Supplementary Appendix
Hierarchical Bayesian inference for model fitting and comparison

Payam Piray, Amir Dezfouli, Tom Heskes, Michael J. Frank, Nathaniel D. Daw

1 Formal derivations of the HBI algorithm

In this appendix, we provide the proof of the results given in the main text. The proof is given in three parts by obtaining 1) the functional form of \( q(H, Z) \); 2) the posterior \( q(\mu, \tau, m) \) and corresponding update equations; and 3) the update equations for the posterior over latent variables, \( q(H, Z) \).

Recall that the HBI generative model is given by

\[
p(Z|m) = \prod_n \prod_k m_k z_{kn}^2. \tag{1}
\]

\[
p(X|H, Z) = \prod_k \prod_n p(x_n|h_{kn}, M_k z_{kn}), \tag{2}
\]

\[
p(H|Z, \mu, T) = \prod_k \prod_n N(h_{kn}|\mu_k, T_k^{-1}) z_{kn}, \tag{3}
\]

\[
p(m) = \text{Dir}(m|\alpha_0) = C(\alpha_0) \prod_k m_k^{\alpha_0 - 1}, \tag{4}
\]

where \( T_k \) is a diagonal matrix. By defining \( T_k = \text{diag}(\tau_k) \), in which \( \text{diag}(.) \) is an operator outputting a diagonal matrix with elements given by \( \tau_k \), we have:

\[
p(\mu, \tau) = \prod_{k=1}^K N(\mu_k|a_0, \text{diag}(b\tau_k)^{-1}) G(\tau_k|v, s), \tag{5}
\]
where we have defined:

\[ G(\tau_k|v, s) = D_k \prod_{i=1}^{D_k} G(\tau_{ki}|v, s). \]

Thus, the full probabilistic model is given by,

\[ p(X, H, Z, \mu, \tau, m) = p(X|H, Z)p(H|Z, \mu, \tau)p(Z|m)p(\mu|\tau)p(\tau)p(m). \quad (6) \]

1.1 The functional form of the posterior over H and Z

Let us first consider the derivation of the functional form for the factor \( q(H, Z) \). According to standard results in variational inference [15,16], the log of this factor is given by:

\[
\log q(H, Z) = E_{\mu, \tau, m}[\log p(X, H, Z, \mu, \tau, m)] + \text{constant}^{\backslash H, \backslash Z},
\]

where the constant term denotes all the terms independent of the corresponding variables. Note that the expectation is taken with respect to the current estimates of \( q(\mu, \tau, m) \). By using equation (6) and absorbing all the terms which are independent of \( H \) and \( Z \) into the additive constant, we have:

\[
\log q(H, Z) = E_{\mu, \tau}[\log p(X, H|Z, \mu, \tau)] + E_m[\log p(Z|m)] + \text{constant}^{\backslash H, \backslash Z}.
\]

Substituting the two conditional distribution on the right-hand side using equations (1-3), we have:

\[
\log q(H, Z) = \sum_k \sum_n z_{kn} (\log I_{kn} + E_m[\log m_k]) + \text{constant}^{\backslash H, \backslash Z},
\]

where,

\[
\log I_{kn} = E_{\mu, \tau}[\log p(x_n|h_{kn}, M_k)N(h_{kn}|\mu_k, T_k^{-1})].
\]

Note that we have defined \( T_k = \text{diag}(\tau_k) \). We assume that there is a quadratic approximation of \( I_{kn} \) with respect to \( h_{kn} \),

\[
I_{kn} \propto \exp(-\frac{1}{2}(h_{kn} - \theta_{kn})^T A_{kn}(h_{kn} - \theta_{kn})),
\]
which gives,

\[ \log q(H, Z) = \sum_k \sum_n z_{kn} \left( -\frac{1}{2} (h_{kn} - \theta_{kn})^T A_{kn} (h_{kn} - \theta_{kn}) + \mathbb{E}_m [\log m_k] \right) \]

+ constant \( H \), \( Z \).

Since \( \log q(H|Z) = \log q(H, Z) - \log q(Z) \), we can read off terms involving \( H \) in \( \log q(H, Z) \) to obtain \( \log q(H|Z) \):

\[ \log q(H|Z) = \sum_k \sum_n z_{kn} \left( -\frac{1}{2} (h_{kn} - \theta_{kn})^T A_{kn} (h_{kn} - \theta_{kn}) \right) + \text{constant} \ H. \]

Requiring that this distribution should be normalized, we obtain:

\[ q(H|Z) = \prod_k \prod_n N(h_{kn}|\theta_{kn}, A_{kn}^{-1})^{z_{kn}}. \quad (7) \]

