Minimising the Kullback–Leibler Divergence

Ralf Herbrich

7th August 2005

Abstract

In this note we show that minimising the Kullback–Leibler divergence over a family in the class of exponential distributions is achieved by matching the *expected natural statistic*. We will also give an explicit update formula for distributions with only one likelihood term.

1 Notation

We use $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ to denote a Gaussian density at $\mathbf{x}$ with a mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$,

$$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) := (2\pi)^{-\frac{n}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})\right) . \tag{1.1}$$

When dealing one dimensional Gaussians the vectors and matrices are replaced by scalars. If p is a density over $\mathbf{x}$, we will write $\langle g(\mathbf{x})\rangle_{p(\mathbf{x})}$ as a shorthand notation for the expectation of g over $\mathbf{x}$, $\int g(\mathbf{x})\, p(\mathbf{x})\, d\mathbf{x}$. Finally, the *Kullback–Leibler* divergence between two densities p and q is defined by

$$\mathrm{KL}(p||q) := \left\langle \log\left(\frac{p(\mathbf{x})}{q(\mathbf{x})}\right)\right\rangle_{p(\mathbf{x})} . \tag{1.2}$$

2 Minimising in the Exponential Family

A set of distributions over $\mathbb{R}^N$ is in the exponential family if its densities can be written as

$$p_{\boldsymbol{\theta}}(\mathbf{x}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\left(\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x})\right) ,$$

where $\boldsymbol{\phi}(\mathbf{x})$ is known as the *natural statistic* of $\mathbf{x}$ and $Z(\boldsymbol{\theta}) := \int \exp(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}))d\mathbf{x}$ ensures normalisation. The exponential family includes many known families of distributions including the Gaussian distribution. For example, in the Gaussian case, the natural statistic $\boldsymbol{\phi}(\mathbf{x})$ is simply the vector of all first and second moments, $\boldsymbol{\phi}(\mathbf{x}) = (x_1, \ldots, x_N, x_1^2, x_1x_2, \ldots, x_Nx_{N-1}, x_N^2)$. Note that the expected natural statistic of $p_{\boldsymbol{\theta}}(\mathbf{x})$ is given in terms of the gradient of $\log(Z(\boldsymbol{\theta}))$ w.r.t. $\boldsymbol{\theta}$, that is,

$$\nabla_{\boldsymbol{\theta}} \log(Z(\boldsymbol{\theta})) = \frac{\int \left[\nabla_{\boldsymbol{\theta}} \exp\left(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})\right)\right] d\mathbf{x}}{Z(\boldsymbol{\theta})} = \langle \boldsymbol{\phi}(\mathbf{x})\rangle_{p_{\boldsymbol{\theta}}(\mathbf{x})} . \tag{2.1}$$

Theorem 1. *For any distribution p, the distribution $p_{\boldsymbol{\theta}^*}$ which minimises the Kullback-Leibler divergence, KL $(p||p_{\boldsymbol{\theta}^*})$, over the exponential family with natural statistic $\boldsymbol{\phi}$ is implicitly given by*

$$\langle \boldsymbol{\phi}(\mathbf{x})\rangle_{p_{\boldsymbol{\theta}^*}(\mathbf{x})} = \langle \boldsymbol{\phi}(\mathbf{x})\rangle_{p(\mathbf{x})} . \tag{2.2}$$

Proof. Let us recall the Kullback-Leibler divergence from (1.2) and consider it as a function f of the parameters $\boldsymbol{\theta}$,

$$\begin{aligned} f(\boldsymbol{\theta}) &= \mathrm{KL}(p\|p_{\boldsymbol{\theta}}) = \left\langle \log\left(\frac{p(\mathbf{x})}{p_{\boldsymbol{\theta}}(\mathbf{x})}\right)\right\rangle_{p(\mathbf{x})} \\ &= \langle \log(p(\mathbf{x}))\rangle_{p(\mathbf{x})} + \langle \log(Z(\boldsymbol{\theta}))\rangle_{p(\mathbf{x})} - \left\langle \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})\right\rangle_{p(\mathbf{x})} \\ &= \langle \log(p(\mathbf{x}))\rangle_{p(\mathbf{x})} + \log(Z(\boldsymbol{\theta})) - \boldsymbol{\theta}^{\mathrm{T}}\langle \boldsymbol{\phi}(\mathbf{x})\rangle_{p(\mathbf{x})} .\end{aligned}$$

Recall that a necessary condition for the minimum $\boldsymbol{\theta}^*$ is $\nabla_{\boldsymbol{\theta}} f\left(\boldsymbol{\theta}^*\right) = \mathbf{0}$. From (2.1) we have

$$\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) = \langle \boldsymbol{\phi}(\mathbf{x})\rangle_{p_{\boldsymbol{\theta}}(\mathbf{x})} - \langle \boldsymbol{\phi}(\mathbf{x})\rangle_{p(\mathbf{x})} .$$

