Appendix C. Quality of the posterior estimates from prior replacement

In the results obtained for single MDN inversions (Figures 2 and 3) we observed qualitatively that prior replacement may out-perform prior-specific training in some aspects of the quality of the estimated posterior distribution. We hypothesised that this effect could be attributed to the difference in the distribution of samples used to train the network in each case. We now test this hypothesis by investigating a Bayesian inverse problem in which *sampling* (rather than a neural network) is used to estimate a single posterior PDF.

To do this we suppose that we have a likelihood distribution which we can only evaluate up to a multiplicative constant, and a prior which we know parametrically. Consequently, we do not know the posterior (equation 3) analytically (i.e., we do not know the normalising constant - as is often the case in practical problems). The usual approach (Mosegaard & Sambridge 2002) is to sample directly from the posterior using McMC methods in order to estimate the posterior density. We call this *direct estimation*. However, since the prior is known analytically, the posterior can also be estimated by prior replacement. To do this we would construct an old posterior using the appropriate likelihood (i.e., that used in direct estimation) and a broad old prior. We would then sample from this old posterior and estimate its density. Then the prior replacement equations would be applied to replace the old prior with the new prior (i.e., the appropriate prior used in direct estimation). Henceforth we refer to this as *indirect estimation*.

Direct and indirect estimation are equivalent to prior-specific training and prior replacement, respectively, in the preceding discussion of MDN inversion. The only difference now is that we assume that the samples are being used to directly estimate a posterior for a given datum, rather than to estimate the parameters of a neural network which will predict the posterior for any data.

Henceforth, we analyse the quality of the posterior estimate obtained using direct and indirect estimation for a single continuous model parameter, m. Furthermore we assume that the forward function is such that it describes an unnormalised Gaussian over m (for an example of such a likelihood function, see Tarantola 2002, pp.64-68). This likelihood may be written as the product of a normalised Gaussian and a constant,

$$p\left(d|m\right) = c_1 \phi\left(m; \mu_L, \Sigma_L\right). \tag{C.1}$$

We assume also that the new prior is Gaussian, thus

$$p_{new}\left(m\right) = \phi\left(m; \mu_B, \Sigma_B\right). \tag{C.2}$$

The new posterior can then be formed by substituting equations C.1 and C.2 into equation 3. Cancelling the c_1 constants from the denominator and numerator of the resulting expression we obtain the normalised product of two Gaussians, which is another Gaussian

$$\begin{aligned} p_{new}\left(m|d\right) &= \frac{\phi\left(m; \mu_B, \Sigma_B\right)\phi\left(m; \mu_L, \Sigma_L\right)}{\int\limits_m \phi\left(m; \mu_B, \Sigma_B\right)\phi\left(m; \mu_L, \Sigma_L\right)} \\ &= \phi\left(m; \mu_P, \Sigma_P\right), \end{aligned} \tag{C.3}$$

where the new posterior Gaussian will have mean and variance given by

$$\Sigma_P = \left({\Sigma_B}^{-1} + {\Sigma_L}^{-1}\right)^{-1}, \quad \mu_P = \Sigma_P\left({\Sigma_B}^{-1}\mu_B + {\Sigma_L}^{-1}\mu_L\right). \tag{C.4}$$

(Bromiley 2003). It should be noted that generally if we assume Gaussian forms for our prior, likelihood and hence posterior there is no need for Monte-Carlo sampling and PDF estimation. However, we use this toy problem to investigate the difference between direct estimation (prior-specific training) and indirect estimation (prior replacement). We now describe direct and indirect estimation in more detail, and then also the methods by which we can compare the quality of the posterior estimates we obtain in each case.

Appendix C.1. Direct estimation

In direct estimation a set of N samples, $M_1, ..., M_i, ..., M_N$, are made directly from the new posterior, i.e., $M_i \sim p_{new}\left(m|d\right)$. These are then used to estimate the parameters of the new posterior distribution. We denote the estimate

$$\hat{p}_D\left(m|d\right) = \phi\left(m; \hat{\mu}_{P_D}, \hat{\Sigma}_{P_D}\right) \approx p_{new}\left(m|d\right) \tag{C.5}$$

where the maximum likelihood estimates (MLE) of the mean and variance are related to the N samples by

$$\hat{\mu}_{P_D} = \frac{1}{N}\sum_i^N M_i, \quad \hat{\Sigma}_{P_D} = \frac{1}{N-1}\sum_i^N \left(M_i - \hat{\mu}_{P_D}\right)^2, \tag{C.6}$$

and these are therefore termed the direct estimators.

Appendix C.2. Indirect estimation

In indirect estimation samples are made from an old posterior and are used to estimate that distribution. Then prior replacement is used to determine an estimate of the new posterior by emplacing the appropriate new prior. Initially, we assume an infinitely-broad, Uniform old prior thus the PDF is constant (and improper, see e.g., Hobert & Casella 1996, Daniels 1999, Sun et al. 2001)

$$p_{old}\left(m\right) = c_2. \tag{C.7}$$

This is then used to construct the old posterior: by substituting equations C.1 and C.7 into equation 2, and cancelling constant terms we obtain a Gaussian

$$\begin{aligned} p_{old}\left(m|d\right) &= \frac{c_1 c_2 \phi\left(m; \mu_L, \Sigma_L\right)}{\int\limits_m c_1 c_2 \phi\left(m; \mu_L, \Sigma_L\right) dm} \\ &= \phi\left(m; \mu_L, \Sigma_L\right). \end{aligned} \tag{C.8}$$

This is simply a normalised version of the likelihood. As in direct estimation, we then use N samples from this distribution to obtain an approximation to it, which we denote

$$\hat{p}_{old}\left(m|d\right) = \phi\left(m; \hat{\mu}_L, \hat{\Sigma}_L\right) \approx p_{old}\left(m|d\right). \tag{C.9}$$

The MLE estimators for the variance and mean parameters are given by

$$\hat{\mu}_L = \frac{1}{N}\sum_i^N M_i, \quad \hat{\Sigma}_L = \frac{1}{N-1}\sum_i^N \left(M_i - \hat{\mu}_L\right)^2 \tag{C.10}$$

where M_i are now samples made from $\hat{p}_{old}\left(m|d\right)$. We now perform prior replacement in order to obtain an estimate of $p_{new}\left(m|d\right)$. To do this we substitute the expressions for the approximate old posterior, the old prior and the new prior (equations C.9, C.7 and C.2 respectively) into equation 5 such that we obtain an approximation for the new posterior given by

$$\hat{p}_I\left(m|d\right) = \frac{1}{k}\frac{\phi\left(m; \mu_B, \Sigma_B\right)}{c_2}\phi\left(m; \hat{\mu}_L, \hat{\Sigma}_L\right) \approx p_{new}\left(m|d\right) \tag{C.11}$$

where $\hat{p}_I\left(m|d\right)$ is used to denote this (indirect) approximation to $p_{new}\left(m|d\right)$. Making the same substitutions in equation 6 yields the approximate normalising constant

$$k \approx \int_{-\infty}^{+\infty} \frac{\phi\left(m; \mu_B, \Sigma_B\right)}{c_2}\phi\left(m; \hat{\mu}_L, \hat{\Sigma}_L\right) dm. \tag{C.12}$$

