark [1961]. To my best knowledge, this is the first publication deriving the full posterior, and the first to report the posterior for the inverse problem (see also Section 2.4.3).

2 The Maximum of Two Gaussian Variables

2.1 Notation

We consider two normally distributed variables x_1 and x_2, forming the vector $\boldsymbol{x}$. Let there be some prior (i.e. outside) information $\mathfrak{I}_\mathfrak{g}$ giving rise to the belief $\boldsymbol{x}, \mathfrak{I}_\mathfrak{g}$

$$p(x_1, x_2|\mathfrak{I}_\mathfrak{g}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_\mathfrak{g}, \boldsymbol{\Sigma}_\mathfrak{g}) = \frac{1}{2\pi\sigma_{\mathfrak{g}1}\sigma_{\mathfrak{g}2}(1-\rho^2)^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_\mathfrak{g})^\mathrm{T}\boldsymbol{\Sigma}_\mathfrak{g}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_\mathfrak{g})\right) \tag{1}$$

over their values. Here we have defined a mean vector $\boldsymbol{\mu}_\mathfrak{g} = (\mu_{\mathfrak{g}1}, \mu_{\mathfrak{g}2})^\mathrm{T}$ and a covariance matrix $\boldsymbol{\Sigma}_\mathfrak{g}$. The latter has the form

$$\boldsymbol{\Sigma}_\mathfrak{g} = \begin{pmatrix} \sigma_{\mathfrak{g}1}^2 & \rho\sigma_{\mathfrak{g}1}\sigma_{\mathfrak{g}2} \\ \rho\sigma_{\mathfrak{g}1}\sigma_{\mathfrak{g}2} & \sigma_{\mathfrak{g}2}^2 \end{pmatrix} \qquad \text{and thus} \qquad \boldsymbol{\Sigma}_\mathfrak{g}^{-1} = \frac{1}{\sigma_{\mathfrak{g}1}^2\sigma_{\mathfrak{g}2}^2(1-\rho^2)} \begin{pmatrix} \sigma_{\mathfrak{g}2}^2 & -\rho\sigma_{\mathfrak{g}1}\sigma_{\mathfrak{g}2} \\ -\rho\sigma_{\mathfrak{g}1}\sigma_{\mathfrak{g}2} & \sigma_{\mathfrak{g}1}^2 \end{pmatrix} \tag{2}$$

with the *linear coefficient of correlation* ρ

$$\rho = \frac{\mathrm{cov}(x_1, x_2)}{\sigma_{\mathfrak{g}1}\sigma_{\mathfrak{g}2}} \tag{3}$$

(for notational convenience, the index $\mathfrak{g}$ is dropped from ρ because there will be no chance for confusion). We further introduce the variable m which is defined through $m = \max(x_1, x_2)$, and we assume that there is some outside prior information $\mathfrak{I}_\mathfrak{m}$ on the value of m as well: $m, \mathfrak{I}_\mathfrak{m}$

$$p(m|\mathfrak{I}_\mathfrak{m}) = \mathcal{N}(m; \mu_\mathfrak{m}, \sigma_\mathfrak{m}^2) \tag{4}$$

The inference problems to be solved are

- The posterior over m given both $\mathfrak{I}_\mathfrak{m}$ and $\mathfrak{I}_\mathfrak{g}$ (jointly called $\mathfrak{I}_\mathfrak{c}$):

$$p(m|\mathfrak{I}_\mathfrak{c}) = \frac{p(m|\mathfrak{I}_\mathfrak{m})\int p(\boldsymbol{x}|m)p(\boldsymbol{x}|\mathfrak{I}_\mathfrak{g})\,\mathrm{d}\boldsymbol{x}}{\int\left[p(m|\mathfrak{I}_\mathfrak{m})\int p(\boldsymbol{x}|m)p(\boldsymbol{x}|\mathfrak{I}_\mathfrak{g})\,\mathrm{d}\boldsymbol{x}\right]\mathrm{d}m} = Z^{-1}p(m|\mathfrak{I}_\mathfrak{m})\int p(\boldsymbol{x}|m)p(\boldsymbol{x}|\mathfrak{I}_\mathfrak{g})\,\mathrm{d}\boldsymbol{x} \tag{5}$$

  with the normalization constant $Z = \iint p(\boldsymbol{x}, m|\mathfrak{I}_\mathfrak{c})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}m$. This problem will be called the "forward" problem here.

- The posterior over $\boldsymbol{x}$ given $\mathfrak{I}_\mathfrak{c}$,

$$p(\boldsymbol{x}|\mathfrak{I}_\mathfrak{c}) = \frac{p(\boldsymbol{x}|\mathfrak{I}_\mathfrak{g})\int p(m|\boldsymbol{x})p(m|\mathfrak{I}_\mathfrak{m})\,\mathrm{d}m}{\int\left[p(\boldsymbol{x}|\mathfrak{I}_\mathfrak{g})\int p(\boldsymbol{x}|m)p(m|\mathfrak{I}_\mathfrak{m})\,\mathrm{d}m\right]\mathrm{d}\boldsymbol{x}} = Z^{-1}p(\boldsymbol{x}|\mathfrak{I}_\mathfrak{g})\int p(m|\boldsymbol{x})p(m|\mathfrak{I}_\mathfrak{m})\,\mathrm{d}m \tag{6}$$

  This problem will be called the "inverse" problem.

Throughout the derivations, the notation $\mathcal{N}, \phi, \Phi$

$$\begin{aligned} \mathcal{N}(x; \mu, \sigma^2) &\equiv \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \\ \phi(x) &\equiv \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \\ \Phi(x) &\equiv \int_{-\infty}^{x} \phi(t)\,\mathrm{d}t = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)\right] \end{aligned} \tag{7}$$

will be used to denote the general and standard normal probability density functions (PDF) and the standard normal cumulative distribution function (CDF).

2.2 Some Integrals

The derivations in this paper will repeatedly feature certain integrals. The first two incomplete moments of the standard Gaussian are

$$\begin{aligned} \int_{-\infty}^{y} t\phi(t)\,\mathrm{d}t &= -\phi(y) \\ \int_{-\infty}^{y} t^2\phi(t)\,\mathrm{d}t &= \Phi(y) - y\phi(y) \end{aligned} \tag{8}$$

this is obvious directly from differentiation. A simple substitution gives

$$\int_{-\infty}^{y} t\mathcal{N}(t; \alpha, \beta^2)\,\mathrm{d}t = \alpha\Phi\left(\frac{y-\alpha}{\beta}\right) - \beta\phi\left(\frac{y-\alpha}{\beta}\right) \tag{9}$$

$$\int_{-\infty}^{y} t^2\mathcal{N}(t; \alpha, \beta^2)\,\mathrm{d}t = (\alpha^2+\beta^2)\Phi\left(\frac{y-\alpha}{\beta}\right) - (\alpha+y)\beta\phi\left(\frac{y-\alpha}{\beta}\right) \tag{10}$$

Further, we will use the integrals

$$\begin{aligned}
\int_{-\infty}^{\infty} \Phi\left(\frac{x-a}{b}\right) \mathcal{N}(x;\alpha,\beta^2)\,\mathrm{d}x &= \Phi(z) \\
\int_{-\infty}^{\infty} x\Phi\left(\frac{x-a}{b}\right) \mathcal{N}(x;\alpha,\beta^2)\,\mathrm{d}x &= \alpha\Phi(z) + \frac{\beta^2}{b\sqrt{1+\beta^2/b^2}}\phi(z) \\
\int_{-\infty}^{\infty} x^2\Phi\left(\frac{x-a}{b}\right) \mathcal{N}(x;\alpha,\beta^2)\,\mathrm{d}x &= (\alpha^2+\beta^2)\Phi(z) + \left[2\alpha\frac{\beta^2}{b\sqrt{1+\beta^2/b^2}} - z\frac{\beta^4}{b^2+\beta^2}\right]\phi(z) \\
\text{where} \qquad z &= \frac{\alpha-a}{b\sqrt{1+\beta^2/b^2}}
\end{aligned} \tag{11}$$