It remains to show that $\boldsymbol{\theta}^*$ such that $\langle\boldsymbol{\phi}(\mathbf{x})\rangle_{p_{\boldsymbol{\theta}^*}(\mathbf{x})} = \langle\boldsymbol{\phi}(\mathbf{x})\rangle_{p(\mathbf{x})}$ is a minimum. To this end, consider the matrix of second derivatives,

$$\begin{aligned} [\nabla\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta})]_{i,j} &= \frac{\partial^2 \log(Z(\boldsymbol{\theta}))}{\partial\theta_i\partial\theta_j} = \frac{\partial}{\partial\theta_j}\frac{\int \phi_i(\mathbf{x})\exp\left(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})\right)d\mathbf{x}}{Z(\boldsymbol{\theta})} \\ &= \left\langle \phi_i(\mathbf{x})\,\phi_j(\mathbf{x})\right\rangle_{p_{\boldsymbol{\theta}}(\mathbf{x})} - \langle\phi_i(\mathbf{x})\rangle_{p_{\boldsymbol{\theta}}(\mathbf{x})}\left\langle\phi_j(\mathbf{x})\right\rangle_{p_{\boldsymbol{\theta}}(\mathbf{x})} .\end{aligned}$$

At the solution $\boldsymbol{\theta}^*$, this is the covariance matrix of the natural statistic $\boldsymbol{\phi}(\mathbf{x})$ over the distribution $p_{\boldsymbol{\theta}^*}$. By definition, this is positive semi-definite matrix (in fact, for every distribution $p_{\boldsymbol{\theta}}$) and thus we have proven the theorem. □

Remark. In the case of the Gaussian family, $\{\mathcal{N}(\cdot;\boldsymbol{\mu},\boldsymbol{\Sigma})\}$, Theorem 1 reduces to matching the mean and covariance (which are related in a one-to-one way to the first and second moments),

$$\boldsymbol{\mu}^* = \langle\mathbf{x}\rangle_{p(\mathbf{x})} , \tag{2.3}$$

$$\boldsymbol{\Sigma}^* = \left\langle \mathbf{x}\mathbf{x}^{\mathrm{T}}\right\rangle_{p(\mathbf{x})} - \langle\mathbf{x}\rangle_{p(\mathbf{x})}\langle\mathbf{x}\rangle_{p(\mathbf{x})}^{\mathrm{T}} . \tag{2.4}$$

3 Matching the Bayesian Posterior

We will now derive an explicit update formula for matching the expected natural statistic if $p(\mathbf{x})$ has the simple form

$$p(\mathbf{x}) = \frac{1}{\tilde{Z}(\boldsymbol{\theta})}\cdot t(\mathbf{x})\,p_{\boldsymbol{\theta}}(\mathbf{x}) ,$$

where $\tilde{Z}(\boldsymbol{\theta}) := \int t(\mathbf{x})p_{\boldsymbol{\theta}}(\mathbf{x})d\mathbf{x}$ ensures normalisation[1]. In fact, similar to (2.1), the expected natural statistic under $p(\mathbf{x})$ can again be expressed solely in terms of the gradient of $\tilde{Z}(\boldsymbol{\theta})$ w.r.t. $\boldsymbol{\theta}$. In order to see this, note that

$$\begin{aligned} \nabla_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\mathbf{x}) &= \left[\nabla_{\boldsymbol{\theta}}\frac{1}{Z(\boldsymbol{\theta})}\right]\exp\left(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})\right) + \frac{1}{Z(\boldsymbol{\theta})}\left[\nabla_{\boldsymbol{\theta}}\exp\left(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})\right)\right] \\ &= -\frac{[\nabla_{\boldsymbol{\theta}} Z(\boldsymbol{\theta})]}{Z(\boldsymbol{\theta})}p_{\boldsymbol{\theta}}(\mathbf{x}) + \boldsymbol{\phi}(\mathbf{x})\,p_{\boldsymbol{\theta}}(\mathbf{x}) \\ &= -\langle\boldsymbol{\phi}(\mathbf{x})\rangle_{p_{\boldsymbol{\theta}}(\mathbf{x})}\cdot p_{\boldsymbol{\theta}}(\mathbf{x}) + \boldsymbol{\phi}(\mathbf{x})\,p_{\boldsymbol{\theta}}(\mathbf{x}) .\end{aligned}$$