Substituting equation C.12 into equation C.11 and cancelling the constant old prior, c_2, we obtain

$$\begin{aligned} \hat{p}_I\left(m|d\right) &= \frac{\phi\left(m; \mu_B, \Sigma_B\right)\phi\left(m; \hat{\mu}_L, \hat{\Sigma}_L\right)}{\int_{-\infty}^{\infty} \phi\left(m; \mu_B, \Sigma_B\right)\phi\left(m; \hat{\mu}_L, \hat{\Sigma}_L\right) dm} \\ &= \phi\left(m; \hat{\mu}_{P_I}, \hat{\Sigma}_{P_I}\right), \end{aligned} \tag{C.13}$$

which we have recognised as a normalised product of two Gaussians, which is a Gaussian. As such we can obtain the mean and variance using the standard identities for a Gaussian multiplication (Bromiley 2003) as

$$\hat{\Sigma}_{P_I} = \left(\Sigma_B{}^{-1} + \hat{\Sigma}_L^{-1}\right)^{-1}, \quad \hat{\mu}_{P_I} = \hat{\Sigma}_{P_I}\left(\Sigma_B{}^{-1}\mu_B + \hat{\Sigma}_L^{-1}\hat{\mu}_L\right), \tag{C.14}$$

and these are therefore termed the indirect estimators.

Appendix C.3. Comparing quality

To compare the quality of the two posterior estimates we calculate the variance and bias of the estimators (the mean and variance parameters) in each case. If we use the example of the variance parameter Σ, and the estimator of it $\hat{\Sigma}$, then the bias and the variance of the estimator are defined as

$$\text{bias}\left(\hat{\Sigma}\right) = \mathrm{E}\left[\hat{\Sigma}\right] - \Sigma, \tag{C.15}$$

$$\text{var}\left(\hat{\Sigma}\right) = \mathrm{E}\left[\left(\mathrm{E}\left[\hat{\Sigma}\right] - \hat{\Sigma}\right)^2\right]. \tag{C.16}$$

Exact analytical expressions exist for these quantities for given N in the case of the direct estimators: they are simply those for a Gaussian which are well known (see e.g., Ulrych et al. 2001), thus the biases are

$$\text{bias}\left(\hat{\mu}_{P_D}\right) = \text{bias}\left(\hat{\Sigma}_{P_D}\right) = 0. \tag{C.17}$$

and the variances are

$$\text{var}\left(\hat{\mu}_{P_D}\right) = \frac{\Sigma_P}{n}, \tag{C.18}$$

$$\text{var}\left(\hat{\Sigma}_{P_D}\right) = \frac{2\Sigma_P{}^2}{n-1}. \tag{C.19}$$

No such exact analytical expressions exist for the bias and variance for the indirect estimators. However, we have derived approximations to these in Appendix D based on third-order Taylor expansions taken about the expected values of the $\hat{\mu}_L$ and $\hat{\Sigma}_L$ estimators (Oehlert 1992, Van der Vaart 2000):

$$\text{bias}\left(\hat{\mu}_{P_I}\right) \approx \frac{1}{n-1}\left(\mu_L - \mu_P\right)\left(\Sigma_L^{-1}\Sigma_P - \Sigma_L^{-2}\Sigma_P^2\right), \tag{C.20}$$

$$\text{bias}\left(\hat{\Sigma}_{P_I}\right) \approx \frac{2}{n-1}\left(\Sigma_L^{-2}\Sigma_P{}^3 - \Sigma_L^{-1}\Sigma_P{}^2\right) \tag{C.21}$$

and the variances are

$$\text{var}\left(\hat{\mu}_{P_I}\right) \approx \left(\mu_P - \mu_L\right)^2 \frac{2\Sigma_P^2\Sigma_L^{-2}}{n-1} + \frac{\Sigma_P^2\Sigma_L^{-1}}{n}, \tag{C.22}$$

$$\mathrm{var}\left(\hat{\Sigma}_{P_I}\right) \approx \frac{2}{n-1}\frac{\Sigma_P^4}{\Sigma_L^2}. \tag{C.23}$$

Another measure of approximation quality is the Kullback-Leibler (KL) divergence (Kullback & Leibler 1951), which measures the difference between two PDFs. Suppose that we make an estimate $\hat{p}(m|d)$ of a distribution $p(m|d)$. The KL divergence between the two, $D_{KL}\left[p(m|d) \,||\, \hat{p}(m|d)\right]$, is given by

$$D_{KL}\left[p\left(m|d\right) \,||\, \hat{p}\left(m|d\right)\right] = \int_{-\infty}^{+\infty} \ln\left(\frac{p(m|d)}{\hat{p}(m|d)}\right) p(m|d)\, dm. \tag{C.24}$$

This quantity is used extensively to measure approximation quality because of its intuitively appealing interpretation as the amount of information lost when approximating $p(m|d)$ by $\hat{p}(m|d)$ (Hershey & Olsen 2007). Thus we interpret the KL divergence as a measure of the overall 'goodness of fit' of an approximate distribution. However, the advantage of the bias and variance quantities is that they are expected measures of the accuracy and precision, respectively, given a certain number of samples N. The KL divergence is only defined between two known distributions, therefore what we require is the expected KL divergence given that $\hat{p}(m|d)$ has been estimated using a certain number of samples N. No analytical expression exists for this quantity, thus we have to obtain an estimate of it empirically. That is to say, we must make a large number, L, of new posterior estimates and use this population to estimate the average value, which would be calculated from the L estimates as

$$\mathrm{E}\left[D_{KL}\left[p\left(m|d\right) \,||\, \hat{p}\left(m|d\right)\right]\right] \approx \frac{1}{L}\sum_{l}^{L} D_{KL}\left[p\left(m|d\right) \,||\, \hat{p}_l\left(m|d\right)\right] \tag{C.25}$$

where $\hat{p}_l(m|d)$ is the l^{th} estimate of the posterior. Thus in practice we made an estimate of the posterior L times using both methods and calculated

$$D_{KL}\left[p(m|d) \,||\, \hat{p}_{I,l}(m|d)\right] \text{ and } D_{KL}\left[p(m|d) \,||\, \hat{p}_{D,l}(m|d)\right]$$

each time. Then from these two sets of L KL divergences we could calculate

$$\mathrm{E}\left[D_{KL}\left[p(m|d) \,||\, \hat{p}_I(m|d)\right]\right] \text{ and } \mathrm{E}\left[D_{KL}\left[p(m|d) \,||\, \hat{p}_D(m|d)\right]\right].$$

The number of estimates of the posterior we made in each case was $L = 1 \times 10^4$, whilst the number of samples made in each method was chosen to be $N = 10$. The analytical quantities (equations C.17 to C.23) can be calculated without any actual sampling. However, they do still require that the number of samples be specified. Thus when calculating these we chose $N = 10$ in both direct and indirect estimation for consistency.

It is clear that the relative properties of the old posterior (that is, the likelihood) and the new prior may effect the quality of the approximation derived by each method. Thus we do not calculate the quantities described above for just a single set of new and old posteriors; instead we vary these distributions systematically. Thus we repeated the above whilst varying the likelihood's parameters (the prior was kept constant since we are only interested in investigating the effect of the relative relationship of new prior and likelihood). We first investigated the effect of μ_L in isolation. To do this μ_L was varied and Σ_L kept constant. Secondly, we investigated the effect of Σ_L in isolation, by varying Σ_L and keeping μ_L constant. The results are described below.