A derivation of these results can, for example, be found in Rasmussen and Williams [2006, section 3.9]

2.3 Analytic Forms

2.3.1 Forward Problem

Neither of the posterior distributions are normal themselves. The forward posterior is ν_1, ν_2

$$\begin{aligned}
p(m|\mathfrak{I}_\mathfrak{c}) &= Z^{-1} p(m|\mathfrak{I}_\mathfrak{m}) \iint_{-\infty}^{\infty} p(\boldsymbol{x}|m) p(\boldsymbol{x}|\mathfrak{I}_\mathfrak{g})\,\mathrm{d}\boldsymbol{x} \\
&= Z^{-1} p(m|\mathfrak{I}_\mathfrak{m}) \int_{-\infty}^{\infty} \left[\int_{-\infty}^{x_1} \delta(x_1-m) p(\boldsymbol{x}|\mathfrak{I}_\mathfrak{g})\,\mathrm{d}x_2 + \int_{x_1}^{\infty} \delta(x_2-m) p(\boldsymbol{x}|\mathfrak{I}_\mathfrak{g})\,\mathrm{d}x_2\right]\mathrm{d}x_1 \\
&= Z^{-1} \underbrace{p(m|\mathfrak{I}_\mathfrak{m}) \int_{-\infty}^{\infty} \delta(x_1-m) \int_{-\infty}^{x_1} p(\boldsymbol{x}|\mathfrak{I}_\mathfrak{g})\,\mathrm{d}x_2\,\mathrm{d}x_1}_{\nu_1} \\
&\quad + Z^{-1} \underbrace{p(m|\mathfrak{I}_\mathfrak{m}) \int_{-\infty}^{\infty} \delta(x_2-m) \int_{-\infty}^{x_2} p(\boldsymbol{x}|\mathfrak{I}_\mathfrak{g})\,\mathrm{d}x_1\,\mathrm{d}x_2}_{\nu_2}
\end{aligned} \tag{12}$$

For a motivation of the change in the integration ranges from the second to the third line in Equation (12), consider the sketch in Figure 1. Since the two summands are related to each

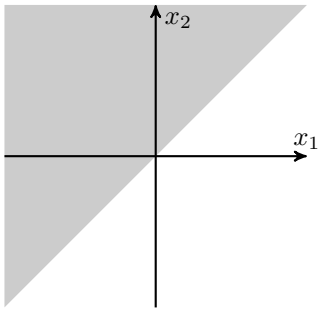

Figure 1: Sketch of the integration range for ν_2. The open set $(x_1, x_2) \in ((-\infty, \infty), (x_1, \infty))$ is identical to the open set $(x_1, x_2) \in ((-\infty, x_2), (-\infty, \infty))$.

other through the symmetry $x_1 \leftrightarrow x_2$, consider only the first term, ν_1. To solve the integrals, note that the bi-variate Gaussian $p(\boldsymbol{x}|\mathfrak{I}_\mathfrak{g})$ can be re-written as

$$\begin{aligned}
p(x_1, x_2|\mathfrak{I}_\mathfrak{g}) &= p(x_1|\mathfrak{I}_\mathfrak{g}) p(x_2|x_1, \mathfrak{I}_\mathfrak{g}) \\
&= \frac{1}{\sqrt{2\pi\sigma_{\mathfrak{g}1}^2}} \exp\left[-\frac{1}{2}\left(\frac{x_1-\mu_{\mathfrak{g}1}}{\sigma_{\mathfrak{g}1}}\right)^2\right] \\
&\quad \frac{1}{\sqrt{2\pi\sigma_{\mathfrak{g}2}^2(1-\rho^2)}} \exp\left[-\frac{1}{2\sigma_{\mathfrak{g}2}^2(1-\rho^2)}\left(x_2 - \left(\mu_{\mathfrak{g}2} + \rho\frac{\sigma_{\mathfrak{g}2}}{\sigma_{\mathfrak{g}1}}(x_1-\mu_{\mathfrak{g}1})\right)\right)^2\right]
\end{aligned} \tag{13}$$

So we can simplify ν_1 to

$$\begin{aligned}\nu_1 &= p(m|\mathfrak{I}_\mathfrak{m})\mathcal{N}(m;\mu_{\mathfrak{g}1},\sigma^2_{\mathfrak{g}1})\int_{-\infty}^{m}\frac{1}{\sqrt{2\pi\sigma^2_{\mathfrak{g}2}(1-\rho^2)}}\exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{x_2-\mu_{\mathfrak{g}2}}{\sigma_{\mathfrak{g}2}}-\rho\frac{m-\mu_{\mathfrak{g}1}}{\sigma_{\mathfrak{g}1}}\right)^2\right]\mathrm{d}x_2 \\ &= p(m|\mathfrak{I}_\mathfrak{m})\mathcal{N}(m;\mu_{\mathfrak{g}1},\sigma^2_{\mathfrak{g}1})\int_{-\infty}^{m}\frac{1}{\sqrt{2\pi\sigma^2_{\mathfrak{g}2}(1-\rho^2)}}\exp\left[-\frac{1}{2}\left(\frac{x_2-\mu_{\mathfrak{g}2}-\rho\frac{\sigma_{\mathfrak{g}2}}{\sigma_{\mathfrak{g}1}}(m-\mu_{\mathfrak{g}1})}{\sigma_{\mathfrak{g}2}(1-\rho^2)^{1/2}}\right)^2\right]\mathrm{d}x_2\end{aligned} \tag{14}$$

We introduce the substitution

$$t(x_2)\equiv\frac{x_2-\mu_{\mathfrak{g}2}-\rho\frac{\sigma_{\mathfrak{g}2}}{\sigma_{\mathfrak{g}1}}(m-\mu_{\mathfrak{g}1})}{\sigma_{\mathfrak{g}2}(1-\rho^2)^{1/2}}\quad\text{with Jacobian}\quad\frac{\mathrm{d}t}{\mathrm{d}x_2}=\frac{1}{\sigma_2(1-\rho^2)^{1/2}} \tag{15}$$

Which allows us to solve the integral and find the posterior up to normalization

$$\begin{aligned}p(m|\mathfrak{I}_\mathfrak{c}) = Z^{-1}&\mathcal{N}(\mu_\mathfrak{m};\mu_{\mathfrak{g}1},\sigma^2_\mathfrak{m}+\sigma^2_{\mathfrak{g}1})\mathcal{N}(m;\mu_{\mathfrak{c}1},\sigma^2_{\mathfrak{c}1})\Phi\left(\frac{(\sigma_{\mathfrak{g}1}-\rho\sigma_{\mathfrak{g}2})m-\sigma_{\mathfrak{g}1}\mu_{\mathfrak{g}2}+\rho\sigma_{\mathfrak{g}2}\mu_{\mathfrak{g}1}}{\sigma_{\mathfrak{g}1}\sigma_{\mathfrak{g}2}(1-\rho^2)^{1/2}}\right) \\ &+ Z^{-1}\mathcal{N}(\mu_\mathfrak{m};\mu_{\mathfrak{g}2},\sigma^2_\mathfrak{m}+\sigma^2_{\mathfrak{g}2})\mathcal{N}(m;\mu_{\mathfrak{c}2},\sigma^2_{\mathfrak{c}2})\Phi\left(\frac{(\sigma_{\mathfrak{g}2}-\rho\sigma_{\mathfrak{g}1})m-\sigma_{\mathfrak{g}2}\mu_{\mathfrak{g}1}+\rho\sigma_{\mathfrak{g}1}\mu_{\mathfrak{g}2}}{\sigma_{\mathfrak{g}2}\sigma_{\mathfrak{g}1}(1-\rho^2)^{1/2}}\right)\end{aligned} \tag{16}$$

