Appendix A. MED in details

A.1 General MED for multi-way classification

Let $\mathbf{x} \in \mathbb{R}^K$ be an input feature vector. We consider general multi-way classification, where the response variable y takes value from a finite set $\{1, \cdots, L\}$. Let $F(y, \mathbf{x}; \boldsymbol{\eta})$ be the discriminant function parameterized by $\boldsymbol{\eta}$. MED learns a distribution $q(\boldsymbol{\eta})$ by solving an entropic regularized risk minimization problem under prior $p_0(\boldsymbol{\eta})$

$$\min_{q(\boldsymbol{\eta})} \quad \mathrm{KL}\left(q(\boldsymbol{\eta}) \| p_0(\boldsymbol{\eta})\right) + C\mathcal{R}(q(\boldsymbol{\eta})), \tag{26}$$

where

$$\mathcal{R}(q(\boldsymbol{\eta})) = \sum_d \max_y \left\{ \ell_d^{\Delta}(y) + \mathbb{E}_q[F(y, \mathbf{x}_d; \boldsymbol{\eta}) - F(y_d, \mathbf{x}_d; \boldsymbol{\eta})] \right\} \tag{27}$$

is the hinge-loss on training data $\mathcal{D} = \{(\mathbf{x}_d, y_d)\}_{d=1}^D$, capturing the large-margin principle underlying the MED prediction rule

$$\hat{y} = \mathrm{argmax}_y \, \mathbb{E}_q[F(y, \mathbf{x}; \boldsymbol{\eta})], \tag{28}$$

and $\ell_d^{\Delta}(y)$ measures how y differs from the true label y_d.

A.2 MED under mean-field assumption (for binary case)

The general solution to the MED problem for binary case (5) is

$$p(\boldsymbol{\eta}) = \frac{1}{Z} p_0(\boldsymbol{\eta}) \exp\left\{ \sum_d \omega_d y_d F(\mathbf{x}_d; \boldsymbol{\eta}) \right\}, \tag{29}$$

where Z is the partition function ensuring $p(\boldsymbol{\eta})$ a valid distribution and ω_d are the Lagrange multipliers, which can be obtained by solving the dual problem

$$\begin{aligned} \max_{\boldsymbol{\omega}} \quad & \ell \sum_d \omega_d - \log Z \\ \text{s.t.} \quad & \forall 1 \leq d \leq D \,:\, 0 \leq \omega_d \leq C. \end{aligned} \tag{30}$$

Very often we adopt mean-field assumptions on $p(\boldsymbol{\eta})$, i.e. $p(\boldsymbol{\eta}) = \prod_i p(\boldsymbol{\eta}_i)$ where $\boldsymbol{\eta}_i$s constitute a partition of $\boldsymbol{\eta}$, to seek a variationally approximated solution which is otherwise intractable to get. As a result, problem (5) is *partially* decomposed, one for $p(\boldsymbol{\eta}_i)$ each, as

$$\min_{p(\boldsymbol{\eta}_i)} \quad -H\left(p(\boldsymbol{\eta}_i)\right) - \mathbb{E}_{\boldsymbol{\eta}_i}[\tilde{E}(\boldsymbol{\eta}_i)] + C \sum_d h_\ell \left(y_d \mathbb{E}_{\boldsymbol{\eta}_i}[\tilde{F}_d(\boldsymbol{\eta}_i)] \right), \tag{31}$$

where $\tilde{E}(\boldsymbol{\eta}_i) = \mathbb{E}_{\boldsymbol{\eta}_{-i}}[\log p_0(\boldsymbol{\eta})]$ and $\tilde{F}_d(\boldsymbol{\eta}_i) = \mathbb{E}_{\boldsymbol{\eta}_{-i}}[F(\mathbf{x}_d; \boldsymbol{\eta})]$. Note that the dependence of $\tilde{E}(\boldsymbol{\eta}_i)$ and $\tilde{F}_d(\boldsymbol{\eta}_i)$ on $p(\boldsymbol{\eta}_{-i})$ disqualifies (31) from being a thoroughly independent decomposition.