Appendix C.4. Results

Firstly we varied μ_L in the range $[0, 4]$ at intervals of 0.1. The variance of the likelihood was kept constant at $\Sigma_L = 0.5$. The prior distribution was fixed with $\mu_B = 2$ and $\Sigma_B = 0.75$. This defined 41 different new posterior distributions, two examples of which are plotted in Figure C1(a) with the prior and likelihood distributions. The approximate expected Kullback-Leibler divergences for each of these scenarios for both methods, $\mathrm{E}\left[D_{KL}\left[p(m|d)\,||\,\hat{p}_I(m|d)\right]\right]$ and $\mathrm{E}\left[D_{KL}\left[p(m|d)\,||\,\hat{p}_D(m|d)\right]\right]$ are plotted in Figure C1(b). The analytically calculated variance and bias of the estimators ($\hat{\Sigma}_{P_D}$, $\hat{\mu}_{P_D}$, $\hat{\Sigma}_{P_I}$ and $\hat{\mu}_{P_I}$) for both methods are plotted for comparison in Figure C1 (c)-(f).

We then carried out exactly the same procedure except varying Σ_L rather than μ_L. Σ_L was varied in the range [0 4] at intervals of 0.1. The mean of the likelihood was kept constant at $\mu_L = 2$. The prior in this case had parameters $\mu_B = 4$ and $\Sigma_B = 1$. Again this defined 41 different new posterior distributions, two of which are plotted in Figure C2(a) with the prior and likelihood distributions. The approximate expected Kullback-Leibler divergences for each of these scenarios, $\mathrm{E}\left[D_{KL}\left[p(m|d)\,||\,\hat{p}_I(m|d)\right]\right]$ and $\mathrm{E}\left[D_{KL}\left[p(m|d)\,||\,\hat{p}_D(m|d)\right]\right]$ are plotted in Figure C2(b). The analytically calculated variance and bias of the estimators ($\hat{\Sigma}_{P_D}$, $\hat{\mu}_{P_D}$, $\hat{\Sigma}_{P_I}$ and $\hat{\mu}_{P_I}$) are plotted for comparison in Figure C2 (c)-(f).

Appendix C.5. Interpretation of quality comparison results

We can make useful observations about the relative values of the variance and bias of the estimators $\hat{\Sigma}_{P_D}$, $\hat{\mu}_{P_D}$, $\hat{\Sigma}_{P_I}$ and $\hat{\mu}_{P_I}$ from their analytical expressions in equations C.17-C.19 and C.20-C.23 and the results in Figures C1 and C2.

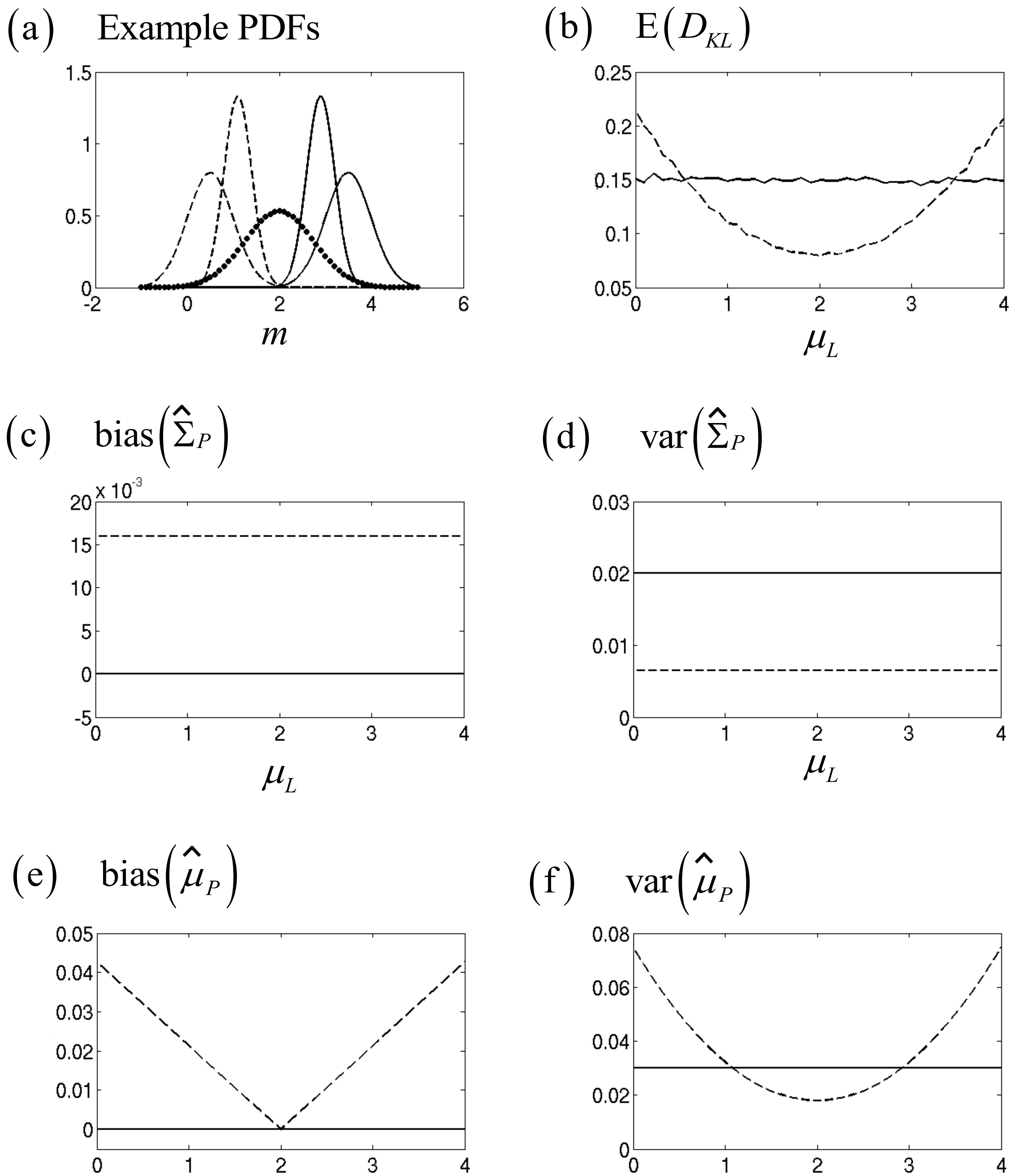

Figure C1. Measures of the quality of the posterior estimate obtained using direct and indirect estimation were calculated for a range of posteriors, defined by: $\mu_L \in \{0, 0.1, ..., 4\}$ whilst $\mu_B = 2$, $\Sigma_B = 0.75$ and $\Sigma_L = 0.5$. (a) the old (ϕ_L) and new posterior (ϕ_P) PDF pairs for $\mu_L = 0.5$ (dashed lines) and $\mu_L = 3.5$ (solid lines). The prior PDF (ϕ_B) is plotted as a dotted bold line. (b) Average Kullback-Leibler divergences for the two methods: $\mathrm{E}\left[D_{KL}\left[p(m|d)||\hat{p}_I(m|d)\right]\right]$ and $\mathrm{E}\left[D_{KL}\left[p(m|d)||\hat{p}_D(m|d)\right]\right]$. (c) bias and (d) variance of $\hat{\Sigma}_{P_D}$ and $\hat{\Sigma}_{P_I}$. (e) bias and (f) variance of $\hat{\mu}_{P_D}$ and $\hat{\mu}_{P_I}$. In plots (b) to (f), solid lines are results obtained for the direct estimation posterior estimate (i.e., P_D), and dashed lines are for the indirect posterior estimate (i.e., P_I).