Where we have used the abbreviations $\mu_\mathfrak{c},\sigma^2_\mathfrak{c}$

$$\sigma^2_{\mathfrak{c}1}\equiv\frac{\sigma^2_{\mathfrak{g}1}\sigma^2_\mathfrak{m}}{\sigma^2_{\mathfrak{c}1}+\sigma^2_\mathfrak{m}}\qquad\text{and}\qquad\mu_{\mathfrak{c}1}\equiv\left(\frac{\mu_{\mathfrak{g}1}}{\sigma^2_{\mathfrak{g}1}}+\frac{\mu_\mathfrak{m}}{\sigma^2_\mathfrak{m}}\right)\sigma^2_{\mathfrak{c}1} \tag{17}$$

for the mean and variance of the product of two Gaussians[1], and analogously for $\mu_{\mathfrak{c}2}$ and $\sigma_{\mathfrak{c}2}$. To find the normalization constant Z, we use the first identity in Equation (11) to get Z

$$Z=\mathcal{N}(\mu_\mathfrak{m};\mu_{\mathfrak{g}1},\sigma^2_\mathfrak{m}+\sigma^2_{\mathfrak{g}1})\Phi(k_1)+\mathcal{N}(\mu_\mathfrak{m};\mu_{\mathfrak{g}2},\sigma^2_\mathfrak{m}+\sigma^2_{\mathfrak{g}2})\Phi(k_2) \tag{19}$$

with k_1,k_2

$$k_1=\frac{(\sigma_{\mathfrak{g}1}-\rho\sigma_{\mathfrak{g}2})\mu_{\mathfrak{c}1}-\sigma_{\mathfrak{g}1}\mu_{\mathfrak{g}2}+\rho\sigma_{\mathfrak{g}2}\mu_{\mathfrak{g}1}}{\left[\sigma^2_{\mathfrak{g}1}\sigma^2_{\mathfrak{g}2}(1-\rho^2)+(\sigma_{\mathfrak{g}1}-\rho\sigma_{\mathfrak{g}2})^2\sigma^2_{\mathfrak{c}1}\right]^{1/2}}\quad\text{and}\quad k_2=\frac{(\sigma_{\mathfrak{g}2}-\rho\sigma_{\mathfrak{g}1})\mu_{\mathfrak{c}2}-\sigma_{\mathfrak{g}2}\mu_{\mathfrak{g}1}+\rho\sigma_{\mathfrak{g}1}\mu_{\mathfrak{g}2}}{\left[\sigma^2_{\mathfrak{g}1}\sigma^2_{\mathfrak{g}2}(1-\rho^2)+(\sigma_{\mathfrak{g}2}-\rho\sigma_{\mathfrak{g}1})^2\sigma^2_{\mathfrak{c}2}\right]^{1/2}} \tag{20}$$

2.3.2 Inverse Problem

The conditional probability of $\boldsymbol{x}$ on m is

$$p(x_1,x_2|m)=\Theta(x_1-x_2)\delta(x_1-m)+\Theta(x_2-x_1)\delta(x_2-m) \tag{21}$$

where $\Theta(y)$ is Heaviside's step function. Therefore, the conditional of $\boldsymbol{x}$ on $\mathfrak{I}_\mathfrak{m}$ is

$$\begin{aligned}p(x_1,x_2|\mathfrak{I}_\mathfrak{m}) &= \int_{-\infty}^{\infty}p(m|x_1,x_2)p(m|\mathfrak{I}_\mathfrak{m})\,\mathrm{d}m \\ &= \Theta(x_1-x_2)\mathcal{N}(x_1;\mu_\mathfrak{m},\sigma^2_\mathfrak{m})+\Theta(x_2-x_1)\mathcal{N}(x_2;\mu_\mathfrak{m},\sigma^2_\mathfrak{m})\end{aligned} \tag{22}$$

which is a proper (i.e. normalizable) distribution, but becomes normalizable after multiplication with the prior:

$$\begin{aligned}p(\boldsymbol{x}|\mathfrak{I}_\mathfrak{c}) = Z^{-1}&\underbrace{\Theta(x_1-x_2)\mathcal{N}(x_1;\mu_\mathfrak{m},\sigma^2_\mathfrak{m})\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu}_\mathfrak{g},\boldsymbol{\Sigma}_\mathfrak{g})}_{\xi_1} \\ &+ Z^{-1}\underbrace{\Theta(x_2-x_1)\mathcal{N}(x_2;\mu_\mathfrak{m},\sigma^2_\mathfrak{m})\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu}_\mathfrak{g},\boldsymbol{\Sigma}_\mathfrak{g})}_{\xi_2}\end{aligned} \tag{23}$$

Figure 2 illustrates the shape of these functions by way of some concrete examples.

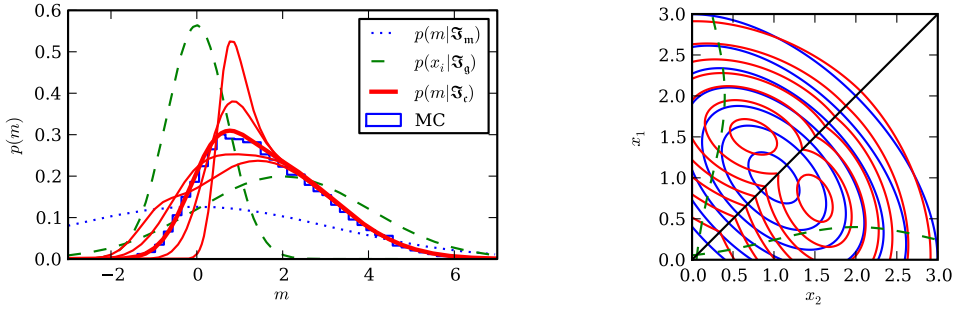

Figure 2: Illustrative plots for the analytical form of the forward and inverse posteriors. Left: Inference on m. Prior distribution and marginals on $\boldsymbol{x_i}$. Posteriors for five different values of ρ: -0.9 (most peaked), -0.5, 0.0 (thick line), 0.5 and 0.9 (broadest). As an experimental verification, a histogram of 20,000 samples from the posterior (generated by rejection sampling, with $\rho = 0$) is shown in blue. Right: Inference on the inverse problem: Prior with $\mu_{\mathfrak{g}} = (1, 1)^{\mathsf{T}}$, $\sigma_{\mathfrak{g}1} = \sigma_{\mathfrak{g}2} = 1$ and $\rho = -0.5$. Data on m with $\mu_{\mathfrak{m}} = 1$, $\sigma_{\mathfrak{m}} = 1$ gives the posterior in red. Note the bimodality arising in this particular case.

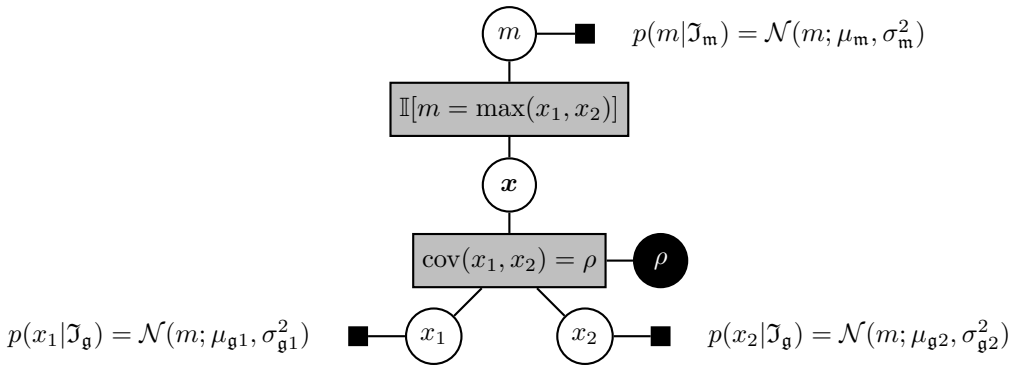

Figure 3: Factor graph representation of the functional relationships in the inference problems

2.4 Moment Matching

The analytical forms derived in the preceeding sections are clearly not members of the normal exponential family. If $\boldsymbol{x}$ has more than two elements, they also quickly take on complicated forms that are expensive to evaluate. If the application in question allows, it might thus be desirable to find Gaussian approximations to the posteriors. This section contains derivations for the first two moments of both posteriors. The Gaussian distributions q matching these moments minimize the Kullback-Leibler divergence $D_{\mathrm{KL}}(p||q) = \int p(y) \log(p(y)/q(y))\, \mathrm{d}y$ to the correct posterior p within the Gaussian family [see, e.g. Bishop, 2006, Section 10.7]).