Multiplying both sides by $\tilde{Z}^{-1}(\boldsymbol{\theta})t(\mathbf{x})$, integrating over $\mathbf{x}$ and re-arranging terms we get

$$\tilde{Z}^{-1}(\boldsymbol{\theta})\,\nabla_{\boldsymbol{\theta}}\tilde{Z}(\boldsymbol{\theta}) = -\langle\boldsymbol{\phi}(\mathbf{x})\rangle_{p_{\boldsymbol{\theta}}(\mathbf{x})} + \langle\boldsymbol{\phi}(\mathbf{x})\rangle_{p(\mathbf{x})}$$

$$\langle\boldsymbol{\phi}(\mathbf{x})\rangle_{p(\mathbf{x})} = \nabla_{\boldsymbol{\theta}}\log\left(\tilde{Z}(\boldsymbol{\theta})\right) + \langle\boldsymbol{\phi}(\mathbf{x})\rangle_{p_{\boldsymbol{\theta}}(\mathbf{x})} . \tag{3.1}$$

[1] Please note that the normalisation constant $\tilde{Z}(\boldsymbol{\theta})$ should not be confused with the normalisation constant $Z(\boldsymbol{\theta})$.

Finally, using Theorem 1 and (2.1) we obtain

$$\nabla_{\boldsymbol{\theta}} \log\left(Z\left(\boldsymbol{\theta}^*\right)\right) = \nabla_{\boldsymbol{\theta}} \log\left(\tilde{Z}\left(\boldsymbol{\theta}\right)\right) + \nabla_{\boldsymbol{\theta}} \log\left(Z\left(\boldsymbol{\theta}\right)\right) .$$

All that is required to solve the above equation for a given exponential family is to know the analytical solution of the gradient equation of $\log(Z(\boldsymbol{\theta}))$ and $\log(\tilde{Z}(\boldsymbol{\theta}))$. These two equations only depend on the particular natural statistic function $\boldsymbol{\phi}$ and the function t. This is applicable, for example, for Gamma densities.

However, some exponential families are usually not parameterised in terms of $\boldsymbol{\theta}$ but rather in terms of $\boldsymbol{\tau}(\boldsymbol{\theta}) := \langle\boldsymbol{\phi}(\mathbf{x})\rangle_{p_{\boldsymbol{\theta}}(\mathbf{x})}$—a parameterisation also known as the *moment representation*. This representation has particular advantages when minimising the KL divergence as Theorem 1 directly specifies the update equation for the parameters. In this case, (3.1) can still be used together with the chain rule of differentiation to obtain the update equation for a particular class of exponential densities if the mapping to $\boldsymbol{\tau} \mapsto \boldsymbol{\theta}$ is easy to differentiate. We can also follow the above argument simply in the new parameterisation. In the next section we give a detailed derivation for the Gaussian family (which is represented in terms of its moments).

4 Matching the Bayesian Posterior in the Gaussian Family

We consider a family of Gaussians parameterised in terms of its mean, $\boldsymbol{\mu}$, and covariance, $\boldsymbol{\Sigma}$,

$$q\left(\mathbf{x}\right) := q\left(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) := \mathcal{N}\left(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) .$$

Our ability to compute (2.3) and (2.4) when $p(\mathbf{x}) \propto t(\mathbf{x})q\left(\mathbf{x}\right)$ depends only on the tractability of the normalisation constant,

$$\tilde{Z} := \tilde{Z}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) := \int t\left(\mathbf{x}\right) q\left(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) d\mathbf{x} .$$

Matching the Mean We will consider the mean of $\mathbf{x}$ under $t(\mathbf{x})q(\mathbf{x})$. First note that

$$\nabla_{\boldsymbol{\mu}} q\left(\mathbf{x}\right) = \boldsymbol{\Sigma}^{-1}\left(\mathbf{x} - \boldsymbol{\mu}\right) q\left(\mathbf{x}\right) ,$$

which can be re-expressed in terms of $\mathbf{x}q\left(\mathbf{x}\right)$,

$$\mathbf{x}q\left(\mathbf{x}\right) = \boldsymbol{\mu} q\left(\mathbf{x}\right) + \boldsymbol{\Sigma}\nabla_{\boldsymbol{\mu}} q\left(\mathbf{x}\right) .$$

Now multiplying both sides by $\tilde{Z}^{-1}t(\mathbf{x})$, integrating over $\mathbf{x}$, and exploiting the linearity of the gradient operator gives

$$\begin{aligned}
\langle\mathbf{x}\rangle_{p(\mathbf{x})} &= \boldsymbol{\mu} + \tilde{Z}^{-1} \cdot \boldsymbol{\Sigma}\left[\nabla_{\boldsymbol{\mu}} \int t\left(\mathbf{x}\right) q\left(\mathbf{x}\right) d\mathbf{x}\right] \\
&= \boldsymbol{\mu} + \tilde{Z}^{-1}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \cdot \boldsymbol{\Sigma}\nabla_{\boldsymbol{\mu}} \tilde{Z}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \\
&= \boldsymbol{\mu} + \boldsymbol{\Sigma}\nabla_{\boldsymbol{\mu}} \log\left(\tilde{Z}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)\right) \\
&= \boldsymbol{\mu} + \boldsymbol{\Sigma}\mathbf{g} ,
\end{aligned} \tag{4.1}$$

where we have defined $\mathbf{g} := \nabla_{\boldsymbol{\mu}} \log(\tilde{Z}(\boldsymbol{\mu}, \boldsymbol{\Sigma}))$.