Since we can always define a distribution $\tilde{p}(\boldsymbol{\eta}_i)$ so that $\log \tilde{p}(\boldsymbol{\eta}_i) = \tilde{E}(\boldsymbol{\eta}_i) + \text{const.}$, we may rewrite subproblem (31) even further as

$$\min_{p(\boldsymbol{\eta}_i)} \quad \mathrm{KL}\left(p(\boldsymbol{\eta}_i) \| \tilde{p}(\boldsymbol{\eta}_i)\right) + C \sum_d h_\ell \left(y_d \mathbb{E}_{\boldsymbol{\eta}_i}[\tilde{F}_d(\boldsymbol{\eta}_i)] \right), \tag{32}$$

whose solution, according to Eq. (29), reads

$$p(\boldsymbol{\eta}_i) = \frac{1}{Z_i} \tilde{p}(\boldsymbol{\eta}_i) \exp\left\{ \sum_d \omega_d Y_d \tilde{F}_d(\boldsymbol{\eta}_i) \right\}, \tag{33}$$

where $\tilde{p}(\boldsymbol{\eta}_i) \propto \exp\{\tilde{E}(\boldsymbol{\eta}_i)\} \propto \exp\{\mathbb{E}_{\boldsymbol{\eta}_{-i}}[\log p_0(\boldsymbol{\eta}_i | \boldsymbol{\eta}_{-i})]\} \propto p_0(\boldsymbol{\eta}_i) \exp\{\mathbb{E}_{\boldsymbol{\eta}_{-i}}[\log p_0(\boldsymbol{\eta}_{-i} | \boldsymbol{\eta}_i)]\}$.

Incidentally, we may *partially* decompose $\mathrm{KL}(p \| p_0)$ as follows,

$$\begin{aligned} & \mathrm{KL}\left(p(\boldsymbol{\eta}) \| p_0(\boldsymbol{\eta})\right) \\ = {} & \mathrm{KL}\left(p(\boldsymbol{\eta}_i) \| \tilde{p}(\boldsymbol{\eta}_i)\right) - \sum_{j \neq i} H\left(p(\boldsymbol{\eta}_j)\right) - \log \int \exp\left\{ \tilde{E}(\boldsymbol{\eta}_i) \right\} d\boldsymbol{\eta}_i \qquad (\forall i) \\ = {} & \mathrm{KL}\left(p(\boldsymbol{\eta}_i) \| \tilde{p}(\boldsymbol{\eta}_i)\right) + \mathrm{KL}\left(p(\boldsymbol{\eta}_{-i}) \| p_0(\boldsymbol{\eta}_{-i})\right) - \log \int \exp\left\{ \mathbb{E}_{\boldsymbol{\eta}_{-i}}[\log p_0(\boldsymbol{\eta}_i | \boldsymbol{\eta}_{-i})] \right\} d\boldsymbol{\eta}_i, \end{aligned} \tag{34}$$

which might help with the calculation of the KL-divergence term, as shown later in D.2.3.

Appendix B. Solving M^{3}F via blockwise coordinate descent

The alternative objective proposed in [11] for ordinal rating is

$$J(U,V,\theta) = \frac{1}{2}\left(\|U\|_F^2 + \|V\|_F^2\right) + C \sum_{ij\in\mathcal{I}} \sum_{r=1}^{L-1} h\left(T_{ij}^r(\theta_{ir} - U_i V_j^\top)\right), \tag{35}$$

which can be efficiently optimized to a local minimum by a gradient descent scheme, the cost of which however is to substitute some *smooth hinge* function for the canonical hinge loss so as to bypass its non-differentiability [11].