2.4.1 Forward Problem

We will denote the mean and variance of the posterior of the max as $\mu_{m(12)}$ and $\sigma^2_{m(12)}$ for reasons that will become clear in Section 3. The corresponding integrals to solve are

$$\begin{aligned} \langle m \rangle \equiv \mu_{m(12)} &= \int_{-\infty}^{\infty} m p(m|\mathfrak{I}_{\mathfrak{c}})\, \mathrm{d}m = Z^{-1} \int m(\nu_1 + \nu_2)\, \mathrm{d}m \\ \langle m^2 \rangle - \langle m \rangle^2 \equiv \sigma^2_{m(12)} &= \int_{-\infty}^{\infty} m^2 p(m|\mathfrak{I}_{\mathfrak{c}})\, \mathrm{d}m \end{aligned} \tag{24}$$

[1]This is using the standard result that

$$\mathcal{N}(x; a_1, b_1^2)\mathcal{N}(x; a_2, b_2^2) = \mathcal{N}(a_1; a_2, b_1^2 + b_2^2)\mathcal{N}\left[x; \left(\frac{a_1}{b_1^2} + \frac{a_2}{b_2^2}\right)\left(\frac{1}{b_1^2} + \frac{1}{b_2^2}\right)^{-1}, \left(\frac{1}{b_1^2} + \frac{1}{b_2^2}\right)^{-1}\right] \tag{18}$$

which can be derived by completing the square, a simple proof that is omitted here

Comparison with Equation (16) shows that these two integrals are solved by Equation (11). The solutions are thus, after some algebra,

$$\begin{aligned}
\mu_{m(12)} &= w_1 \left[\mu_{\mathfrak{c}1} + \sigma_{\mathfrak{c}1}\frac{b_1}{a_1}\frac{\phi(k_1)}{\Phi(k_1)}\right] + w_2 \left[\mu_{\mathfrak{c}2} + \sigma_{\mathfrak{c}2}\frac{b_2}{a_2}\frac{\phi(k_2)}{\Phi(k_2)}\right] \\
\sigma^2_{m(12)} &= w_1 \left\{\left[\mu^2_{\mathfrak{c}1} + \sigma^2_{\mathfrak{c}1}\right] + \left[2\mu_{\mathfrak{c}1}\sigma_{\mathfrak{c}1}\frac{b_1}{a_1} - k_1\sigma^2_{\mathfrak{c}1}\frac{b_1^2}{a_1^2}\right]\frac{\phi(k_1)}{\Phi(k_1)}\right\} \\
&\quad + w_2 \left\{\left[\mu^2_{\mathfrak{c}2} + \sigma^2_{\mathfrak{c}2}\right] + \left[2\mu_{\mathfrak{c}2}\sigma_{\mathfrak{c}2}\frac{b_2}{a_2} - k_2\sigma^2_{\mathfrak{c}2}\frac{b_2^2}{a_2^2}\right]\frac{\phi(k_2)}{\Phi(k_2)}\right\} - \mu^2_{m(12)}
\end{aligned} \tag{25}$$

where w_i, a_i, b_i

$$w_1 = Z^{-1}\mathcal{N}(\mu_{\mathfrak{m}}; \mu_{\mathfrak{g}1}, \sigma^2_{\mathfrak{m}} + \sigma^2_1)\Phi(k_1) \qquad w_2 = Z^{-1}\mathcal{N}(\mu_{\mathfrak{m}}; \mu_{\mathfrak{g}2}, \sigma^2_{\mathfrak{m}} + \sigma^2_2)\Phi(k_2) \tag{26}$$

$$a_1 = \left[\sigma^2_{\mathfrak{g}1}\sigma^2_{\mathfrak{g}2}(1-\rho^2) + (\sigma_{\mathfrak{g}1} - \rho\sigma_{\mathfrak{g}2})^2\sigma^2_{\mathfrak{c}1}\right]^{1/2} \qquad a_2 = \left[\sigma^2_{\mathfrak{g}1}\sigma^2_{\mathfrak{g}2}(1-\rho^2) + (\sigma_{\mathfrak{g}2} - \rho\sigma_{\mathfrak{g}1})^2\sigma^2_{\mathfrak{c}2}\right]^{1/2} \tag{27}$$

$$b_1 = \sigma_{\mathfrak{c}1}(\sigma_{\mathfrak{g}1} - \rho\sigma_{\mathfrak{g}2}) \qquad b_2 = \sigma_{\mathfrak{c}2}(\sigma_{\mathfrak{g}2} - \rho\sigma_{\mathfrak{g}1}) \tag{28}$$

2.4.2 Inverse Problem

The derivation for the inverse problem is just slightly more involved. We are interested in the moments of the marginals $p(x_1|\mathfrak{I}_{\mathfrak{c}})$ and $p(x_2|\mathfrak{I}_{\mathfrak{c}})$, and will denote these means and variances with $\mu_{1(m2)}$, $\sigma^2_{1(m2)}$, et cetera. From Equation (23), we get

$$\begin{aligned}
\mu_{1(m2)} = \langle x_1\rangle_{\mathfrak{I}_{\mathfrak{c}}} &= \int_{-\infty}^{\infty} x_1 \int_{-\infty}^{x_1} \mathcal{N}(x_1; \mu_{\mathfrak{m}}, \sigma^2_{\mathfrak{m}})\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_{\mathfrak{g}}, \boldsymbol{\Sigma}_{\mathfrak{g}})\,\mathrm{d}x_2\,\mathrm{d}x_1 \\
&\quad + \int_{-\infty}^{\infty}\int_{-\infty}^{x_2} x_1\mathcal{N}(x_2; \mu_{\mathfrak{m}}, \sigma^2_{\mathfrak{m}})\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_{\mathfrak{g}}, \boldsymbol{\Sigma}_{\mathfrak{g}})\,\mathrm{d}x_1\,\mathrm{d}x_2
\end{aligned} \tag{29}$$

The first integral is in fact identical to the first term of $\mu_{m(12)}$. The second term, however, involves the first *incomplete* moment:

$$\begin{aligned}
&\int_{-\infty}^{\infty}\int_{-\infty}^{x_2} x_1\mathcal{N}(x_2; \mu_{\mathfrak{m}}, \sigma^2_{\mathfrak{m}})\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_{\mathfrak{g}}, \boldsymbol{\Sigma}_{\mathfrak{g}})\,\mathrm{d}x_1\,\mathrm{d}x_2 \\
&= \int_{-\infty}^{\infty}\mathcal{N}(x_2; \mu_{\mathfrak{m}}, \sigma^2_{\mathfrak{m}})\mathcal{N}(x_2; \mu_{\mathfrak{g}2}, \sigma^2_{\mathfrak{g}2})\int_{-\infty}^{x_2} x_1\mathcal{N}\left[x_1; \mu_{\mathfrak{g}1} + \rho\frac{\sigma_{\mathfrak{g}1}}{\sigma_{\mathfrak{g}2}}(x_2 - \mu_{\mathfrak{g}2}), \sigma^2_{\mathfrak{g}1}(1-\rho^2)\right]\mathrm{d}x_1\,\mathrm{d}x_2
\end{aligned} \tag{30}$$

