Biostatistics 202C: Weeks 3 and 4

Andrew Holbrook

Fall 2022

1 Normal approximations to posterior

Again, we have $\mathbf{y} = y_1, \ldots, y_N \overset{iid}{\sim} p(y|\theta)$. By independence, the likelihood takes the form

$$p(\mathbf{y}|\theta) = \prod_{n=1}^{N} p(y_n|\theta)\,.$$

Defining the MLE

$$\hat{\theta}_N = \arg\max_{\theta} p(\mathbf{y}|\theta)\,,$$

it can be shown that (under regularity conditions)

$$\sqrt{N}(\hat{\theta}_N - \theta_0) \xrightarrow{D} \mathrm{N}(0, \mathcal{I}^{-1}(\theta_0))\,,$$

where θ_0 is the true parameter value and $\mathcal{I}(\theta_0)$ is the Fisher information

$$\mathcal{I}(\theta_0) = -\mathrm{E}_{y|\theta}\left(\frac{\partial^2}{\partial\theta^2}\log p(y|\theta_0)\right)\,.$$

Convergence in distribution leads to the approximation

$$\hat{\theta}_N \dot{\sim} \mathrm{N}\left(\theta, \frac{1}{N}\mathcal{I}^{-1}(\theta_0)\right)\,.$$

Convergence in distribution also implies $\hat{\theta}_N \xrightarrow{P} \theta_0$, so we can use Slutsky's theorem to write the more useful result

$$\hat{\theta}_N \dot{\sim} \mathrm{N}\left(\theta, \frac{1}{N}\mathcal{I}^{-1}(\hat{\theta}_N)\right)\,.$$

In machine learning, it is common practice to use optimization to obtain the posterior mode or MAP (maximum a posteriori) estimator

$$\bar{\theta} = \arg\max_{\theta} p(\theta|\mathbf{y})$$

use the MAP estimator in the context of the normal approximation[1]

$$\theta|\mathbf{y} \dot{\sim} \mathrm{N}(\bar{\theta}, \bar{V})\,,$$

where

$$\bar{V} = \left(-\frac{\partial^2}{\partial\theta^2}\log p(\theta|\mathbf{y})\Big|_{\theta=\bar{\theta}}\right)^{-1}\,.$$

[1] Note that, assuming regularity conditions, the Bernstein-von Mises theorem (see previous lecture) says that a similar normal approximation about the MLE will also prove accurate.

We can write the second derivative as

$$\frac{\partial^2}{\partial\theta^2}\log p(\theta|\mathbf{y}) = \frac{\partial^2}{\partial\theta^2}\log p(\mathbf{y}|\theta) + \frac{\partial^2}{\partial\theta^2}\log p(\theta) - \frac{\partial^2}{\partial\theta^2}\log p(\mathbf{y}) = \frac{\partial^2}{\partial\theta^2}\log p(\mathbf{y}|\theta) + \frac{\partial^2}{\partial\theta^2}\log p(\theta)\,.$$

It is then reasonable to quantify the influence of the prior using the fraction of information in the prior (FIP)

$$\text{FIP} = \left.\frac{\frac{\partial^2}{\partial\theta^2}\log p(\theta)}{\frac{\partial^2}{\partial\theta^2}\log p(\theta|\mathbf{y})}\right|_{\theta=\bar{\theta}},$$

or, when $\boldsymbol{\theta}$ is a vector:

$$\text{FIP} = \text{tr}\left[\left(\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\log p(\boldsymbol{\theta})\right)\left(\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\log p(\boldsymbol{\theta}|\mathbf{y})\right)^{-1}\right]\Bigg|_{\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}}.$$

2 Variational inference (VI)

Recall that the KL divergence between two probability distributions is defined

$$D_{KL}(p||q) = \int p(\theta)\log\left(\frac{p(\theta)}{q(\theta)}\right)d\theta\,.$$

Gibbs' inequality says

$$D_{KL}(p||q) \geq 0\,,$$

and $D_{KL}(p||q) = 0$ if and only if $p = q$. To see this, note that $\log(x) \leq x - 1$ for all $x > 0$, and so

$$\begin{aligned} D_{KL}(p||q) &= -\int p(\theta)\log\left(\frac{q(\theta)}{p(\theta)}\right)d\theta \geq -\int p(\theta)\left(\frac{q(\theta)}{p(\theta)} - 1\right)d\theta \\ &= -\int q(\theta)d\theta + \int p(\theta)d\theta \geq 0\,. \end{aligned}$$