Subtracting \( \log q(H|Z) \) from \( \log q(H, Z) \) cancels out the quadratic component and yields \( \log q(Z) \), which is a linear function with respect to \( z_{kn} \). Therefore, we have:

\[ q(Z) = \prod_k \prod_n r_{kn}^{z_{kn}}. \quad (8) \]

The functional form of \( q(H, Z) \) is then given by,

\[ q(H, Z) = \prod_k \prod_n r_{kn}^{z_{kn}} N(h_{kn}|\theta_{kn}, A_{kn}^{-1})^{z_{kn}}. \quad (9) \]

Here our goal was to obtain the functional form of the posterior over latent variables. We will obtain values of \( r_{kn}, \theta_{kn} \) and \( A_{kn} \) in section 1.3.

### 1.2 The posterior over \( \mu, \tau \) and \( m \)

We continue with obtaining the functional form and update equations for the other variational factor \( q(\mu, \tau, m) \). The posterior of \( m \) is independent from the posterior over \( \mu \) and \( \tau \) because the log-posterior decomposes into the terms that
only depend on \( m \) and terms that only depend on \( \mu \) and \( \tau \):

\[
\log q(\mu, \tau, m) = \mathbb{E}_{H, Z}[\log p(X, H, Z, \mu, \tau, m)] + \text{constant}^{\mu, \tau, m} = \mathbb{E}_{H, Z}[\log p(H|Z, \mu, \tau) + \log p(\mu, \tau)] + \mathbb{E}_Z[\log p(Z|m) + p(m)] + \text{constant}^{\mu, \tau, m},
\]

where we have used equation (6). This implies that the variational posterior \( q(\mu, T, m) \) factorizes to give \( q(\mu, T)q(m) \). Thus, the posterior over \( \mu \) and \( \tau \) is given by:

\[
\log q(\mu, \tau) = \mathbb{E}_{H, Z}[\log p(H|Z, \mu, \tau)] + \log p(\mu, \tau) + \text{constant}^{\mu, \tau},
\]

in which we absorbed any terms independent of \( \mu, \tau \) into the additive constant.

Substituting for the distributions on the right-hand side, we have:

\[
\log q(\mu_k) = \sum_n \frac{1}{2} r_{kn} \log |T_k| - \sum_n \frac{1}{2} (\mu_k - \theta_{kn})^\top r_{kn} T_k (\mu_k - \theta_{kn}) + \\
- \sum_n \frac{1}{2} r_{kn} \text{Tr}(A_{kn}^{-1} T_k) + \frac{1}{2} \log |T_k| + \\
- \frac{1}{2} (\mu_k - a_0)^\top b T_k (\mu_k - a_0) + \\
+ \sum_{i=1}^{D_k} \log G(v, s) + (v - 1) \log \tau_{ki} - s \tau_{ki} + \text{constant}^{\mu_k, \tau_k}.
\]

As the right-hand side is quadratic with respect to \( \mu_k \), the posterior over \( \mu_k \) also takes the form of a Gaussian with a variance depending on \( \tau_k \):

\[
q(\mu_k | \tau_k) = N(\mu_k | a_k, (\beta_k T_k)^{-1}),
\]

where

\[
a_k = \frac{1}{N_k + b} \left( \sum_n r_{kn} \theta_{kn} + ba_0 \right)
\]
\[ \beta_k = b + \overline{N}_k, \]

and \( N_k \) is given by:

\[ \overline{N}_k = \sum_n r_{kn}. \]

By subtracting \( \log q(\mu_k | \tau_k) \) from \( \log q(\mu_k, \tau_k) \), we obtain the posterior over \( \tau_k \):

\[ q(\tau_k) = \mathcal{G}(\tau_k | \nu_k, \sigma_k), \]

where

\[ \sigma_k = \frac{1}{2} \sum_n \text{diag}(r_{kn}[(\theta_{kn} - \bar{\theta}_k)(\theta_{kn} - \bar{\theta}_k)^T + A^{-1}_{kn}]) \]
\[ + \frac{1}{2} \frac{b \overline{N}_k}{b + \overline{N}_k} \text{diag}((\bar{\theta}_k - a_0)(\bar{\theta}_k - a_0)^T) + s \]
\[ \nu_k = v + \frac{1}{2} \overline{N}_k, \]

and \( \bar{\theta}_k \) is given by:

\[ \bar{\theta}_k = \frac{1}{N_k} \sum_n r_{kn} \theta_{kn}. \]

Finally, we consider the factor \( q(m) \):

\[ \log q(m) = \mathbb{E}_Z[\log p(Z|m)] + \log p(m) + \text{constant}^m. \]