The inner integral can be solved using the result given in Equation (10), leading to an expression solved by Equation (11). After a bit of algebra, we arrive at the final result

$$\begin{aligned}
\mu_{1(m2)} &= w_1 \left[\mu_{\mathfrak{c}1} + \sigma_{\mathfrak{c}1}\frac{b_1}{a_1}\frac{\phi(k_1)}{\Phi(k_1)}\right] + w_2 \left[\left(\mu_{\mathfrak{g}1} + \rho\frac{\sigma_{\mathfrak{g}1}}{\sigma_{\mathfrak{g}2}}(\mu_{\mathfrak{c}2} - \mu_{\mathfrak{g}2})\right) + \frac{A}{a_2}\frac{\phi(k_2)}{\Phi(k_2)}\right] \\
\sigma^2_{1(m2)} &= w_1 \left\{\left[\mu^2_{\mathfrak{c}1} + \sigma^2_{\mathfrak{c}1}\right] + \left[2\mu_{\mathfrak{c}1}\sigma_{\mathfrak{c}1}\frac{b_1}{a_1} - k_1\sigma^2_{\mathfrak{c}1}\frac{b_1^2}{a_1^2}\right]\frac{\phi(k_1)}{\Phi(k_1)}\right\} \\
&\quad + w_2 \left\{\sigma^2_{\mathfrak{g}1}\left[\left(\frac{\mu_{\mathfrak{g}1}}{\sigma_{\mathfrak{g}1}} + \rho\frac{(\mu_{\mathfrak{c}2} - \mu_{\mathfrak{g}2})}{\sigma_{\mathfrak{g}2}}\right)^2 + (1-\rho^2) + \rho^2\frac{\sigma^2_{\mathfrak{c}2}}{\sigma^2_{\mathfrak{g}2}}\right]\right. \\
&\quad \left. + \left[\frac{B}{h(1+\sigma^2_{\mathfrak{c}2}/h^2)^{1/2}} - \frac{C}{h^3(1+\sigma^2_{\mathfrak{c}2}/h^2)^{3/2}}\right]\frac{\phi(k_2)}{\Phi(k_2)}\right\} - \mu_1^2
\end{aligned} \tag{31}$$

where

$$\begin{aligned}
A &= \rho\sigma^2_{\mathfrak{c}2}\sigma_{\mathfrak{g}1}\left(1 - \rho\frac{\sigma_{\mathfrak{g}1}}{\sigma_{\mathfrak{g}2}}\right) - \sigma^2_{\mathfrak{g}1}\sigma_{\mathfrak{g}2}(1-\rho^2) \\
B &= 2\rho^2\frac{\sigma^2_{\mathfrak{g}1}}{\sigma^2_{\mathfrak{g}2}}\sigma^2_{\mathfrak{c}2}(\mu_{\mathfrak{c}2} - \mu_{\mathfrak{g}2}) + \rho\frac{\sigma_{\mathfrak{g}1}}{\sigma_{\mathfrak{g}2}}\left(2\sigma^2_{\mathfrak{c}2}\mu_{\mathfrak{g}1} + \mu_{\mathfrak{g}2}\frac{\sigma^2_{\mathfrak{g}1}\sigma_{\mathfrak{g}2}(1-\rho^2)}{\sigma_{\mathfrak{g}2} - \rho\sigma_{\mathfrak{g}1}}\right) - \mu_{\mathfrak{g}1}\sigma^2_{\mathfrak{g}1}(1-\rho^2)\frac{\sigma_{\mathfrak{g}2}}{\sigma_{\mathfrak{g}2} - \rho\sigma_{\mathfrak{g}1}} \\
C &= \rho^2\frac{\sigma^2_{\mathfrak{g}1}}{\sigma^2_{\mathfrak{g}2}}\sigma^4_{\mathfrak{c}2}(\mu_{\mathfrak{c}2} - f) + \sigma^2_{\mathfrak{g}1}(1-\rho^2)\left(1 + \rho\frac{\sigma_{\mathfrak{g}1}}{\sigma_{\mathfrak{g}2}}\right)\frac{\sigma_{\mathfrak{g}2}}{\sigma_{\mathfrak{g}2} - \rho\sigma_{\mathfrak{g}1}}(\mu_{\mathfrak{c}2}h^2 + f\sigma^2_{\mathfrak{c}2})
\end{aligned} \tag{32}$$

with

$$f = \frac{\sigma_{\mathfrak{g}2}\mu_{\mathfrak{g}1} - \rho\sigma_{\mathfrak{g}1}\mu_{\mathfrak{g}2}}{\sigma_{\mathfrak{g}2} - \rho\sigma_{\mathfrak{g}1}} \qquad \text{and} \qquad h = \frac{\sigma_{\mathfrak{g}1}\sigma_{\mathfrak{g}2}(1-\rho^2)^{1/2}}{\sigma_{\mathfrak{g}2} - \rho\sigma_{\mathfrak{g}1}} \tag{33}$$

The corresponding result for the posterior marginal on x_2 can be derived trivially from these results by exchanging the indices 1 and 2. Note that, as mentioned above, the first terms of these mixtures are shared with the posterior for m. Intuitively, this can be interpreted as follows: For the posterior on m, the first term (ν_1) corresponds to the statement that "if $x_1 > x_2$" (the probability of this is encoded by the cumulative density term in Equation (16)) "then m is distributed like x_1" (represented by the product of the probability density functions in (16)). This part of the relationship features in the inverse problem as well: If $x_1 > x_2$, then x_1 is distributed like m. The second term in the posterior marginal on x_1 corresponds to the statement that "if $x_1 < x_2$, then x_2 is distributed like m and x_1 is distributed such that its distribution fits with the updated marginal of x_2 given the correlation between x_1 and x_2 and the prior marginal on x_1.

2.4.3 Related Work

The moments of the likelihood of the max have been derived before by Clark [1961]. That is, for $\sigma_{\mathfrak{m}} \to \infty$, the posterior $p(m|\mathfrak{I}_{\mathfrak{c}})$ reported here simplifies to a result reported by Clark:

$$\begin{aligned}
\mu_{m(12)} &\to \Phi(k)\left[\mu_{\mathfrak{g}1} + \sigma_{\mathfrak{g}1}\frac{(\sigma_{\mathfrak{g}1} - \rho\sigma_{\mathfrak{g}2})}{a}\frac{\phi(k)}{\Phi(k)}\right] + \Phi(-k)\left[\mu_{\mathfrak{g}2} + \sigma_{\mathfrak{g}2}\frac{(\sigma_{\mathfrak{g}2} - \rho\sigma_{\mathfrak{g}1})}{a}\frac{\phi(-k)}{\Phi(-k)}\right] \\
\sigma^2_{m(12)} &\to \Phi(k)\left\{\left[\mu^2_{\mathfrak{g}1} + \sigma^2_{\mathfrak{g}1}\right] + \left[2\mu_{\mathfrak{g}1}\sigma_{\mathfrak{g}1}\frac{(\sigma_{\mathfrak{g}1} - \rho\sigma_{\mathfrak{g}2})}{a} - k\sigma^2_{\mathfrak{g}1}\frac{(\sigma_{\mathfrak{g}1} - \rho\sigma_{\mathfrak{g}2})^2}{a^2}\right]\frac{\phi(k)}{\Phi(k)}\right\} \\
&\quad + \Phi(-k)\left\{\left[\mu^2_{\mathfrak{g}2} + \sigma^2_{\mathfrak{g}2}\right] + \left[2\mu_{\mathfrak{g}2}\sigma_{\mathfrak{g}2}\frac{(\sigma_{\mathfrak{g}2} - \rho\sigma_{\mathfrak{g}1})}{a} + k\sigma^2_{\mathfrak{g}2}\frac{(\sigma_{\mathfrak{g}2} - \rho\sigma_{\mathfrak{g}1})^2}{a^2}\right]\frac{\phi(-k)}{\Phi(-k)}\right\} \\
&\quad - \mu^2_{m(12)} \\
\text{where} \quad a &= \sqrt{\sigma^2_{\mathfrak{g}1} + \sigma^2_{\mathfrak{g}2} - 2\rho\sigma_{\mathfrak{g}1}\sigma_{\mathfrak{g}2}} \qquad \text{and} \qquad k = \frac{\mu_{\mathfrak{g}1} - \mu_{\mathfrak{g}2}}{a}
\end{aligned} \tag{34}$$