In VI, we approximate the posterior distribution $p(\theta|\mathbf{y})$ with a distribution $q(\theta|\psi)$, where ψ are 'variational parameters' we optimize thus:

$$\begin{aligned} \hat{\psi} &= \arg\min_{\psi} D_{KL}(q(\theta|\psi)||p(\theta|\mathbf{y})) = \arg\min_{\psi}\int q(\theta|\psi)\log\left(\frac{q(\theta|\psi)}{p(\theta|\mathbf{y})}\right)d\theta \\ &= \arg\min_{\psi}\int q(\theta|\psi)\log\left(\frac{q(\theta|\psi)p(\mathbf{y})}{p(\theta,\mathbf{y})}\right)d\theta \\ &= \arg\min_{\psi}\int q(\theta|\psi)\left(\log q(\theta|\psi) - \log p(\theta,\mathbf{y})\right)d\theta + \int q(\theta|\psi)\log p(\mathbf{y})d\theta \\ &= \arg\min_{\psi}\int q(\theta|\psi)\left(\log q(\theta|\psi) - \log p(\theta,\mathbf{y})\right)d\theta + \log p(\mathbf{y}) \\ &= \arg\min_{\psi} -\mathcal{L}(\psi,\mathbf{y}) + \log p(\mathbf{y}) \\ &= \arg\max_{\psi}\mathcal{L}(\psi,\mathbf{y})\,, \end{aligned}$$

where we define

$$\mathcal{L}(\psi,\mathbf{y}) := -\int q(\theta|\psi)\left(\log q(\theta|\psi) - \log p(\theta,\mathbf{y})\right)d\theta\,.$$

Rearranging terms, we note that

$$\log p(\mathbf{y}) = D_{KL}(q(\theta|\psi)||p(\theta|\mathbf{y})) + \mathcal{L}(\psi,\mathbf{y})\,,$$

and therefore

$$\log p(\mathbf{y}) \geq \mathcal{L}(\psi,\mathbf{y})\,.$$

For this reason, $\mathcal{L}$ is sometimes referred to as the ELBO (Evidence Lower BOund).[2]

[2] Example on Prof. Shahbaba's notes.

3 Another approach

Alternatively, we may obtain an approximating distribution $q(\theta|\psi)$ by minimizing the (different!) objective function:

$$\hat{\psi} = \arg\min_{\psi} D_{KL}(p(\theta|\mathbf{y})||q(\theta|\psi))\,. \tag{1}$$

Let us assume that the approximating distribution is an exponential family distribution, i.e.,

$$q(\theta|\psi) = h(\theta)\exp\left(\sum_{\alpha}\psi_\alpha\phi_\alpha(\theta) - \Phi(\psi)\right)\,.$$

Here, ψ_α are the elements of the natural parameter, ϕ_α are the elements of the sufficient statistic, and Φ is the log-partition function that enforces integration constraints. The solution to (1) is given by moment matching, i.e., $\hat{\psi}$ satisfies

$$\int q(\theta|\hat{\psi})\phi_\alpha(\theta)d\theta = \int p(\theta|\mathbf{y})\phi_\alpha(\theta)d\theta\,.$$

To see this, we use the fact that the log-partition function of an exponential family distribution satisfies

$$\frac{\partial}{\partial\psi_\alpha}\Phi(\psi_\alpha) = \mathrm{E}_{\theta|\psi}\left(\phi_\alpha\right)\,.$$

This itself may be seen by noting that

$$1 = \int q(\theta|\psi)d\theta = e^{-\Phi(\psi)}\int h(\theta)\exp\left(\sum_{\alpha}\psi_\alpha\phi_\alpha(\theta)\right)d\theta$$

and differentiating both sides:

$$\begin{aligned}0 &= \frac{\partial}{\partial\psi_\alpha}e^{-\Phi(\psi)}\int h(\theta)\exp\left(\sum_{\alpha}\psi_\alpha\phi_\alpha(\theta)\right)d\theta + e^{-\Phi(\psi)}\int h(\theta)\frac{\partial}{\partial\psi_\alpha}\exp\left(\sum_{\alpha}\psi_\alpha\phi_\alpha(\theta)\right)d\theta \\ &= -\frac{\partial}{\partial\psi_\alpha}\Phi(\psi) + \mathrm{E}_{\theta|\psi}\left(\phi_\alpha\right)\,.\end{aligned}$$