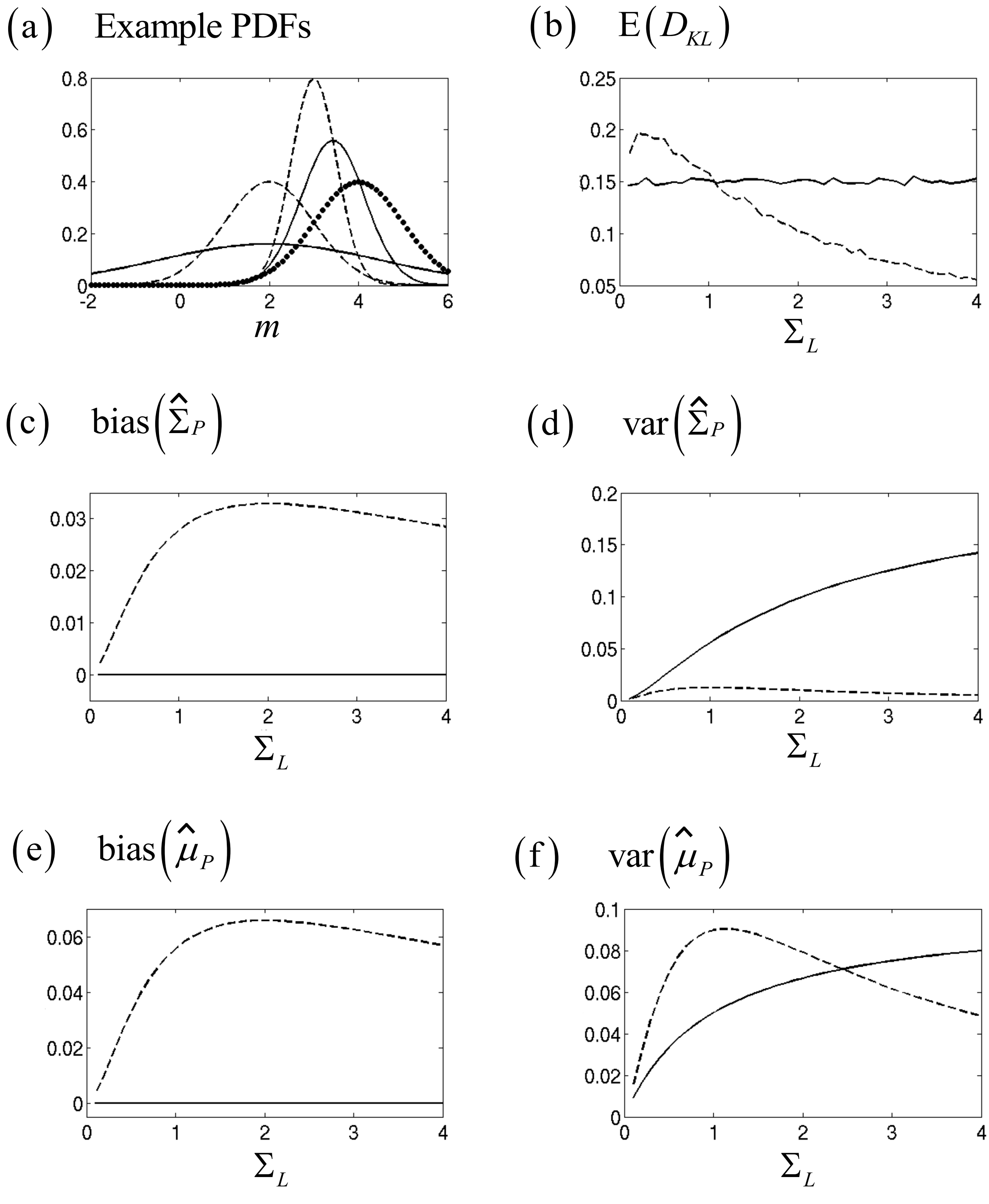

Figure C2. Measures of the quality of the posterior estimate obtained using direct and indirect estimation were calculated for a range of posteriors, defined by: $\Sigma_L \in \{0, 0.1, ..., 4\}$ whilst $\mu_B = 4$, $\Sigma_B = 1$ and $\mu_L = 2$. (a) the old (ϕ_L) and new posterior (ϕ_P) PDF pairs for $\Sigma_L = 1$ (dashed lines) and $\Sigma_L = 2.5$ (solid lines). The prior PDF (ϕ_B) is plotted as a dotted bold line. (b) Average Kullback-Leibler divergences for the two methods: $\mathrm{E}\left[D_{KL}\left[p(m|d)||\hat{p}_I(m|d)\right]\right]$ and $\mathrm{E}\left[D_{KL}\left[p(m|d)||\hat{p}_D(m|d)\right]\right]$. (c) bias and (d) variance of $\hat{\Sigma}_{P_D}$ and $\hat{\Sigma}_{P_I}$. (e) bias and (f) variance of $\hat{\mu}_{P_D}$ and $\hat{\mu}_{P_I}$. In plots (b)-(f) solid lines are results obtained for the direct estimation posterior estimate (i.e., P_D), and dashed lines are for the indirect posterior estimate (i.e., P_I).

When comparing equation C.17 to equation C.20, we see that $|\text{bias}\left(\hat{\mu}_{P_I}\right)| > |\text{bias}\left(\hat{\mu}_{P_D}\right)|$. However, $\text{bias}\left(\hat{\mu}_{P_I}\right)$ will be zero in two non-trivial cases: where either (i) $\mu_P = \mu_L$, or (ii) $\Sigma_P = \Sigma_L$. From equation C.4 we can see that the former case implies that $\mu_B = \mu_L$, and that the latter case implies $\Sigma_B = \infty$ (i.e., the prior is flat). In Figures C1(e) and C2(e) we observe those estimators behaving in this way.

Inspecting equation equation C.18 and C.22 we see that it is possible that $\text{var}\left(\hat{\mu}_{P_I}\right) < \text{var}\left(\hat{\mu}_{P_D}\right)$, and that this will tend to be the case where (i) $\mu_P \to \mu_L$ or (ii) $\Sigma_L >> \Sigma_P$. Again from equation C.4 we can see that the former case implies that $\mu_B \to \mu_L$, and that the latter case implies $\Sigma_L >> \Sigma_B$. We can observe this behaviour in Figures C1(f) and C2(f).