As expected, the posterior of the inverse problem simply becomes equal to the prior in this case. From Equation (31) we find

$$\begin{aligned}
\mu_{1(m2)} &\to \Phi(k)\mu_1 + \sigma_1\frac{\sigma_1 - \rho\sigma_2}{a}\phi(k) + \Phi(-k)\mu_1 - \sigma_1\frac{\sigma_1 - \rho\sigma_2}{a}\phi(-k) \\
&= \Phi(k)\mu_1 + \sigma_1\frac{\sigma_1 - \rho\sigma_2}{a}\phi(k) + (1 - \Phi(k))\mu_1 - \sigma_1\frac{\sigma_1 - \rho\sigma_2}{a}\phi(k) \\
&= \mu_1
\end{aligned} \tag{35}$$

and similarly for the variance.

The max-factor is also part of the Infer.NET software package [Minka and Winn, 2008] (to my knowledge, the derivations for this code have not been published yet). However, their implementation can only handle two independent Gaussian inputs (Section 3 introduces the max over a finite set of correlated variables). So their implementation corresponds to the case of $\rho = 0$, which leads to the following simplifications, presented here for reference:

$$k_1 = \frac{\mu_{\mathfrak{c}1} - \mu_{\mathfrak{g}2}}{(\sigma_{\mathfrak{g}1} + \sigma_{\mathfrak{c}2})^{1/2}} \qquad a_1 = \sigma_{\mathfrak{g}1}(\sigma_{\mathfrak{g}1} + \sigma_{\mathfrak{c}2})^{1/2} \qquad b_1 = \sigma_{\mathfrak{c}1}\sigma_{\mathfrak{g}1} \tag{36}$$

$$A = \sigma^2_{\mathfrak{g}1}\sigma_{\mathfrak{g}2} \qquad B = -\mu_{\mathfrak{g}1}\sigma^2_{\mathfrak{g}1} \qquad C = \sigma^2_{\mathfrak{g}1}(\mu_{\mathfrak{c}2}\sigma^2_{\mathfrak{g}1} + \mu_{\mathfrak{g}1}\sigma^2_{\mathfrak{c}2}) \tag{37}$$

$$f = \mu_{\mathfrak{g}1} \qquad h = \sigma_{\mathfrak{g}1} \tag{38}$$

Figure 4 shows some of these approximations. The parameter settings used in this figure represent a worst case (e.g., the posterior over $\boldsymbol{x}$ is rarely so strongly bimodal.)

3 The Maximum of a Finite Set

3.1 Analytic Form

Extending the analysis of Section 2.3, we can write the posterior over the max m of a finite set $\{x_i\}_{i=1,\ldots,N}$ of variables, distributed according to an N-dimensional version of Equation

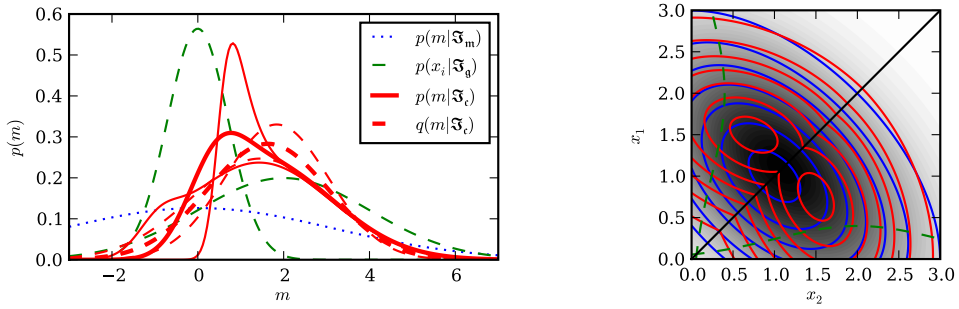

Figure 4: Illustrative plots for the Gaussian approximations to the posteriors. Same beliefs in $\mathfrak{I}_\mathfrak{c}$ as in Figure 2. Left: For the sake of readability, only the cases $\rho = -0.9$ (broadest), $\rho = 0$ and $\rho = 0.9$ are plotted here. In red, dashed lines the corresponding three Gaussian approximations. Note the varying quality of fit. Right: Gaussian approximation (with $\boldsymbol{\mu}_{1(m2)} = 1.06$ and $\sigma^2_{1(m2)} = 0.94$) indicated by shaded area.

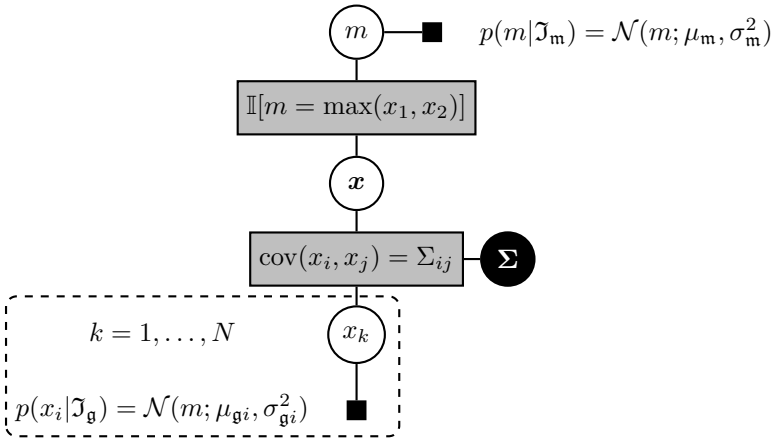

Figure 5: Factor graph representation of the inference problem on a finite set. The dashed "plate" represents N copies of generating variable nodes.

(1), with a new normalization constant Z_N, as

$$\begin{aligned} p(m|\mathfrak{I}_\mathfrak{c}) &= Z_N p(m|\mathfrak{I}_\mathfrak{m}) \int p(\boldsymbol{x}|m) p(\boldsymbol{x}|\mathfrak{I}_\mathfrak{g}) \,\mathrm{d}\boldsymbol{x} \\ &= Z_N \mathcal{N}(m; \mu_\mathfrak{m}, \sigma_\mathfrak{m}) \left[\sum_{i=1}^{N} \int_{-\infty}^{\infty} \delta(m - x_i) p(x_i|\mathfrak{I}_\mathfrak{g}) \int \cdots \int_{-\infty}^{x_i} p(\{x_j\}_{j\neq i}|x_i, \mathfrak{I}_\mathfrak{g}) \prod_{j\neq i} \mathrm{d}x_j \,\mathrm{d}x_i \right] \\ &= Z_N \sum_i \left[\mathcal{N}(\mu_\mathfrak{m}; \mu_{\mathfrak{g}i}, \sigma^2_\mathfrak{m} + \sigma^2_{\mathfrak{g}i}) \mathcal{N}(m; \mu_{\mathfrak{c}i}, \sigma^2_{\mathfrak{c}i}) \int \cdots \int_{-\infty}^{x_i} \mathcal{N}(\boldsymbol{x}_{\setminus i}; \boldsymbol{\mu}_{\mathfrak{g}\setminus i}(x_i), \boldsymbol{\Sigma}_{\mathfrak{g}\setminus i}) \,\mathrm{d}\boldsymbol{x}_{\setminus i} \right] \end{aligned} \tag{39}$$