We therefore have

$$\begin{aligned}\frac{\partial}{\partial\psi_\alpha}D_{KL}(p(\theta|\mathbf{y})||q(\theta|\psi)) &= \frac{\partial}{\partial\psi_\alpha}\int p(\theta|\mathbf{y})\left(\log p(\theta|\mathbf{y}) - \log q(\theta|\psi)\right)d\theta \\ &= -\int p(\theta|\mathbf{y})\frac{\partial}{\partial\psi_\alpha}\log q(\theta|\psi)d\theta \\ &= -\int p(\theta|\mathbf{y})\frac{\partial}{\partial\psi_\alpha}\left(\sum_{\alpha}\psi_\alpha\phi_\alpha(\theta) - \Phi(\psi)\right)d\theta \\ &= -\int p(\theta|\mathbf{y})\left(\phi_\alpha(\theta) - \frac{\partial}{\partial\psi_\alpha}\Phi(\psi)\right)d\theta \\ &= -\int p(\theta|\mathbf{y})\left(\phi_\alpha(\theta) - \mathrm{E}_{\theta|\psi}(\phi_\alpha)\right)d\theta \\ &= -\int p(\theta|\mathbf{y})\phi_\alpha(\theta)d\theta - \mathrm{E}_{\theta|\psi}(\phi_\alpha)\,.\end{aligned}$$

Setting this derivative to 0 leads to the moment matching solution to (1):

$$\int q(\theta|\hat{\psi})\phi_\alpha(\theta)d\theta = \int p(\theta|\mathbf{y})\phi_\alpha(\theta)d\theta\,.$$

3.1 Assumed density filtering

Note that the posterior distribution may be written as the product of N 'compatibility functions'

$$p(\theta|\mathbf{y}) \propto p(\theta) \prod_{n=1}^{N} p(y_n|\theta) = \prod_{n=0}^{N} p_n(\theta)\,.$$

where $p_0(\theta) := p(\theta)$ and $p_n(\theta) := p(y_n|\theta)$. Assumed density filtering (ADF) constructs an approximation of $p(\theta|\mathbf{y})$ by first finding

$$\psi^0 = \arg\min_{\psi} D_{KL}(p_0(\theta)||q(\theta|\psi))$$

and then, in order, iteratively updating the approximation by finding

$$\psi^n = \arg\min_{\psi} D_{KL}(p_n(\theta)q(\theta|\psi^{n-1})||q(\theta|\psi))\,.$$

When the prior $p(\theta)$ is a member of the exponential family $q(\theta|\psi)$, the initial update ψ^0 is given by the prior parameters. Subsequent ψ^n are obtained using moment matching.

ADF thus uses the current approximation to help guide the construction of the next approximation and generally performs better than taking the product of N independent approximations. On the other hand, the method is sensitive to the ordering and can be thrown off by bad starting approximations.

3.2 Expectation propagation

Expectation propagation (EP) iteratively improves an approximation to the posterior $q(\theta|\psi) \propto \prod_n r_n(\theta)$. The algorithm begins by setting each $r_n = 1$ and performs the following steps until convergence:

1. Randomly select $n \sim \text{Unif}\{0, 1, \ldots, N\}$;
2. Remove r_n from current posterior by dividing and normalizing

   $$q(\theta|\psi^{-n}) \propto \frac{q(\theta|\psi)}{r_n(\theta)}\,.$$

   Note that this 'cavity distribution' is in the same exponential family distribution if r_n is in it or is constant (although it might not be normalizable).
3. Update exponential family approximation to posterior $q(\theta|\psi^*)$ by finding

   $$\psi^* = \arg\min_{\psi} D_{KL}(p_n(\theta)q(\theta|\psi^{-n})||q(\theta|\psi))\,.$$

   Again, this step performed by moment matching.
4. Update exponential family approximation of r_n as

   $$r_n(\theta) \propto \frac{q(\theta|\psi^*)}{q(\theta|\psi^{-n})}\,.$$

   Note that this step enforces r_n's membership within the (not necessarily normalizable) exponential family.

EP often outperforms ADU, but there is no guarantee of convergence. Fixed points exist for approximation distributions belonging to exponential family. Lack of normalization guarantees means possibility of, e.g., negative variances for normal approximations.