Here we propose an alternative efficient blockwise coordinate descent algorithm that directly solves problem (35) and obtains comparable results. It involves $(N+M)$ SVM solvers for optimizing U and V, and $N \times (L-1)$ linear programming solvers for θ, each of which reduces to a binary search as explained later in B.3.

B.1 Optimizing U given V & θ

Observing that $\min_{U|V,\theta} J(U,V,\theta)$ can be decomposed into N independent subproblems, one for U_i each, as

$$\min_{U|V,\theta} J(U,V,\theta) = \sum_{i=1}^{N} \min_{U_i} J(U_i|V,\theta) + \text{const.}, \tag{36}$$

where

$$J(U_i|V,\theta) = \frac{1}{2}\|U_i\|_F^2 + C \sum_{j|ij\in\mathcal{I}} \sum_{r=1}^{L-1} h\left(T_{ij}^r(\theta_{ir} - U_i V_j^\top)\right) \tag{37}$$

and const. denotes the remaining term, which is independent of U_is ($\|V\|_F^2/2$ in this case), we may achieve a considerable speedup by solving these downsized subproblems in a parallel fashion.

And $J(U_i|V,\theta)$, under a slight reformulation, can be efficiently solved by *SVM^{struct}*, an existing high performance structural SVM solver.

The prototype structural SVM problem: The primal objective of the n-slack structural SVM with margin rescaling [6] is given by

$$\begin{aligned} \min_{\mathbf{w},\xi\geq\mathbf{0}} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + \frac{C}{n}\sum_{i=1}^{n}\xi_i \\ \text{s.t.} \quad & \forall i \in \{1,\dots,n\}, \forall \bar{y}_i \in \mathcal{Y} : \mathbf{w}^T[\mathbf{\Psi}(x_i,y_i) - \mathbf{\Psi}(x_i,\bar{y}_i)] \geq \Delta(y_i,\bar{y}_i) - \xi_i, \end{aligned} \tag{38}$$

where $\mathbf{w}$ is the weight vector, $\mathbf{\Psi}(x,y)$ the feature vector relating input data vector x and output y, (x_i, y_i) the training data, and $\Delta(y_i,\bar{y}_i)$ the prediction loss.

The hinge loss thereof is thus $\max_{\bar{y}\in\mathcal{Y}}\{\Delta(y,\bar{y}) - \mathbf{w}^\top\mathbf{\Psi}(x,y) + \mathbf{w}^\top\mathbf{\Psi}(x,\bar{y})\}$, which subsumes the canonical hinge loss $h(x) = \max(1-x,0)$ for binary classification problem $\mathcal{Y} = \{-1,1\}$ when we take $\Delta(y,\bar{y}) = 1 - \delta(y,\bar{y})$ and $\mathbf{\Psi}(x,y) = yx/2$:

$$\begin{aligned} & \max_{\bar{y}\in\{-1,1\}} \left(1 - \delta(y,\bar{y}) - \frac{1}{2}y\mathbf{w}^\top x + \frac{1}{2}\bar{y}\mathbf{w}^\top x\right) \\ = & \max_{\bar{y}\in\{-y,y\}} \left(1 - \delta(y,\bar{y}) - \frac{1}{2}y\mathbf{w}^\top x + \frac{1}{2}\bar{y}\mathbf{w}^\top x\right) \\ = & \max(1 - y\mathbf{w}^\top x, 0). \end{aligned} \tag{39}$$

Reformulation: To better disclose the correspondence, we first rewrite each subproblem $\min_{U_i} J(U_i|V,\theta)$ by introducing slack variables ξ_{ij}^r as follows:

$$\begin{aligned} \min_{U_i,\xi_i\geq\mathbf{0}} \quad & \frac{1}{2}\|U_i\|_F^2 + C \sum_{j|ij\in\mathcal{I}} \sum_{r=1}^{L-1} \xi_{ij}^r \\ \text{s.t.} \quad & \forall j \in \{j \mid ij \in \mathcal{I}\}, \forall r \in \{1,\dots,L-1\} : -T_{ij}^r U_i V_j^\top \geq (1 - T_{ij}^r\theta_{ir}) - \xi_{ij}^r. \end{aligned} \tag{40}$$