Substituting for the two distributions on the right-hand side, we have

\[ \log q(m) = \sum_k \sum_n r_{kn} \log m_k + \log C(\alpha_0) + \sum_k (\alpha_0 - 1) \log m_k + \text{constant}^m. \]

Therefore \( q(m) \) takes the form of Dirichlet distribution:

\[ q(m) = \text{Dir}(m|\alpha), \]

where \( \alpha \) has components \( \alpha_k \) given by,

\[ \alpha_k = \alpha_0 + \overline{N}_k. \]
1.3 The posterior over $H$ and $Z$

We have already seen in section 1.1 that $q(H, Z)$ could be written,

$$\log q(H, Z) = \sum_k \sum_n z_{kn} (\log I_{kn} + E_m [\log m_k]) + \text{constant}$$

where,

$$\log I_{kn} = E_{\mu, \tau} [\log p(x_n | h_{kn}, M_k) N(h_{kn} | \mu_k, T_k^{-1})].$$

Since we have already obtained $q(\mu, \tau)$, we can now compute $I_{kn}$:

$$\log I_{kn} = \log p(x_n | h_{kn}, M_k) - \frac{1}{2} D_k \log 2\pi + \frac{1}{2} \mathbb{E} [\log |T_k|]$$

$$- \frac{1}{2} \mathbb{E}_{\mu} [(h_{kn} - \mu_k)^T T_k (h_{kn} - \mu_k)].$$

Therefore, $\log I_{kn}$ is given by:

$$\log I_{kn} = \log p(x_n | h_{kn}, M_k) - \frac{1}{2} D_k \log 2\pi + \frac{1}{2} \mathbb{E} [\log |T_k|]$$

$$- \frac{1}{2} (h_{kn} - a_k)^T \mathbb{E} [T_k] (h_{kn} - a_k) - \frac{1}{2} \mathbb{E} [\text{Tr}(T_k (\beta_k T_k)^{-1})],$$

which can be written in the form,

$$\log I_{kn} = \log p(x_n | h_{kn}, M_k) N(h_{kn} | a_k, \mathbb{E}[T_k]^{-1}) + \lambda_k,$$

where $\lambda_k$ is independent of $h_{kn}$ and is given by

$$\lambda_k = \frac{1}{2} \mathbb{E} [\log |T_k|] - \frac{1}{2} \mathbb{E} [\log |T_k|] - \frac{1}{2} D_k \beta_k,$$

Substituting the moments of $T_k = \text{diag}(\tau_k)$ with their values under $q(\tau_k)$,

$$\log \mathbb{E} [\tau_k] = D_k \log \nu_k - \sum_i \log \sigma_{ki},$$

$$\mathbb{E} [\log \tau_k] = D_k \psi(\nu_k) - \sum_i \log \sigma_{ki},$$

gives

$$\lambda_k = \frac{D_k}{2} (\psi(\nu_k) - \log \nu_k - \frac{1}{\beta_k})$$
Now, we make a quadratic approximation of \( p(x_n|h_{kn}, M_k)N(h_{kn}|a_k, E[T_k]^{-1}) \) with respect to \( h_{kn} \) (for example using Laplace approximation or any other method):

\[
p(x_n|h_{kn}, M_k)N(h_{kn}|a_k, E[T_k]^{-1}) \simeq f_{kn} \exp\left(-\frac{1}{2}(h_{kn} - \theta_{kn})^T A_{kn}(h_{kn} - \theta_{kn})\right).
\]

Substituting this approximation into \( \log I_{kn} \), we obtain

\[
\log I_{kn} = \log f_{kn} - \frac{1}{2}(h_{kn} - \theta_{kn})^T A_{kn}(h_{kn} - \theta_{kn}) + \lambda_k.
\]

Therefore, we have:

\[
\log q(H, Z) = \sum_k \sum_n z_{kn} \left( + \log f_{kn} - \frac{1}{2}(h_{kn} - \theta_{kn})^T A_{kn}(h_{kn} - \theta_{kn}) \\
+ \lambda_k + E[\log m_k] \right) + \text{constant}\{H\setminus\{Z\}.}
\]

Subtracting this equation from \( \log q(H|Z) \) given by the log of equation (7), we have:

\[
\log q(Z) = \sum_k \sum_n z_{kn} \log \rho_{kn} + \text{constant}\{\{H\setminus\{Z}\},
\]

where

\[
\log \rho_{kn} = \log f_{kn} + \frac{1}{2} D_k \log 2\pi - \frac{1}{2} \log |A_{kn}| + \lambda_k + E[\log m_k].
\]

Requiring that \( q(Z) \) be normalized, we obtain equation (8), where

\[
r_{kn} = \frac{\rho_{kn}}{\sum_{j=1}^K \rho_{jn}},
\]

which completes the proof.