3.3 The clutter problem

Suppose we have Gaussian observations $\mathbf{Y} = \mathbf{y}_1, \ldots, \mathbf{y}_N$ in a 'sea of unrelated clutter', which we model using the mixture distribution

$$p(\mathbf{y}|\boldsymbol{\theta}) = (1-w)\text{N}(\mathbf{y}|\boldsymbol{\theta}, \mathbf{I}_D) + w\text{N}(\mathbf{y}|\mathbf{0}, 10\mathbf{I}_D)\,.$$

We also specify the prior $p(\boldsymbol{\theta}) \sim \text{N}(\mathbf{0}, 100\mathbf{I}_D)$. Let $p_0(\boldsymbol{\theta}) = p(\boldsymbol{\theta})$, and $p_n(\boldsymbol{\theta}) := p(\mathbf{y}_n|\boldsymbol{\theta})$ for $n > 0$. Finally, we specify the Gaussian approximation distribution $q(\boldsymbol{\theta}) = \text{N}(\mathbf{m}_\theta, v_\theta\mathbf{I}_D)$.

3.3.1 Assumed density filtering

At iteration 0, we minimize the KL divergence between p_0 and q, subject to the restriction that q be Gaussian with mean $\mathbf{m}_\theta$ and covariance $v_\theta \mathbf{I}_D$. Thus, iteration 0 sets q to be equal to the prior $p = p_0$. At each subsequent nth step, we obtain the 'exact posterior'

$$\hat{p}(\boldsymbol{\theta}) = \frac{p_n(\boldsymbol{\theta}) q^{-n}(\boldsymbol{\theta})}{\int p_n(\boldsymbol{\theta}) q^{-n}(\boldsymbol{\theta}) d\boldsymbol{\theta}}$$

and minimize $D_{KL}(\hat{p}||q)$ subject to the spherical Gaussian specification of q. Due to the fact that the Gaussian distribution belongs to the exponential family, this solution is obtained by moment matching:

$$\begin{aligned} \mathrm{E}_q(\boldsymbol{\theta}) &= \mathrm{E}_{\hat{p}}(\boldsymbol{\theta}) \,, \\ \mathrm{E}_q(\boldsymbol{\theta}^T \boldsymbol{\theta}) &= \mathrm{E}_{\hat{p}}(\boldsymbol{\theta}^T \boldsymbol{\theta}) \,. \end{aligned}$$

Each step also produces the normalization factor

$$z_n = \int p_n(\boldsymbol{\theta}) q^{-n}(\boldsymbol{\theta}) d\boldsymbol{\theta} \,,$$

and we use the product of these factors to estimate $p(\mathbf{Y})$. Here, we have

$$z_n = (1-w)\mathrm{N}(\mathbf{y}_n | \mathbf{m}_\theta^{-n}, (v_\theta^{-n} + 1)\mathbf{I}_D) + w\mathrm{N}(\mathbf{y}_n | \mathbf{0}, 10\mathbf{I}_D) \,.$$

For the clutter problem, the ADF algorithm follows these steps:

1. Initialize $\mathbf{m}_\theta = \mathbf{0}$, $v_\theta^{-n} = 100$, $s = 1$;
2. For $n = 1, \ldots, N$, update $(\mathbf{m}_\theta, v_\theta, s)$ according to
   (a) $s = s^{-n} z_n$;
   (b) $\pi_n = 1 - \frac{w}{z_n}\mathrm{N}(\mathbf{y}_n | \mathbf{0}, 10\mathbf{I}_D)$;
   (c) $\mathbf{m}_\theta = \mathbf{m}_\theta^{-n} + v_\theta^{-n} \pi_n \frac{\mathbf{y}_n - \mathbf{m}_\theta^{-n}}{v_\theta^{-n}+1}$;
   (d) $v_\theta = v_\theta^{-n} - \pi_n \frac{(v_\theta^{-n})^2}{v_\theta^{-n}+1} + \pi_n(1-\pi_n)\frac{(v_\theta^{-n})^2 \|\mathbf{y}_n - \mathbf{m}_\theta^{-n}\|^2}{D(v_\theta^{-n}+1)^2}$.

3.3.2 Expectation propagation

For the same problem, the EP term approximations take the form

$$r_n(\boldsymbol{\theta}) = s_n \exp\left(-\frac{1}{2v_n}(\boldsymbol{\theta} - \mathbf{m}_n)^T (\boldsymbol{\theta} - \mathbf{m}_n) \right) \,,$$

where $q(\boldsymbol{\theta}) \propto \prod_n r_n(\boldsymbol{\theta})$. The EP algorithm proceeds as follows.