Then we may use the similar technique as in (39) to reduce (40) to (38), specifically by taking

$$\begin{cases} \mathbf{w} & = U_i^\top \\ x_{ij}^r & = V_j^\top \\ y_{ij}^r & = -T_{ij}^r \\ \mathcal{Y} & = \{-1, 1\} \\ \mathbf{\Psi}(x, y) & = \frac{1}{2} yx \\ \Delta(y_{ij}^r, \bar{y}_{ij}^r, \theta_{ir}) & = \left(1 - \delta(y_{ij}^r, \bar{y}_{ij}^r)\right)(1 + y_{ij}^r \theta_{ir}) \end{cases}$$

as well as a scaled C. Note that $\mathbf{w}$ is of dimensionality K, the number of latent factors, while the number of constraints $|\{j \mid ij \in \mathcal{I}\}| \times (L-1)$, upper-bounded by $M \times (L-1)$, might vary diversely across different rows in Y according to how many items the corresponding user has rated.

Also note that we've extended the loss Δ in a sense by taking into account not only the prediction $\bar{y}$ and the training output y, but also another sample-specific constant θ_{ir}. This change might bring about a negative loss, but it will not impact the optimization process.

B.2 Optimizing V given U & θ

Similar to B.1, we may decompose $\min_{V|U,\theta} J(U, V, \theta)$ into M independent subproblems, one for V_j each, and then solve them in parallel by *SVM*^{*struct*}.

B.3 Optimizing θ given U & V

Again the special structure of the problem allows us to decompose it into $N \times (L-1)$ independent subproblems, one for θ_{ir} each, as

$$\min_{\theta|U,V} J(U, V, \theta) = C \sum_{i=1}^{N} \sum_{r=1}^{L-1} \min_{\theta_{ir}} J(\theta_{ir}|U, V) + \text{const.}, \tag{41}$$

where

$$J(\theta_{ir}|U, V) = \sum_{j|ij \in \mathcal{I}} h\left(T_{ij}^r (\theta_{ir} - U_i V_j^\top)\right) \tag{42}$$

and const. denotes the remaining term, which is independent of θ_{ir}s $((\|U\|_F^2 + \|V\|_F^2)/2)$.

Rewrite $\min_{\theta_{ir}} J(\theta_{ir}|U, V)$ by introducing slack variables ξ_{ij}^r and after some rearrangement, we have

$$\begin{aligned} \min_{\theta_{ir}, \xi_{\mathbf{i}}^{\mathbf{r}} \geq \mathbf{0}} \quad & \sum_{j|ij \in \mathcal{I}} \xi_{ij}^r \\ \text{s.t.} \quad & \xi_{ij}^r \geq 1 + U_i V_j^\top - \theta_{ir} \quad (\text{if } T_{ij}^r = 1) \\ & \xi_{ij}^r \geq 1 - U_i V_j^\top + \theta_{ir} \quad (\text{if } T_{ij}^r = -1). \end{aligned} \tag{43}$$

While (43) is solvable by any general linear programming solver, we find it to be an innate binary search problem over θ_{ir} and thus can be solved far more efficiently.