The Second Moment Matrix Once again we take gradients[2] of $q(\mathbf{x})$, but this time with respect to the covariance matrix $\boldsymbol{\Sigma}$,

$$\nabla_{\boldsymbol{\Sigma}} q\left(\mathbf{x}\right) = \frac{1}{2}\left(-\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-1}\left(\mathbf{x} - \boldsymbol{\mu}\right)\left(\mathbf{x} - \boldsymbol{\mu}\right)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}\right) q\left(\mathbf{x}\right) ,$$

[2]It helps to remember that $\nabla_{\boldsymbol{\Sigma}} \log(q(\mathbf{x})) = (q(\mathbf{x}))^{-\mathbf{1}} \cdot \nabla_{\boldsymbol{\Sigma}} q(\mathbf{x}))$.

which can be re-arranged, as we did before, in order to obtain

$$\mathbf{x}\mathbf{x}^{\mathrm{T}} q\left(\mathbf{x}\right) = 2\boldsymbol{\Sigma}\left[\nabla_{\boldsymbol{\Sigma}} q\left(\mathbf{x}\right)\right]\boldsymbol{\Sigma} + \left(\boldsymbol{\Sigma} + \mathbf{x}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\mu}\mathbf{x}^{\mathrm{T}} - \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}\right) q\left(\mathbf{x}\right) \ .$$

Multiplying both sides by $\tilde{Z}^{-1} t\left(\mathbf{x}\right)$, integrating over $\mathbf{x}$ and exploiting the linearity of the gradient operator gives

$$\begin{aligned}
\left\langle \mathbf{x}\mathbf{x}^{\mathrm{T}} \right\rangle_{p(\mathbf{x})} &= \boldsymbol{\Sigma} + 2\boldsymbol{\Sigma}\left(\tilde{Z}^{-1}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)\nabla_{\boldsymbol{\Sigma}}\tilde{Z}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)\right)\boldsymbol{\Sigma} + \left\langle \mathbf{x}\right\rangle_{p(\mathbf{x})}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\mu}\left\langle \mathbf{x}\right\rangle_{p(\mathbf{x})}^{\mathrm{T}} - \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} \\
&= \boldsymbol{\Sigma} + 2\boldsymbol{\Sigma}\left(\nabla_{\boldsymbol{\Sigma}}\log\left(\tilde{Z}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)\right)\right)\boldsymbol{\Sigma} + \left\langle \mathbf{x}\right\rangle_{p(\mathbf{x})}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\mu}\left\langle \mathbf{x}\right\rangle_{p(\mathbf{x})}^{\mathrm{T}} - \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} \\
&= \boldsymbol{\Sigma} + 2\boldsymbol{\Sigma}\mathbf{G}\boldsymbol{\Sigma} + \left\langle \mathbf{x}\right\rangle_{p(\mathbf{x})}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\mu}\left\langle \mathbf{x}\right\rangle_{p(\mathbf{x})}^{\mathrm{T}} - \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} \, ,
\end{aligned}$$

where we have defined $\mathbf{G} := \nabla_{\boldsymbol{\Sigma}}\log(\tilde{Z}(\boldsymbol{\mu}, \boldsymbol{\Sigma}))$.

Matching the Covariance The update (2.4) for the covariance requires to compute

$$\left\langle \mathbf{x}\mathbf{x}^{\mathrm{T}} \right\rangle_{p(\mathbf{x})} - \left\langle \mathbf{x}\right\rangle_{p(\mathbf{x})}\left\langle \mathbf{x}\right\rangle_{p(\mathbf{x})}^{\mathrm{T}} = \boldsymbol{\Sigma} - \boldsymbol{\Sigma}\left(\mathbf{g}\mathbf{g}^{\mathrm{T}} - 2\mathbf{G}\right)\boldsymbol{\Sigma} \, , \tag{4.2}$$

where we used (4.1). Substituting (4.1) and (4.2) into (2.3) and (2.4) we obtain the required updates for the mean and covariance:

$$\begin{aligned}
\boldsymbol{\mu}^{*} &= \boldsymbol{\mu} + \boldsymbol{\Sigma}\mathbf{g} \, , \\
\boldsymbol{\Sigma}^{*} &= \boldsymbol{\Sigma} - \boldsymbol{\Sigma}\left(\mathbf{g}\mathbf{g}^{\mathrm{T}} - 2\mathbf{G}\right)\boldsymbol{\Sigma} \, .
\end{aligned}$$