1. Initialize the prior terms $v_0 = 100$, $\mathbf{m}_0 = \mathbf{0}$, $s_0 = (2\pi v_0)^{-D/2}$ and the data terms so that $r_n(\boldsymbol{\theta}) \propto 1$: $v_n = \infty$, $\mathbf{m}_n = \mathbf{0}$ and $s_n = 1$;
2. $\mathbf{m}_\theta = \mathbf{m}_0$, $v_\theta = v_0$;
3. Until all $(\mathbf{m}_n, v_n, s_n)$ converge, loop $n = 1, \ldots, N$:
   (a) Remove r_n from the posterior to get 'old' posterior:

$$\begin{aligned} (v_\theta^{-n})^{-1} &= v_\theta^{-1} - v_n^{-1} \\ \mathbf{m}_\theta^{-n} &= \mathbf{m}_\theta + \frac{v_\theta^{-n}}{v_n}(\mathbf{m}_\theta - \mathbf{m}_n) \,; \end{aligned}$$

   (b) Recompute $(\mathbf{m}_\theta, v_\theta, z_n)$ from $(\mathbf{m}_\theta^{-n}, v_\theta^{-n})$ as in ADF:
      i. $z_n = (1-w)\mathrm{N}(\mathbf{y}_n | \mathbf{m}_\theta^{-n}, (v_\theta^{-n} + 1)\mathbf{I}_D) + w\mathrm{N}(\mathbf{y}_n | \mathbf{0}, 10\mathbf{I}_D)$;

ii. $\pi_n = 1 - \frac{w}{z_n}\mathrm{N}(\mathbf{y}_n|\mathbf{0}, 10\mathbf{I}_D)$;

iii. $\mathbf{m}_\theta = \mathbf{m}_\theta^{-n} + v_\theta^{-n}\pi_n \frac{\mathbf{y}_n - \mathbf{m}_\theta^{-n}}{v_\theta^{-n}+1}$;

iv. $v_\theta = v_\theta^{-n} - \pi_n \frac{(v_\theta^{-n})^2}{v_\theta^{-n}+1} + \pi_n(1-\pi_n)\frac{(v_\theta^{-n})^2\|\mathbf{y}_n - \mathbf{m}_\theta^{-n}\|^2}{D(v_\theta^{-n}+1)^2}$;

(c) Update r_n:

$$
\begin{aligned}
v_n^{-1} &= v_\theta^{-1} - (v_\theta^{-n})^{-1} \\
\mathbf{m}_n &= \mathbf{m}_\theta^{-n} + \frac{(v_n + v_\theta^{-n})}{v_\theta^{-n}}(\mathbf{m}_\theta - \mathbf{m}_\theta^{-n}) \\
s_n &= \frac{z_n}{(2\pi v_n)^{D/2}\mathrm{N}(\mathbf{m}_n|\mathbf{m}_\theta^{-n}, (v_n + v_\theta^{-n})\mathbf{I})}\,;
\end{aligned}
$$

4. Compute normalizing constant (if you want it):

$$
\begin{aligned}
V &= \frac{\mathbf{m}_\theta^T\mathbf{m}_\theta}{v_\theta} - \sum_n \frac{\mathbf{m}_n^T\mathbf{m}_n}{v_n} \\
p(\mathbf{Y}) &\approx (2\pi v_\theta)^{D/2} e^{V/2} \prod_{n=0}^{N} s_n\,.
\end{aligned}
$$

4 Score matching

Here, we take a break from Bayesian inference but continue to use information theoretic tools. Suppose our data $\mathbf{y} \in \mathbb{R}^D$ follow a true distribution with pdf $p_\mathbf{y}(\cdot)$ and that we model these data using the distribution $p(\cdot|\boldsymbol{\theta})$ for $\boldsymbol{\theta} \in \mathbb{R}^P$. That is, we wish to approximate $p_\mathbf{y}(\cdot)$ with $p(\cdot|\hat{\boldsymbol{\theta}})$ for some $\hat{\boldsymbol{\theta}}$ that we estimate using the data $\mathbf{y}$. Further, assume we can only compute the model pdf up to a constant:

$$
p(\mathbf{y}|\boldsymbol{\theta}) = \frac{1}{z(\boldsymbol{\theta})} q(\mathbf{y}|\boldsymbol{\theta})\,.
$$