We group the constraints according to T_{ij}^r and denote them by $\mathcal{C}^1$ and $\mathcal{C}^{-1}$ respectively, and then define

$$\begin{cases} \mathcal{Z} & \doteq \{z \mid \exists j, \text{ s.t. } z = z_j \doteq U_i V_j^\top + T_{ij}^r\} \\ \mathcal{J}^1(\theta) & \doteq \{j \in \mathcal{C}^1 \mid z_j > \theta\} \\ \mathcal{J}^{-1}(\theta) & \doteq \{j \in \mathcal{C}^{-1} \mid z_j < \theta\} \\ \Delta_l(\theta) & \doteq |\mathcal{J}^1(\theta - \epsilon)| - |\mathcal{J}^{-1}(\theta - \epsilon)| \\ \Delta_r(\theta) & \doteq |\mathcal{J}^{-1}(\theta + \epsilon)| - |\mathcal{J}^1(\theta + \epsilon)| \end{cases}$$

Note that we've omitted the index i, r whenever possible to imply the same process applies to any subproblem of θ_{ir}. Think of each constraint in problem (43) as a clipped half-plane in the $(\theta_{ir}, \xi_{ij}^r \geq 0)$ space, then $\mathcal{Z}$ denotes all the unique zero-crossings (or θ_{ir} intercepts) and it's easy to see that $\epsilon\Delta_l(\theta)$ and $\epsilon\Delta_r(\theta)$ equals the change in the objective when taking a small enough step ϵ to the left and right from θ respectively.

We sort $\mathcal{Z}$ in ascending order so that $z'_1 < z'_2 < \cdots < z'_{|\mathcal{Z}|}$, on which we conduct a binary search to find the optimal θ_{ir} as follows.

(a) Start with $s = 1$, $t = |\mathcal{Z}|$, and $j = \lceil \frac{s+t}{2} \rceil$.

(b) If $\Delta_l(z'_j) < 0$, take $t = j - 1$ and goto (c);
If $\Delta_r(z'_j) < 0$, take $s = j + 1$ and goto (c);
Otherwise, goto (d).

(c) Take $j = \lceil \frac{s+t}{2} \rceil$ and goto (b).

(d) If $\Delta_l(z'_j) = 0$, $\theta_{ir} \in [z'_{j-1}, z'_j]$;
If $\Delta_r(z'_j) = 0$, $\theta_{ir} \in [z'_j, z'_{j+1}]$;
Otherwise, $\theta_{ir} = z'_j$.

Hence the overall time complexity of optimizing each θ_{ir} is $O(|\mathcal{Z}| \log |\mathcal{Z}|)$, which is further bounded by $O(M \log M)$.

B.4 Influence of the margin parameter on M^{3}F

Both the original M^{3}F model and its fast version adopted hinge loss, which is appropriate for discrete ordinal ratings. A simple generalization to the canonical hinge loss would be

$$h_\ell(x) = \max(\ell - x, 0) \quad (\ell > 0), \tag{44}$$

where ℓ is the margin parameter and it's fairly natural to wonder what influences, if any, the margin parameter might exert on the solution and the performance of the model.

Denote the new objective by $J_\ell(U, V, \theta)$ and we have the following observation:

$$\begin{aligned} J_\ell(\sqrt{\ell}U, \sqrt{\ell}V, \ell\theta) &= \frac{1}{2}\left(\|\sqrt{\ell}U\|_F^2 + \|\sqrt{\ell}V\|_F^2\right) + C \sum_{ij\in\mathcal{I}} \sum_{r=1}^{L-1} h_\ell\left(T_{ij}^r(\ell\theta_{ir} - \ell U_i V_j^\top)\right) \\ &= \frac{\ell}{2}\left(\|U\|_F^2 + \|V\|_F^2\right) + \ell C \sum_{ij\in\mathcal{I}} \sum_{r=1}^{L-1} h\left(T_{ij}^r(\theta_{ir} - U_i V_j^\top)\right) \\ &= \ell J(U, V, \theta). \end{aligned} \tag{45}$$

Therefore the original minimizer $(\tilde{U}, \tilde{V}, \tilde{\theta})$ of J, when scaled to $(\sqrt{\ell}\tilde{U}, \sqrt{\ell}\tilde{V}, \ell\tilde{\theta})$, becomes the minimizer of J_ℓ and what's more, these 2 minimizers yield exactly the same prediction rating matrix. That's to say M^{3}F is in a way invariant to the margin parameter, which is a desirable property.