Similarly, we see that generally $\text{bias}\left(\hat{\Sigma}_{P_I}\right) > \text{bias}\left(\hat{\Sigma}_{P_D}\right)$, and the only non-trivial exception to this is when $\Sigma_L = \Sigma_P$, where $\text{bias}\left(\hat{\Sigma}_{P_I}\right) = 0$. Such behaviour can be observed in Figures C1(c) and C2(c). However, a more useful observation can be made about the variance of the variance estimator, by beginning with the observation that multiplication of two Gaussians always yields a Gaussian with lower variance than either of the two Gaussians which were multiplied together (this can be seen immediately from equation C.4). This implies (given the Gaussian multiplication in equation C.3) that $\Sigma_P < \Sigma_L$. Therefore dividing equation C.23 by equation C.19 we find that

$$\frac{\text{var}\left(\hat{\Sigma}_{P_I}\right)}{\text{var}\left(\hat{\Sigma}_{P_D}\right)} = \frac{\Sigma_P^2}{\Sigma_L^2} < 1. \tag{C.26}$$

Thus to third order the variance on the variance estimator in the indirect estimation method is always less than that in the direct estimation (i.e., $\text{var}\left(\hat{\Sigma}_{P_I}\right) < \text{var}\left(\hat{\Sigma}_{P_D}\right)$). Such behaviour can be observed in Figures C1(d) and C2(d).

The curves corresponding to the indirect estimators in Figure C2(c), (d), (e) and (f) all show similar features. All increase (relatively) rapidly from zero at $\Sigma_L = 0$ to reach a maximum approximately where $\Sigma_B = \Sigma_L$ and then decrease (relatively slowly) as $\Sigma_L \to \infty$. This effect can be understood, equally well for the bias and the variance, if we consider two end-member examples. The first is when the prior has infinite variance (it is 'flat') and the likelihood has zero variance (it is a delta function). No error (which would give rise to bias or variance) can be made when sampling from the old posterior (i.e., the likelihood), and we obtain a perfect posterior upon applying Bayes' rule. The second end member case is when the likelihood is flat and the prior is a delta function. In this case errors can be made when sampling the likelihood, but they are irrelevant since the prior (which we multiply by in Bayes' rule) is a delta function, and again we obtain a perfect posterior. The variance and bias of the estimators must go to zero at these end members (which correspond to either end of the horizontal axes). Between these two end members two processes compete: (i) as the likelihood variance decreases (relative to the prior variance) fewer errors occur in sampling, and (ii) as the likelihood variance increases (relative to the prior variance) these errors matter less. Thus one might expect these two competing effects to balance around the point at which the variances are equal (which is what we observe at the maxima where $\Sigma_L = 1 = \Sigma_B$).

In Figure C1(f) we observe that $\mathrm{var}\left(\hat{\mu}_{P_I}\right)$ tends to be lower than $\mathrm{var}\left(\hat{\mu}_{P_D}\right)$ where the likelihood mean approaches the prior mean. This makes intuitive sense since indirect estimation makes an unbiased estimate of the mean of the likelihood. Equation C.4 shows that as the likelihood mean approaches the prior mean, the posterior mean approaches the likelihood (and prior) mean. Thus the indirect estimate approaches a point at which it is making a direct and unbiased estimate of the posterior mean. Similarly the same mechanisms can used to explain the behaviour of $\mathrm{bias}\left(\hat{\mu}_{P_I}\right)$ in Figure C1(e). Here we see that $\mathrm{bias}\left(\hat{\mu}_{P_I}\right)$ goes (linearly) to zero when the likelihood and prior means are equal.

As a consequence of the behaviour of the estimators described above, the overall goodness-of-fit measure, D_{KL}, tends to be lower in the indirect estimation (prior replacement) than in the direct estimation method whenever the likelihood variance is relatively large (compared to the prior variance) and/or the likelihood mean approaches the prior mean. This behaviour can be seen in Figures C1(b) and C2(b). Although D_{KL} is a useful, well-understood measure, it is of limited analytical use here as it cannot easily be related to the parameters of the Gaussian distributions (in the indirect estimation method). However, it neatly encapsulates the other results derived above for the biases and variances.

We found (in results not reproduced here) that varying N had little impact on the relative properties of direct and indirect estimation. N simply acts as a scaling factor in equations C.17-C.19 and C.20-C.23 (thus the variances, biases and D_{KL} all reduced with increasing N in both methods). It should also be noted that the choice of a maximum likelihood estimator here is somewhat at odds with the Bayesian framework used thus far (Ulrych et al. 2001). However we do not anticipate that attaching prior distributions to the parameters (the variance and means) could change the outcome of the analysis. For example, we could choose to use a Bayesian estimator such as the maximum a posteriori

(MAP) estimator for the variance assuming a Jeffrey's prior (Lupton 1993, Jeffreys 1998) but we would not see any practical difference in the analytical results since this would simply change N to $N + 1$ in the expressions above (Ulrych et al. 2001).

In the next section we discuss the implications of these results for the application of prior replacement in MDN inversion, and possible general implications for Bayesian inversion. Thus we reiterate here that direct estimation is equivalent to prior-specific training since no old posterior is used: samples are made directly from the new posterior. Also indirect estimation is equivalent to prior replacement since samples are initially made from the normalised likelihood distribution (equivalent to the old posterior with a flat old prior); then the old prior is replaced by the new prior analytically.

Appendix C.6. Discussion

It is important to note that in the results above we have assumed that we have a likelihood distribution for which we only know the unnormalised density; thus we may only estimate an old or new posterior density by first sampling from it. In contrast, any prior distribution we use (whether it be the old or the new) is assumed to be known parametrically, thus we may manipulate it algebraically with respect to the estimate of the old posterior. In principle prior replacement may be performed even if the old and/or new prior is not known analytically, but the results we have obtained for this investigation of the quality of the approximation in either case would not be relevant. This is because the results assume that the Gaussian new prior is known exactly, and therefore the mean and the variance of the new prior are not random variables in our formulation. However, prior information is very often specified parametrically in geophysical inverse problems. For example, spatial correlation is often specified using Gaussian Markov random fields (Dong et al. 1997, Rue & Held 2005, Eidsvik et al. 2012, Sun et al. 2012). Hence this is not a major practical limitation to the significance of these results.

If we assume that these results relating prior replacement and quality of the final posterior estimate are applicable not just for single Gaussians but for GMMs then we can explain qualitatively the results observed when prior replacement is applied to the results of MDN inversion (compared to the results of prior-specific training). In Figure 3 we saw that a low probability lobe was better resolved by prior replacement than by prior-specific training. In that case the old posterior in Figure 1 (equivalent to the likelihood in the results above) had higher variance than the new prior in Figure

3(a), but had a similar mean. Thus from the estimation quality results we expect that if the old posterior variance is sufficiently large and the means sufficiently similar, that not only would the new posterior variance be more certain but also less biased in the prior replacement result (indirect estimation) than in prior-specific training (direct estimation) result. However, we would also expect that the mean would be more biased and uncertain when using prior replacement since the means of the old posterior and new prior are not identical. Thus the overall shape of the new posterior distribution should be better resolved at the expense of the exact shape of the high probability density area(s). In Figure 3(c) and Figure 3(d) this is what we observe: the peak is less well defined but the low probability lobe is much better defined when using prior replacement.