That is, we know how to compute $q(\mathbf{y}|\boldsymbol{\theta})$ but we do not know how to compute

$$
z(\boldsymbol{\theta}) = \int_{\mathbf{y}\in\mathbb{R}^D} q(\mathbf{y}|\boldsymbol{\theta}) d\mathbf{y}\,.
$$

Define the model 'score function' as the gradient of the model log-pdf with respect to the data $\mathbf{y}$:

$$
\boldsymbol{\psi}(\mathbf{y}, \boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial \log p(\mathbf{y}|\boldsymbol{\theta})}{\partial y_1} \\ \vdots \\ \frac{\partial \log p(\mathbf{y}|\boldsymbol{\theta})}{\partial y_D} \end{pmatrix} = \begin{pmatrix} \psi_1(\mathbf{y}, \boldsymbol{\theta}) \\ \vdots \\ \psi_D(\mathbf{y}, \boldsymbol{\theta}) \end{pmatrix} = \nabla_\mathbf{y} \log p(\mathbf{y}|\boldsymbol{\theta}) = \nabla_\mathbf{y} \log q(\mathbf{y}|\boldsymbol{\theta})\,.
$$

Similarly, define the data score function $\boldsymbol{\psi}_\mathbf{y}(\cdot) = \nabla_\mathbf{y} \log p_\mathbf{y}(\cdot)$ as the gradient for the true distribution. Now define the expected squared distance between scores as

$$
j(\boldsymbol{\theta}) = \frac{1}{2}\int_{\mathbf{y}\in\mathbb{R}^D} p_\mathbf{y}(\mathbf{y})\|\boldsymbol{\psi}(\mathbf{y}|\boldsymbol{\theta}) - \boldsymbol{\psi}_\mathbf{y}(\mathbf{y})\|^2 d\mathbf{y}\,.
$$

Then the score matching estimator is defined

$$
\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} j(\boldsymbol{\theta})\,. \tag{2}
$$

This estimator does not require knowledge of $z(\boldsymbol{\theta})$, but computation of $j(\boldsymbol{\theta})$ apparently requires non-parametric estimation of $p_\mathbf{y}$ and its derivative. But wait...

Theorem 1. *Assume that the model score function $\boldsymbol{\psi}(\mathbf{y}|\boldsymbol{\theta})$ is differentiable and that weak regularity conditions hold. Then the objective function $j(\boldsymbol{\theta})$ of (2) may be expressed as*

$$\begin{aligned} j(\boldsymbol{\theta}) &\propto \int_{\mathbf{y}\in\mathbb{R}^D} p_{\mathbf{y}}(\mathbf{y}) \sum_{d=1}^{D} \left(\partial_d \psi_d(\mathbf{y}|\boldsymbol{\theta}) + \frac{1}{2}\psi_d(\mathbf{y}|\boldsymbol{\theta})^2 \right) d\mathbf{y} \\ &= \int_{\mathbf{y}\in\mathbb{R}^D} p_{\mathbf{y}}(\mathbf{y}) \sum_{d=1}^{D} \left(\frac{\partial^2 \log q(\mathbf{y}|\boldsymbol{\theta})}{\partial y_d^2} + \frac{1}{2}\left(\frac{\partial \log q(\mathbf{y}|\boldsymbol{\theta})}{\partial y_d} \right)^2 \right) d\mathbf{y}\,, \end{aligned}$$