Appendix C. A specific binary search solver

Here we propose a general binary search algorithm to solve problems of the following form:

$$\min_{x\in\mathbb{R}} \quad g(x) + \sum_{i=1}^{N} h_{\ell_i}(a_i x), \tag{46}$$

where $g(x)$ is a *strictly* convex function whose first-order derivative is continuous and easy to get, $h_\ell(x) = \max(0, \ell - x)$ the generalized hinge loss and $a_i \in \mathbb{R}$ ($a_i \neq 0$) the coefficients.

A first observation is that problem (46) is a strictly convex optimization problem of x and has thus a unique optimal solution. Another reason why we're interested in this specific kind of problem is that it actually serves as the conditional subproblem to nearly all the optimization problems (including SVM) that we'll encounter when performing variational inference in PM^{3}F models (as shown later in Appendix D), and therefore naturally fits into the coordinate descent solver for these problems.

We cannot simply take the gradient of the objective function $f(x)$ and set it to zero to get the optimal $\hat{x}$ due to the special form of hinge loss. Alternatively one starts with an initial x_0, and update the value with $x_{(n+1)} = g'^{-1}(-\sum_i h_i'^{(n)})$ where $h_i'^{(n)}$ is the subgradient of the hinge loss $h_{\ell_i}(x)$ at $x_{(n)}$[19], and so on and so forth. However this iteration process does not guarantee a convergence.

Here we introduce an intuitive binary search algorithm that exactly solves problem (46). Actually it is just a slightly varied version of what we did in B.3. We sort all the *zero-crossings* of the hinge loss terms in ascending order so that $z_{i_1} \leq z_{i_2} \leq \cdots \leq z_{i_N}$ where $z_i = \ell_i / a_i$. Then it's obvious that the sequence

$$-\infty, f'^{-}_{i_1}, f'^{+}_{i_1}, \ldots, f'^{-}_{i_N}, f'^{+}_{i_N}, +\infty,$$

where $f'^{-}_{i_k}$ and $f'^{+}_{i_k}$ are the left and right derivatives of f at z_{i_k} respectively:

$$f'^{-}_{i_k} = g'(z_{i_k}) + h'^{-}_{i_k}, \quad h'^{-}_{i_k} = \sum_{j=1}^{i_k - 1} - a_j \mathbb{I}(a_j < 0) + \sum_{j=i_k}^{N} - a_j \mathbb{I}(a_j > 0) \tag{47}$$

$$f'^{+}_{i_k} = g'(z_{i_k}) + h'^{+}_{i_k}, \quad h'^{+}_{i_k} = \sum_{j=1}^{i_k} - a_j \mathbb{I}(a_j < 0) + \sum_{j=i_k+1}^{N} - a_j \mathbb{I}(a_j > 0), \tag{48}$$

is monotonically ascending and there must exist a unique k so that either $f'^{-}_{i_k} < 0 < f'^{+}_{i_k}$, which indicates $\hat{x} = z_{i_k}$, or $f'^{+}_{i_k} < 0 < f'^{-}_{i_{k+1}}$, which indicates $\hat{x} = g'^{-1}(h'_{i_k}) \in (z_{i_k}, z_{i_{k+1}})$ where $h'_{i_k} = h'^{+}_{i_k} = h'^{-}_{i_{k+1}}$ is the constant gradient of the hinge loss terms over section $(z_{i_k}, z_{i_{k+1}})$.

The overall complexity of the algorithm is $O(N \log N)$ due to the sorting step.

Appendix D. Variational inference details

D.1 Inference in the iPM^{3}F model

D.1.1 Solving for $p(V)$

Subproblem:

$$\min_{p(V)} \quad \mathrm{KL}(p(V) \| p_0(V)) + C \sum_{ij \in \mathcal{I}} h_\ell \left(Y_{ij} \bar{Z}_i \mathbb{E}_p[V_j]^\top \right), \tag{49}$$

where $\bar{Z}_i = \mathbb{E}_{p(Z)}[Z_i] = \psi_i$.