where $\boldsymbol{x}_{\setminus i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_N)$. The conditional mean is [see e.g. Bishop, 2006, Section 2.3.2]

$$\left(\boldsymbol{\mu}_{\setminus i}(x_i)\right)_j = \mu_{\mathfrak{g}j} + \Sigma_{\mathfrak{g}ji}\Sigma^{-1}_{\mathfrak{g}ii}(x_i - \mu_{\mathfrak{g}i}) = \mu_{\mathfrak{g}j} + \rho_{ij}\frac{\sigma_{\mathfrak{g}j}}{\sigma_{\mathfrak{g}i}}(x_i - \mu_{\mathfrak{g}i}) \tag{40}$$

with the linear coefficient of correlation $\rho_{ij} = \Sigma_{\mathfrak{g}ij}/(\sigma_{\mathfrak{g}i}\sigma_{\mathfrak{g}j})$. The conditional covariance matrix is the Schur complement of $\Sigma_{\mathfrak{g}ii} = \sigma^2_{\mathfrak{g}i}$ in $\boldsymbol{\Sigma}_\mathfrak{g}$:

$$\Sigma_{\mathfrak{g}\setminus i,kj} = \Sigma_{\mathfrak{g}kj} - \Sigma_{\mathfrak{g}ki}\sigma^{-2}_{\mathfrak{g}i}\Sigma_{\mathfrak{g}ij} \tag{41}$$

In principle, it would be possible to follow the path laid out in the previous sections to calculate the first two moments of this distribution. However, while the univariate Gaussian CDF (essentially an evaluation of the error function) has computational cost comparable to evaluating an exponential function, computationally efficient ways of calculating a multivariate

Gaussian CDF are not generally available. So this approximation would need to involve an undesirable numerical integration.

3.2 A Heuristic Approximation

Another, cheaper option is to use an iterative procedure initially proposed by Clark [1961]. The idea is to start out with the approximation for only two of the generating variables. W.l.o.g., let these be x_1 and x_2, resulting in $m_{(12)} = \max(x_1, x_2)$. Next, estimate $m_{(123)} = \max(x_3, m_{(12)})$ and so on up to $m_{(1\dots N)}$. For the intermediate maxima, the likelihoods presented in Equation (34) suffice, and the prior is included in the last step (using Equation (25)) to gain an approximate posterior over the maximum of the whole set. Of course, this necessitates an analytic expression for the correlation coefficient $\rho_{i(1\dots i-1)}$ between the i-th variable and the max over the preceding variables. This was derived by Clark. Adopted to the notation used here and made more explicit, his result is

$$\rho_{3(12)} = \sigma_{(12)}^{-1} \left(\sigma_1 \rho_{31} \Phi(k_{(12)}) + \sigma_2 \rho_{32} \Phi(-k_{(12)})\right) \tag{42}$$

where $\rho_{ij} = \Sigma_{ij}/\sigma_i\sigma_j$, the index $\mathfrak{g}$ has been dropped for simplicity and $k_{(12)} = (\mu_1 - \mu_2)/\sqrt{\sigma_1^2 + \sigma_2^2}$ is the simplified version of k_1 arising from Equation (20) under $\sigma_\mathfrak{m} \to \infty$. Using this, we can build a recursive algorithm to calculate $\rho_{i(1\dots j)}$ with $j < i$ as

$$\rho_{i(1\dots j)} = \sigma_{(1\dots j)}^{-1} \cdot \begin{cases} \sigma_j \rho_{ij} & \text{if } j = 1 \\ \Phi(-k_{(1\dots j)})\sigma_j \rho_{ij} + \Phi(k_{(1\dots j)})\rho_{i(1\dots j-1)} & \text{else} \end{cases} \tag{43}$$

this necessitates a list $\boldsymbol{k}_{[j]} = (k_{(12)}, \dots k_{(1\dots i-1)})$ which is available at the necessary point in time from the calculation of previous maxima over the preceding parts of the set. Note that calculating $\rho_{i(1\dots i-1)}$ involves $i-1$ recursive function calls, so building the full approximation over the max of N variables is of complexity $\mathcal{O}(N^2)$, as might be expected (although there are only $(N-1)$ uses of the results in Equation (25)). If all correlation coefficients are the same, $\rho_{ij} = \rho \; \forall ij$, then the recursive evaluations can be re-used in consecutive evaluations and the complexity drops to $\mathcal{O}(N)$.

3.2.1 Inverse Problem

The same iterative scheme can be used to provide an approximation for the inverse problem's posterior. First, the list $\boldsymbol{k}_{[j]}$ is build as in the preceding section. Then, approximations to the posterior marginals are build iteratively, starting with $q(x_N|\mathfrak{I}_\mathfrak{c})$, ending with $q(x_2|\mathfrak{I}_\mathfrak{c})$ and $q(x_1|\mathfrak{I}_\mathfrak{c})$. At each intermediate step, we use the EP approximation [Minka, 2001]: To get $q(x_i|\mathfrak{I}_\mathfrak{c})$, use $q(m_{(1\dots i)}|\mathfrak{I}_\mathfrak{m}) = q(m_{(1\dots i)}|\mathfrak{I}_\mathfrak{c})/q(m_{(1\dots i)}|\mathfrak{I}_\mathfrak{g})$ as an approximation to the prior over the subset max, and $q(m_{(1\dots i-1)}|\mathfrak{I}_\mathfrak{g})$ as the approximation on the max over the subset up to x_{i-1}.

4 Discussion of the Approximation's Quality

Figure 6 gives some intuition on the quality of the approximation. For the purpose of this comparison, uncorrelated Gaussians were used because this allows the analytic evaluation of the true posterior (the CDF factorises into individual one-dimensional CDFs). The fit is reasonably good if the beliefs over the x_i are either very similar (Figure 6 top right), or if the beliefs are "separated", in the sense that one of the x_i provides a dominant contribution to the overall mixture (top left). The fit becomes bad when the mixture has many modes (bottom left) or a strong asymmetry (bottom right). The corresponding worst case distributions shown here were generated by setting $\mu_{\mathfrak{g}i} = a + b^{-i}$ and $\sigma_{\mathfrak{g}i}^2 = b^{-i}$ (left, $a = -1, b = 16$) or $\mu_{\mathfrak{g}i} = ci$ and $\sigma_{\mathfrak{g}i}^2 = i^d + 1$ (right, $c = -1, d = 16$). More quantitatively, consider Equation (34) or Equation (16), the case of the max of only two Gaussians. The two cases of good fit described above correspond to

1. one mixture component dominating the mixture

$$|k_{12}| = \frac{|\mu_{\mathfrak{g}1} - \mu_{\mathfrak{g}2}|}{\sqrt{\sigma_{\mathfrak{g}1}^2 + \sigma_{\mathfrak{g}2}^2 - 2\rho\sigma_{\mathfrak{g}1}\sigma_{\mathfrak{g}2}}} \gg 0 \tag{44}$$