Proof. The formula in (2) becomes

$$\begin{aligned} j(\boldsymbol{\theta}) &= \int p_{\mathbf{y}}(\mathbf{y}) \left(\frac{1}{2}\|\boldsymbol{\psi}_{\mathbf{y}}(\mathbf{y})\|^2 + \frac{1}{2}\|\boldsymbol{\psi}(\mathbf{y}|\boldsymbol{\theta})\|^2 - \boldsymbol{\psi}_{\mathbf{y}}(\mathbf{y})^T \boldsymbol{\psi}(\mathbf{y}|\boldsymbol{\theta}) \right) d\mathbf{y} \\ &\propto \int p_{\mathbf{y}}(\mathbf{y}) \left(\frac{1}{2}\|\boldsymbol{\psi}(\mathbf{y}|\boldsymbol{\theta})\|^2 - \boldsymbol{\psi}_{\mathbf{y}}(\mathbf{y})^T \boldsymbol{\psi}(\mathbf{y}|\boldsymbol{\theta}) \right) d\mathbf{y} \\ &= \int p_{\mathbf{y}}(\mathbf{y}) \frac{1}{2}\|\boldsymbol{\psi}(\mathbf{y}|\boldsymbol{\theta})\|^2 d\mathbf{y} - \sum_{d=1}^{D} \int p_{\mathbf{y}}(\mathbf{y}) \psi_d(\mathbf{y}) \psi_d(\mathbf{y}|\boldsymbol{\theta}) d\mathbf{y} \\ &= \int p_{\mathbf{y}}(\mathbf{y}) \frac{1}{2}\|\boldsymbol{\psi}(\mathbf{y}|\boldsymbol{\theta})\|^2 d\mathbf{y} - \sum_{d=1}^{D} \int \frac{p_{\mathbf{y}}(\mathbf{y})}{p_{\mathbf{y}}(\mathbf{y})} \frac{\partial p_{\mathbf{y}}(\mathbf{y})}{\partial y_d} \psi_d(\mathbf{y}|\boldsymbol{\theta}) d\mathbf{y} \\ &= \int p_{\mathbf{y}}(\mathbf{y}) \frac{1}{2}\|\boldsymbol{\psi}(\mathbf{y}|\boldsymbol{\theta})\|^2 d\mathbf{y} - \sum_{d=1}^{D} \int \frac{\partial p_{\mathbf{y}}(\mathbf{y})}{\partial y_d} \psi_d(\mathbf{y}|\boldsymbol{\theta}) d\mathbf{y} \\ &= \int p_{\mathbf{y}}(\mathbf{y}) \frac{1}{2}\|\boldsymbol{\psi}(\mathbf{y}|\boldsymbol{\theta})\|^2 d\mathbf{y} + \sum_{d=1}^{D} \int p_{\mathbf{y}}(\mathbf{y}) \frac{\partial \psi_d(\mathbf{y}|\boldsymbol{\theta})}{\partial y_d} d\mathbf{y} \\ &= \int_{\mathbf{y}\in\mathbb{R}^D} p_{\mathbf{y}}(\mathbf{y}) \sum_{d=1}^{D} \left(\partial_d \psi_d(\mathbf{y}|\boldsymbol{\theta}) + \frac{1}{2}\psi_d(\mathbf{y}|\boldsymbol{\theta})^2 \right) d\mathbf{y} \end{aligned}$$

The penultimate step follows from integration by parts and the assumption that $p_{\mathbf{y}}$ goes to 0 at $\pm\infty$. Actually, Hyvarinen (2005) proves the more general multivariate integration by parts required. □

In practice, one observes $\mathbf{y}_1, \ldots, \mathbf{y}_N \sim p_{\mathbf{y}}(\mathbf{y})$. The empirical equivalent to (2) is given by

$$\tilde{j}(\boldsymbol{\theta}) \propto \frac{1}{N} \sum_{n=1}^{N} \sum_{d=1}^{D} \left(\partial_d \psi_d(\mathbf{y}_n|\boldsymbol{\theta}) + \frac{1}{2}\psi_d(\mathbf{y}_n|\boldsymbol{\theta})^2 \right)\,,$$

and this converges to $j(\boldsymbol{\theta})$ by the law of large numbers.

4.1 Multivariate Gaussian

Consider the multivariate Gaussian model

$$p(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Omega}) = \frac{1}{z(\boldsymbol{\mu}, \boldsymbol{\Omega})} \exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^T \boldsymbol{\Omega}(\mathbf{y}-\boldsymbol{\mu}) \right)\,.$$

Then

$$q(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Omega}) = \exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^T \boldsymbol{\Omega}(\mathbf{y}-\boldsymbol{\mu}) \right)\,,$$

$$\boldsymbol{\psi}(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Omega}) = -\boldsymbol{\Omega}(\mathbf{y}-\boldsymbol{\mu})\,,$$

and

$$\partial_d \psi_d(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Omega}) = -\omega_{dd} := -[\boldsymbol{\Omega}]_{dd}\,.$$