The results of the investigation into estimation quality may have further implications. For example, suppose that we were performing such an inversion with Gaussians: can we predict a-priori whether the solution quality will be better if we do direct estimation or indirect estimation (prior replacement)? This depends on what aspect of the quality of the posterior estimate is desirable. The bias of the sample mean and variance is always lower for direct estimation than indirect estimation, but the variance of the sample variance is always less in the latter. Which method yields lower variance on the sample mean depends upon the posterior, thus we cannot predict this a-priori. Similarly, we cannot predict which method will yield the smallest Kullback-Leibler divergence without calculating the posterior distribution's parameters.

Of course if we were to ask which method is better for a realistic inversion for a non-Gaussian likelihood we cannot conclude anything definitive from our results. They do support the intuitive supposition that prior replacement would yield more biased results than direct estimation. However, they also show that prior replacement can yield lower variance estimators and better overall goodness-of-fit (Kullback-Liebler divergence) for the Gaussian case. Thus for the general case it is not obvious which method to choose if these criteria are deemed to be important. In cases where the likelihood is not known analytically and we must use sampling methods, this conundrum would be useful to resolve, especially if only a limited number of samples can be made, such as in tomography problems in geophysics (Zhang et al. 2013). To our knowledge this is the first time that this issue has been raised in the literature; it should be investigated in future studies, as it may allow us to perform some Bayesian inversions more efficiently.

There is clearly similarity between prior replacement and the well-known Monte-Carlo technique of importance sampling. Importance sampling transforms samples made from a sampling (so-called 'instrumental') distribution such that they may be used to estimate the properties of another (so-called 'target') distribution. Each individual sample made from the instrumental distribution is transformed by weighting it by the ratio of its probability evaluated using the target distribution, to its probability evaluated using the instrumental distribution. Then calculation of the estimator is made using these transformed samples. As we have demonstrated for prior replacement, importance sampling can be used as a variance reduction technique. To do this the instrumental distribution should be chosen such that samples are made more frequently if they are (somehow) more important to the estimate required of the target distribution (compared to simply sampling directly from the target distribution). A trivial example is when attempting to estimate the mean of a target distribution. In this case if we choose an instrumental distribution which is non-zero only at the target distribution's mean value then this makes the mean estimator's variance zero (when such samples are used to estimate the mean after multiplication with the appropriate weights).

Given the above definition of importance sampling, one might expect it to yield similar, perhaps identical, results to prior replacement if the instrumental distribution is made equal to the old posterior as used in prior replacement. We now explore this hypothesis in the context of the Gaussian posterior estimation problem used above to compare direct and indirect sampling (i.e., prior replacement). We do this by describing in more detail the importance sampling method for estimating μ_P (i.e., the mean of $p_{new}(m|d) = \phi(m; \mu_P, \Sigma_P)$).

As stated above we use as the instrumental distribution the old posterior distribution which, as defined earlier, is simply the normalised likelihood $p_{old}(m|d) = \phi(m; \mu_L, \Sigma_L)$. To estimate μ_P using importance sampling we begin by making N samples of m, $M_1, ..., M_i., ..., M_N$, from the instrumental distribution, where

$$M_i \sim p_{old}(m|d). \tag{C.27}$$

Then a weight value is calculated for each of these samples using

$$w_i = \frac{p_{new}(m = M_i|d)}{p_{old}(m = M_i|d)}. \tag{C.28}$$

Using these weights, the normalised importance sampling method (Bishop 2006, p.533),

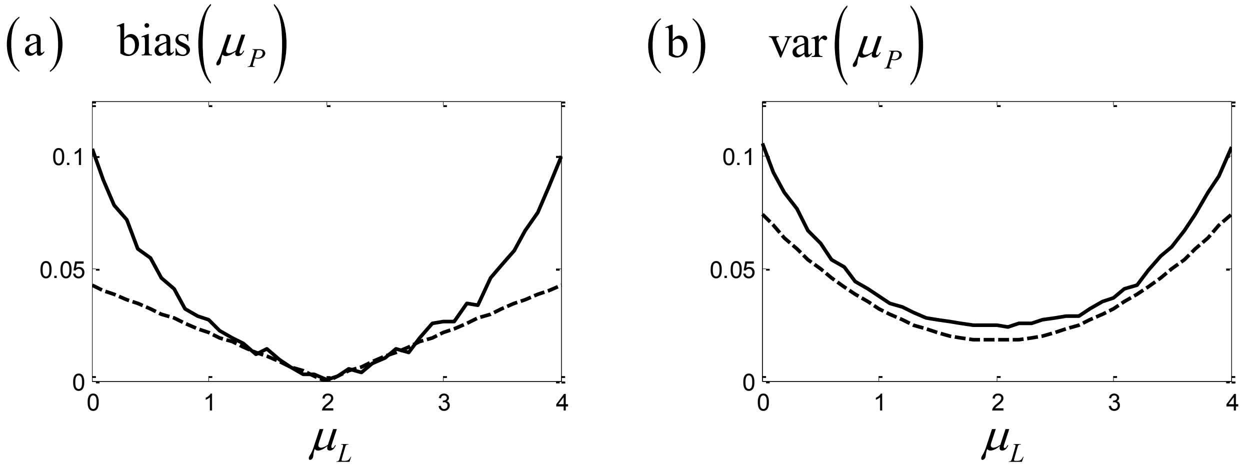

Figure C3. An empirical comparison of prior replacement and importance sampling. The results for indirect estimation (prior replacement) have been replicated from Figure C1 (dotted line): (a) bias and (b) variance of $\hat{\mu}_{P_I}$ for a range of posteriors, defined by $\mu_L \in \{0, 0.1, ..., 4\}$ whilst $\mu_B = 2$, $\Sigma_B = 0.75$ and $\Sigma_L = 0.5$. The equivalent results for the importance sampling estimator $\hat{\mu}_{P_{IS}}$ have been superimposed (solid line), where the old posterior (that used in indirect estimation) has been used as instrumental distribution.

gives the importance sampling estimator of μ_P as

$$\hat{\mu}_{P_{IS}} = \frac{1}{\sum_{i=1}^{N} w_i} \sum_{i=1}^{N} w_i M_i \tag{C.29}$$

where the IS subscript denotes the importance sampling estimate. Although the mathematical procedure of importance sampling (equations C.27 to C.29) is similar to the equivalent prior replacement operation (equations C.10 to C.14) there is a clear difference: prior replacement acts only upon the estimated mean of the old posterior to obtain the new posterior mean, whereas importance sampling acts (by applying a weighting factor) to each sample and then uses this to obtain the new posterior mean estimate. More succinctly, importance sampling transforms individual samples for later use in estimation whereas prior replacement acts to transform distributions directly.

In Figure C3(a) and (b) we demonstrate the empirical affect of this difference between the two methods. To do this we reproduce the results obtained for the variance and bias of the prior replacement (indirect) estimate of μ_P given in Figure C1(e) and (f) for varying values of μ_L. We compare these to the equivalent variance and bias of the $\hat{\mu}_{P_{IS}}$ estimate acquired using importance sampling using the old posterior, $p_{old}(m|d) = \phi(m; \mu_L, \Sigma_L)$, as the instrumental distribution.