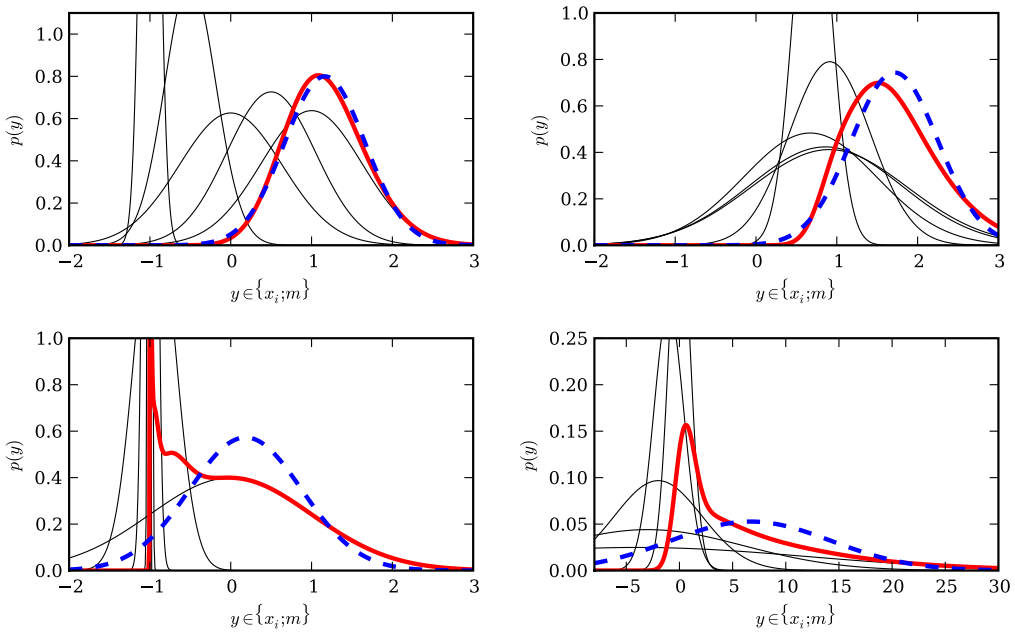

Figure 6: Quality and failure modes of the EP approximation. Max of five uncorrelated Gaussians. Top row: examples of good fits. Left: well separated beliefs. Right: similar beliefs. Bottom row: worst case examples. Left: high certainty contributions within the center. Right: high uncertainty in one tail. In all plots, beliefs over the x_i as slim black lines. True posterior over m in thick red, approximation in thick dashed blue. For simplicity, $p(m|\mathfrak{I}_\mathfrak{m})$ was set to an uninformative value. See text for details.

The likelihood then has one clearly dominating Gaussian component and the fit is good. In this case, the inverse problem is also a good fit, as each of the generating variables x_1, x_2 has one dominating component in its posterior.

2. the two mixture components being almost identical:

$$\mu_{\mathfrak{g}1} \approx \mu_{\mathfrak{g}2} \qquad \text{and} \qquad \sigma_{\mathfrak{g}1} \approx \sigma_{\mathfrak{g}2} \tag{45}$$

the likelihood then consists of two roughly identical Gaussian components with roughly the same weights, and is therefore roughly Gaussian. However, the approximation is *bad* for the inverse problem here, as the true posterior marginals become bimodal (c.f. Figure 2, right). This effect is particularly pronounced if the mean of the prior and the likelihood differ significantly.

These observations suggest a potential increase in the quality of the approximation to be gained from calculating all $N(N-1)$ weight-generators k_{ij} as defined in Equation (44) and iteratively choosing the pair ij with maximal k_{ij}. However, this re-ordering has to be updated after each incremental two-component max operation, involving a re-calculation of up to N correlation coefficients. It thus raises the complexity of calculating the approximation for the overall max from $\mathcal{O}(N^2)$ to $\mathcal{O}(N^3)$. Initial experiments suggest that the potential gain in fit is almost always negligible.

5 Conclusion

This technical report derived the first two moments of the posterior over the maximum of a pair of Gaussian variables, and over the posterior over the two generating variables. These moments can be used for approximate Inference on their own, or as part of a larger graphical model using Expectation Propagation. I have also shown how to extend the usefulness of these approximations to finite sets of Gaussian variables using a heuristic iterative approximation. The quality of the approximation depends on the location and precision of the belief over the generating variables relative to each other, but is always good enough to provide a meaningful point estimate and error measure. It is sufficiently robust to deal with inconsistent belief assignments and large numbers of generating variables (see Figure 7).

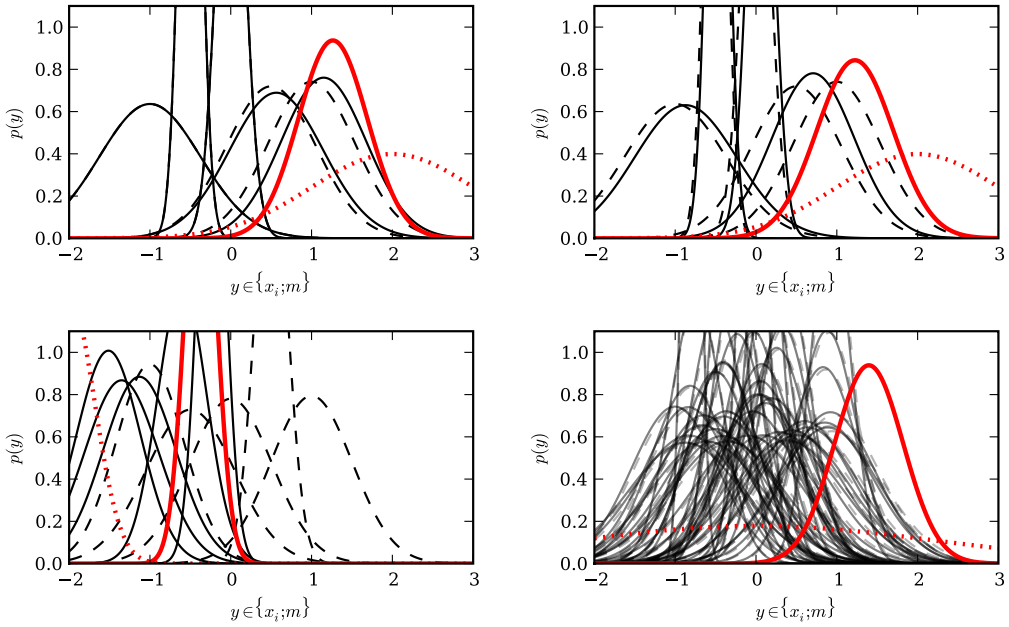

Figure 7: Illustrative examples for the use of the approximation in EP message passing. $p(x_i|\mathfrak{I}_{\mathfrak{g}})$ in black dashed lines. $p(m|\mathfrak{I}_{\mathfrak{m}})$ in red dotted. Marginals after EP message passing as corresponding solid lines. Top left: Max over 5 uncorrelated variables. Only the two variables contributing significantly to the max change their beliefs. Top right: same as previous, but with $\rho_{ij} = 0.9$ for all ij. The change in belief over the dominating x_i now also effects the other beliefs, as expected. Bottom left: The approximation is well-behaved under inconsistent beliefs. $p(m|\mathfrak{I}_{\mathfrak{m}})$ was set inconsistently low relative to the beliefs on the x_i (all $\rho_{ij} = 0.2$). Note that the belief over the largest x_i extends beyond the belief over m as a result of the moment-matching. Bottom right: The approximation is stable for large values of N. Maximum over 50 correlated normals, all ρ_{ij} were set to 0.5.

References

R.K. Ahuja and J.B. Orlin. Inverse optimization. *Operations Research*, pages 771–783, 2001.

C.M. Bishop. *Pattern recognition and machine learning*. Springer New York., 2006.

D. Burton and P.L. Toint. On an instance of the inverse shortest paths problem. *Mathematical Programming*, 53(1):45–61, 1992.

Charles E. Clark. The greatest of a finite set of random variables. *Operations Research*, 9 (2):145–162, 1961.

Richard Dearden, Nir Friedman, and Stuart Russell. Bayesian Q-learning. In *In AAAI/IAAI*, pages 761–768. AAAI Press, 1998.

C. Heuberger. Inverse combinatorial optimization: A survey on problems, methods, and results. *Journal of Combinatorial Optimization*, 8(3):329–361, 2004.

Thomas Minka and John Winn. *infer.NET software package*. Microsoft Research Ltd., 2008.

T.P. Minka. Expectation Propagation for approximate Bayesian inference. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence table of contents*, pages 362–369. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, 2001.

G. Neumann-Denzau and J. Behrens. Inversion of seismic data using tomographical reconstruction techniques for investigations of laterally inhomogeneous media. *Geophysical Journal International*, 79(1):305–315, 1984.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.