Solution:

$$p(V) \propto p_0(V) \exp \left\{ \sum_{ij \in \mathcal{I}} \omega_{ij} Y_{ij} \psi_i V_j^\top \right\}. \tag{50}$$

Note that $p(V)$ remains an isotropic Gaussian

$$p(V) = \prod_{j=1}^{M} \prod_{k=1}^{K} \mathcal{N}(V_{jk} | \Lambda_{jk}, \sigma^2), \tag{51}$$

where $\Lambda_{jk} = \sum_{i | ij \in \mathcal{I}} \omega_{ij} Y_{ij} \psi_{ik}$.

Dual:

$$\begin{aligned} \max_{\boldsymbol{\omega}} \quad & \ell \sum_{ij \in \mathcal{I}} \omega_{ij} - \frac{\sigma^2}{2} \sum_{j=1}^{M} \| \sum_{i,\, ij \in \mathcal{I}} \omega_{ij} Y_{ij} \psi_i \|^2 \\ \text{s.t.} \quad & \forall i, j \in \mathcal{I} \, : \, 0 \leq \omega_{ij} \leq C. \end{aligned} \tag{52}$$

It's obvious that the dual naturally decomposes into M independent box-constrained quadratic programming subproblems, one for $\boldsymbol{\omega}_j$ each.

Equivalent primal: From either the solution or the dual above, one can easily prove that the primal problem can actually be decomposed into M independent binary SVM problems, one for Λ_j each, as follows,

$$\min_{\Lambda_j} \quad \frac{1}{2\sigma^2} \| \Lambda_j \|^2 + C \sum_{i | ij \in \mathcal{I}} h_\ell \left(Y_{ij} \Lambda_j \psi_i^\top \right). \tag{53}$$

And thus we can use some existing high-performance SVM solver (e.g. *SVM^{struct}*) to efficiently, and in a parallel fashion, solve for $p(V)$.

D.1.2 Solving for $p(\nu)$

Since ν is marginalized before exerting any influence in the loss, this part is independent of the loss.

Subproblem:

$$\min_{p(\nu)} \quad \mathrm{KL}(p(\nu)p(Z)\|p_0(\nu, Z)) \tag{54}$$

Solution: We adopt the same multivariate lower bound [2] to seek an approximate solution, which is exactly the same as in [2].

$$\begin{aligned} \gamma_{k1} &= \alpha + \sum_{\kappa=k}^{K}\sum_{i=1}^{N}\psi_{i\kappa} + \sum_{\kappa=k+1}^{K}\left(N - \sum_{i=1}^{N}\psi_{i\kappa}\right)\left(\sum_{i=k+1}^{\kappa} q_{\kappa i}\right) \\ \gamma_{k2} &= 1 + \sum_{\kappa=k}^{K}\left(N - \sum_{i=1}^{N}\psi_{i\kappa}\right) q_{\kappa k}, \end{aligned} \tag{55}$$

where the variational parameter $q_{\kappa\cdot} = (q_{\kappa 1}, \dots, q_{\kappa\kappa})^\top$ lies on a κ-simplex and

$$q_{\kappa i} \propto \tilde{q}_i = \exp\left\{\psi(\gamma_{i2}) + \sum_{j=1}^{i-1}\psi(\gamma_{j1}) - \sum_{j=1}^{i}\psi(\gamma_{j1} + \gamma_{j2})\right\}. \tag{56}$$