It follows that

$$\tilde{j}(\boldsymbol{\mu}, \boldsymbol{\Omega}) = \frac{1}{N} \sum_{n=1}^{N} \left(\left(\sum_{d=1}^{D} -\omega_{dd} \right) + \frac{1}{2} (\mathbf{y}_n - \boldsymbol{\mu})^T \boldsymbol{\Omega}\boldsymbol{\Omega} (\mathbf{y}_n - \boldsymbol{\mu}) \right) .$$

Solve for $\hat{\boldsymbol{\mu}}$ by setting the following gradient to 0:

$$\nabla_{\boldsymbol{\mu}} \tilde{j}(\boldsymbol{\mu}, \boldsymbol{\Omega}) = \boldsymbol{\Omega}\boldsymbol{\Omega}\boldsymbol{\mu} - \boldsymbol{\Omega}\boldsymbol{\Omega} \left(\frac{1}{N} \sum_{n=1}^{N} \mathbf{y}_n \right) .$$

Because $\boldsymbol{\Omega}$ is positive definite, and so invertible, we obtain the familiar MLE $\hat{\boldsymbol{\mu}} = \bar{\mathbf{y}}$. Next,

$$\nabla_{\boldsymbol{\Omega}} \tilde{j}(\boldsymbol{\mu}, \boldsymbol{\Omega}) = -\mathbf{I} + \boldsymbol{\Omega} \left(\frac{1}{2N} \sum_{n=1}^{N} (\mathbf{y}_n - \boldsymbol{\mu})(\mathbf{y}_n - \boldsymbol{\mu})^T \right) + \left(\frac{1}{2N} \sum_{n=1}^{N} (\mathbf{y}_n - \boldsymbol{\mu})(\mathbf{y}_n - \boldsymbol{\mu})^T \right) \boldsymbol{\Omega} .$$

I recommend using the matrix differential to calculate the above gradient. This is equal to 0 when $\boldsymbol{\Omega}$ is the inverse of the MLE.

4.2 Matrix differential briefly

Suppose we have a function $h : \mathbf{M}^{n \times n} \to \mathbb{R}$. Then the matrix differential d relates to the gradient by the following identity:

$$dh(\boldsymbol{\Omega}) = \operatorname{tr}\left((d\boldsymbol{\Omega}) \mathbf{G} \right) \iff \nabla h(\boldsymbol{\Omega}) = \mathbf{G} .$$

Before we calculate the differential, we have

$$\begin{aligned}
\tilde{j}(\boldsymbol{\mu}, \boldsymbol{\Omega}) &= \frac{1}{N} \sum_{n=1}^{N} \left(\left(\sum_{d=1}^{D} -\omega_{dd} \right) + \frac{1}{2} (\mathbf{y}_n - \boldsymbol{\mu})^T \boldsymbol{\Omega}\boldsymbol{\Omega} (\mathbf{y}_n - \boldsymbol{\mu}) \right) \\
&= \frac{1}{N} \sum_{n=1}^{N} \left(-\operatorname{tr}(\boldsymbol{\Omega}) + \frac{1}{2} \operatorname{tr}\left((\mathbf{y}_n - \boldsymbol{\mu})^T \boldsymbol{\Omega}\boldsymbol{\Omega} (\mathbf{y}_n - \boldsymbol{\mu}) \right) \right) \\
&= \frac{1}{N} \sum_{n=1}^{N} \left(+\frac{1}{2} \operatorname{tr}\left(\boldsymbol{\Omega} (\mathbf{y}_n - \boldsymbol{\mu})(\mathbf{y}_n - \boldsymbol{\mu})^T \boldsymbol{\Omega} \right) \right) \\
&= -\operatorname{tr}(\boldsymbol{\Omega}) + \operatorname{tr}\left(\boldsymbol{\Omega} \left(\frac{1}{2N} \sum_{n=1}^{N} (\mathbf{y}_n - \boldsymbol{\mu})(\mathbf{y}_n - \boldsymbol{\mu})^T \right) \boldsymbol{\Omega} \right) .
\end{aligned}$$

The rest follows from the linearity of the trace and differential operators followed by the product rule.

4.3 Note on terminology

The term 'score function' typically denotes the gradient of the log-likelihood w.r.t. the model parameter. Here, it is used to refer to the gradient w.r.t. the data. In fact, the two are the same for location family distributions of the following form:

$$p(\mathbf{y}|\boldsymbol{\mu}) = p(\|\mathbf{y} - \boldsymbol{\mu}\|) ,$$

where $\|\cdot\|$ denotes the L2 norm.