There are significant differences between the variance and bias when using prior replacement and importance sampling. In general, prior replacement yields lower bias and variance. However, the overall behaviour of the bias and variance with respect to the change in μ_L is similar (i.e., the shape of the curves is similar). This suggests that we can use the same intuitive interpretation of importance sampling to understand prior replacement in terms of the importance of certain sample values in determining the required estimate, the only difference being in the way that these samples are used to obtain the final estimate. It should be noted that although this comparison has been made only for estimates of μ_P, similar results exist for the Σ_P estimators. We have omitted these for the sake of brevity since the derivation of importance sampling for the variance estimator is not as easily exposited as that for the mean. Also we have not investigated comparison to other possible implementations of importance sampling such as the non-normalised importance sampling scheme (Bishop 2006, p.533).

Appendix C.7. Summary

We have derived approximations for the variance and bias of estimators using prior replacement (termed indirect estimation) and compared these to sampling directly from the corresponding posterior distribution (termed direct estimation) for Gaussian prior and likelihood. Indirect estimation can outperform direct estimation when prior and likelihood have sufficiently similar means, or when the likelihood has a sufficiently large variance compared to the prior variance. Similar results were observed for the expected Kullback-Leibler divergence in each case. These results not only support our proposed use of prior replacement as a useful method for enhancing MDN training, but also highlighted possible benefits of using prior replacement rather than direct estimation in a variety of other situations where sampling is required to determine a posterior distribution. A mathematical comparison of prior replacement and the well-known Monte-Carlo technique of importance sampling was made. They were shown to be quite distinct: the former is applied to distributions, the latter to individual samples. However, empirical studies showed some similarities between results obtained with both methods suggesting that they are indeed related.

Appendix D. Bias and variance of the estimators

Appendix D.1. Preliminaries

The required quantities for the indirect estimators, $\hat{\mu}_{P_I}$ and $\hat{\Sigma}_{P_I}$, are the bias of the mean, bias $(\hat{\mu}_{P_I})$; the variance of the mean, var $(\hat{\mu}_{P_I})$; the bias of the variance, bias $\left(\hat{\Sigma}_{P_I}\right)$; and the variance of the variance, var $\left(\hat{\Sigma}_{P_I}\right)$. $\hat{\mu}_{P_I}$ and $\hat{\Sigma}_{P_I}$ are functions of the random variables $\hat{\mu}_L$ and $\hat{\Sigma}_L$ (see equation C.14). Thus, in order to estimate the required quantities we will need to be able to approximate the expectation of a function of these random variables. Generally, if we have a function of two random variables, $f\left(\hat{\Sigma}_L, \hat{\mu}_L\right)$, then we can obtain an estimate of the expected value of that function, $\mathrm{E}\left[f\left(\hat{\Sigma}_L, \hat{\mu}_L\right)\right]$, by using a Taylor expansion expanded around the expected value of those variables, $\boldsymbol{\theta} = \left[\mathrm{E}\left[\hat{\Sigma}_L\right], \mathrm{E}\left[\hat{\mu}_L\right]\right]$. The third order Taylor expansion approximation of $\mathrm{E}\left[f\left(\hat{\Sigma}_L, \hat{\mu}_L\right)\right]$ is given by Van der Vaart (2000) as

$$\begin{aligned}\mathrm{E}\left[f\left(\hat{\Sigma}_L, \hat{\mu}_L\right)\right] \approx f(\boldsymbol{\theta}) + \frac{1}{2}\frac{d^2 f(\boldsymbol{\theta})}{d\hat{\Sigma}_L^2}\mathrm{var}\left(\hat{\Sigma}_L\right) + \frac{1}{2}\frac{d^2 f(\boldsymbol{\theta})}{d\hat{\mu}_L^2}\mathrm{var}\left(\hat{\mu}_L\right)\\ + \frac{d^2 f(\boldsymbol{\theta})}{d\hat{\Sigma}_L d\hat{\mu}_L}\mathrm{covar}\left(\hat{\Sigma}_L, \hat{\mu}_L\right). \qquad \text{(D.1)}\end{aligned}$$

Since $\hat{\mu}_L$ and $\hat{\Sigma}_L$ are estimators for a Gaussian distribution (equation C.9) we have the elementary results

$$\mathrm{E}\left[\hat{\Sigma}_L\right] = \Sigma_L, \qquad \text{(D.2)}$$

$$\mathrm{E}\left[\hat{\mu}_L\right] = \mu_L, \qquad \text{(D.3)}$$

$$\mathrm{var}\left(\hat{\Sigma}_L\right) = \frac{2\Sigma_L^2}{N-1}, \text{ and} \qquad \text{(D.4)}$$

$$\mathrm{var}\left(\hat{\mu}_L\right) = \frac{\Sigma_L}{N}, \qquad \text{(D.5)}$$

where N is the number of samples made from the old posterior. The sample mean and sample variance are independent (Riley et al. 2006, p.1230), consequently

$$\mathrm{covar}\left(\hat{\Sigma}_L, \hat{\mu}_L\right) = 0 \qquad \text{(D.6)}$$

and we may disregard this term henceforth in the Taylor expansion.

Appendix D.2. Bias of the indirect mean

The indirect mean estimator is a function of two random variables which we write as

$$\hat{\mu}_{P_I} = \left(\Sigma_B^{-1} + \hat{\Sigma}_L^{-1}\right)^{-1}\left(\Sigma_B^{-1}\mu_B + \hat{\Sigma}_L^{-1}\hat{\mu}_L\right) = f\left(\hat{\Sigma}_L, \hat{\mu}_L\right). \qquad \text{(D.7)}$$

We need to estimate its expected value using the Taylor expansion such that we can estimate the bias. To do this we must first determine the derivatives. The first order derivatives of this function, f, are

$$\frac{df}{d\hat{\Sigma}_L} = \hat{\Sigma}_L^{-2}\left(\Sigma_B^{-1} + \hat{\Sigma}_L^{-1}\right)^{-2}\left(\Sigma_B^{-1}\mu_B + \hat{\Sigma}_L^{-1}\hat{\mu}_L\right) - \hat{\Sigma}_L^{-2}\mu_L\left(\Sigma_B^{-1} + \hat{\Sigma}_L^{-1}\right)^{-1}, \tag{D.8}$$

and

$$\frac{df}{d\hat{\mu}_L} = \left(\Sigma_B^{-1} + \hat{\Sigma}_L^{-1}\right)^{-1}\hat{\Sigma}_L^{-1}. \tag{D.9}$$

Thus the required second order derivatives are

$$\begin{aligned}\frac{d^2 f}{d\hat{\Sigma}_L^2} = & -2\hat{\Sigma}_L^{-3}\left(\Sigma_B^{-1} + \hat{\Sigma}_L^{-1}\right)^{-2}\left(\Sigma_B^{-1}\mu_B + \hat{\Sigma}_L^{-1}\hat{\mu}_L\right) \\ & + 2\hat{\Sigma}_L^{-4}\left(\Sigma_B^{-1} + \hat{\Sigma}_L^{-1}\right)^{-3}\left(\Sigma_B^{-1}\mu_B + \hat{\Sigma}_L^{-1}\hat{\mu}_L\right) \\ & + 2\hat{\Sigma}_L^{-3}\hat{\mu}_L\left(\Sigma_B^{-1} + \hat{\Sigma}_L^{-1}\right)^{-1} - \hat{\Sigma}_L^{-4}\hat{\mu}_L\left(\Sigma_B^{-1} + \hat{\Sigma}_L^{-1}\right)^{-2}\end{aligned} \tag{D.10}$$

and

$$\frac{d^2 f}{d\hat{\mu}_L^2} = 0. \tag{D.11}$$

Substituting the mean vector, $\boldsymbol{\theta} = \left[\mathrm{E}\left[\hat{\Sigma}_L\right], \mathrm{E}\left[\hat{\mu}_L\right]\right] = [\Sigma_L, \mu_L]$, and D.11 into D.1 we obtain

$$\begin{aligned}\mathrm{E}\left[\hat{\mu}_{P_I}\right] &= \mathrm{E}\left[f\left(\hat{\Sigma}_L, \hat{\mu}_L\right)\right] \\ &\approx f\left(\Sigma_L, \mu_L\right) + \frac{1}{2}\frac{d^2 f\left(\Sigma_L, \mu_L\right)}{d\hat{\Sigma}_L^2}\mathrm{var}\left(\hat{\Sigma}_L\right).\end{aligned} \tag{D.12}$$

Then substituting D.10 into this we obtain

$$\begin{aligned}E\left[\hat{\mu}_{P_I}\right] \approx & \left(\Sigma_B^{-1} + \Sigma_L^{-1}\right)^{-1}\left(\Sigma_B^{-1}\mu_B + \Sigma_L^{-1}\mu_L\right) \\ & + \frac{2\Sigma_L^2}{N-1}\Big(- \Sigma_L^{-3}\left(\Sigma_B^{-1} + \Sigma_L^{-1}\right)^{-2}\left(\Sigma_B^{-1}\mu_B + \Sigma_L^{-1}\mu_L\right) \\ & + \Sigma_L^{-4}\left(\Sigma_B^{-1} + \Sigma_L^{-1}\right)^{-3}\left(\Sigma_B^{-1}\mu_B + \Sigma_L^{-1}\mu_L\right) \\ & + \Sigma_L^{-3}\mu_L\left(\Sigma_B^{-1} + \Sigma_L^{-1}\right)^{-1} - \Sigma_L^{-4}\mu_L\left(\Sigma_B^{-1} + \Sigma_L^{-1}\right)^{-2}\Big).\end{aligned} \tag{D.13}$$

Noting that the posterior mean may be written

$$\mu_P = f\left(\Sigma_L, \mu_L\right) = \left(\Sigma_B^{-1} + \Sigma_L^{-1}\right)^{-1}\left(\Sigma_B^{-1}\mu_B + \Sigma_L^{-1}\mu_L\right), \tag{D.14}$$

and the posterior variance as

$$\Sigma_P = \left(\Sigma_B^{-1} + \Sigma_L^{-1}\right)^{-1} \tag{D.15}$$

we may then write the expected mean as

$$\begin{aligned}\mathrm{E}\left[\hat{\mu}_{P_I}\right] &\approx \mu_P + \frac{1}{N-1}\left(-\Sigma_L^{-1}\Sigma_P\mu_P + \Sigma_L^{-2}\Sigma_P^2\mu_P + \Sigma_L^{-1}\Sigma_P\mu_L - \Sigma_L^{-2}\mu_L\Sigma_P^2\right)\\ &= \mu_P + \frac{1}{N-1}\left(\mu_L - \mu_P\right)\left(\Sigma_L^{-1}\Sigma_P - \Sigma_L^{-2}\Sigma_P^2\right). \qquad \text{(D.16)}\end{aligned}$$

Therefore the bias may be approximated, using its definition, as

$$\begin{aligned}\mathrm{bias}\left(\hat{\mu}_{P_I}\right) &= \mathrm{E}\left[\hat{\mu}_{P_I}\right] - \mu_P\\ &\approx \frac{1}{N-1}\left(\mu_L - \mu_P\right)\left(\Sigma_L^{-1}\Sigma_P - \Sigma_L^{-2}\Sigma_P^2\right). \qquad \text{(D.17)}\end{aligned}$$

Appendix D.3. Variance of the indirect mean

When calculating the variance we wish to obtain the expected value of the squared difference between the sample mean and expected sample mean. We may write this 'residual' function g as

$$g\left(\hat{\Sigma}_L, \hat{\mu}_L\right) = \left(\hat{\mu}_{P_I} - \mathrm{E}\left[\hat{\mu}_{P_I}\right]\right)^2 = \left(f\left(\hat{\Sigma}_L, \hat{\mu}_L\right) - \mathrm{E}\left[\hat{\mu}_{P_I}\right]\right)^2. \qquad \text{(D.18)}$$

The expected value of this function is the variance, that is

$$\mathrm{var}\left(\hat{\mu}_{P_I}\right) = \mathrm{E}\left[g\left(\hat{\Sigma}_L, \hat{\mu}_L\right)\right]. \qquad \text{(D.19)}$$

We can use a Taylor expansion to approximate this variance. After calculating derivatives and then following a similar procedure to that in Appendix D.2, we find an approximation for the variance of the mean estimate as

$$\mathrm{var}\left(\hat{\mu}_{P_I}\right) \approx \left(\mu_P - \mu_L\right)^2 \frac{2\Sigma_P^2\Sigma_L^{-2}}{N-1} + \frac{\Sigma_P^2\Sigma_L^{-1}}{N}. \qquad \text{(D.20)}$$

Appendix D.4. The bias of the indirect variance

We can use a similar analysis to that in Appendix D.2 to calculate an approximation for the bias of the variance. We begin with the function which gives the indirect posterior variance estimator,

$$\hat{\Sigma}_{P_I} = \left({\Sigma_B}^{-1} + \hat{\Sigma}_L^{-1}\right)^{-1} = h\left(\hat{\Sigma}_L\right). \qquad \text{(D.21)}$$

Again, we need to estimate its expected value using the Taylor expansion such that we can estimate the bias. After doing this we find an approximation for the bias of the indirect variance as

$$\mathrm{bias}\left(\hat{\Sigma}_{P_I}\right) \approx \Sigma_P - \mathrm{E}\left[\hat{\Sigma}_{P_I}\right] = \frac{2}{N-1}\left(\Sigma_L^{-2}\Sigma_P^3 - \Sigma_L^{-1}\Sigma_P^2\right). \qquad \text{(D.22)}$$

Appendix D.5. Variance of the indirect variance

When calculating the variance we wish to obtain the expected value of the squared difference between sample variance and expected sample variance. We may write this 'residual' function as

$$r\left(\hat{\Sigma}_L\right) = \left(\hat{\Sigma}_{P_I} - \mathrm{E}\left[\hat{\Sigma}_{P_I}\right]\right)^2 = \left(h\left(\hat{\Sigma}_L\right) - \mathrm{E}\left[\hat{\Sigma}_{P_I}\right]\right)^2. \tag{D.23}$$

As previously we obtain the variance by taking the expectation

$$\mathrm{var}\left(\hat{\Sigma}_{P_I}\right) = \mathrm{E}\left[r\left(\hat{\Sigma}_L\right)\right] \tag{D.24}$$

which can be approximated using the Taylor expansion, thus we need to calculate the derivatives of r. After doing this, in a similar manner to that in Appendix D.2, we find an approximation of the variance of the indirect sample variance of the posterior as

$$\mathrm{var}\left(\hat{\Sigma}_{P_I}\right) \approx \frac{2\Sigma_L^{-2}\Sigma_P^4}{N-1}. \tag{D.25}$$