Recurrent calculation: We may rearrange (55) to allow for a quick recurrent calculation as follows,

$$\begin{aligned} \gamma_{k1} &= \alpha + \sum_{\kappa=k}^{K} P(\kappa) + \sum_{i=k+1}^{K} Q(i) \\ \gamma_{k2} &= 1 + Q(k), \end{aligned} \tag{57}$$

where

$$\begin{aligned} P(\kappa) &= \sum_{i=1}^{N}\psi_{i\kappa} \\ Q(i) &= \sum_{\kappa=i}^{K}\left(N - P(\kappa)\right) q_{\kappa i}. \end{aligned}$$

Furthermore, since $q_{\kappa i} = C_\kappa \tilde{q}_i$, C_κ being the normalization constant, we can even calculate $Q(i)$ (or $Q(i)/\tilde{q}_i$ actually) recurrently.

D.1.3 Solving for $p(Z)$

Subproblem:

$$\min_{p(Z)} \quad \mathrm{KL}(p(\nu)p(Z)\|p_0(\nu, Z)) + C\sum_{ij\in\mathcal{I}} h_\ell\left(Y_{ij}\mathbb{E}_p[Z_i]\bar{V}_j^\top\right), \tag{58}$$

where $\bar{V}_j = \mathbb{E}_{p(V)}[V_j] = \Lambda_j$ and $\mathbb{E}_p[Z_i] = \psi_i$.

This is a convex optimization problem of ψ and it decomposes into N independent subproblems, one for ψ_i each, as (after dropping irrelevant terms)

$$\min_{\psi_i} \quad \sum_{k=1}^{K}\Big(\mathbb{E}_Z[\log p(Z_{ik})] - \mathbb{E}_{\nu,Z}[\log p_0(Z_{ik}|\nu)]\Big) + C\sum_{j|ij\in\mathcal{I}} h_\ell\left(Y_{ij}\psi_i\Lambda_j^\top\right), \tag{59}$$

where

$$\begin{aligned} \mathbb{E}_Z[\log p(Z_{ik})] &= \psi_{ik}\log\psi_{ik} + (1-\psi_{ik})\log(1-\psi_{ik}) \\ \mathbb{E}_{\nu,Z}[\log p_0(Z_{ik}|\nu)] &= \psi_{ik}\sum_{j=1}^{k}\mathbb{E}_\nu[\log\nu_j] + (1-\psi_{ik})\mathbb{E}_\nu[\log(1-\prod_{j=1}^{k}\nu_j)] \\ \mathbb{E}_\nu[\log\nu_j] &= \psi(\gamma_{k1}) - \psi(\gamma_{k1}+\gamma_{k2}) \\ \mathbb{E}_\nu[\log(1-\prod_{j=1}^{k}\nu_j)] &\geq \mathcal{L}_k^\nu, \end{aligned}$$

where

$$\mathcal{L}_k^{\nu} = H(q_{k.}) + \sum_{i=1}^{k} q_{ki}\psi(\gamma_{i2}) + \sum_{i=1}^{k-1}\left(\sum_{j=i+1}^{k} q_{kj}\right)\psi(\gamma_{i1}) - \sum_{i=1}^{k}\left(\sum_{j=i}^{k} q_{kj}\right)\psi(\gamma_{i1}+\gamma_{i2}) \quad (60)$$

is in turn the multivariate lower bound as in [2].

Solution: By use of the binary search algorithm introduced in Appendix C, we solve problem (59) in a coordinate descent manner, with iteration number ever increasing during the learning process.

Recurrent calculation: Again we may rearrange Eq. (60) to allow for a quick recurrent calculation of $\mathcal{L}_k^{\nu}$ (or $(\mathcal{L}_k^{\nu} + \log C_k)/C_k$ actually) as follows,

$$\frac{\mathcal{L}_k^{\nu} + \log C_k}{C_k} = \sum_{i=1}^{k} \tilde{q}_i \psi(\gamma_{i2}) + \sum_{j=2}^{k} \tilde{q}_j \sum_{i=1}^{j-1} \psi(\gamma_{i1}) - \sum_{j=1}^{k} \tilde{q}_j \sum_{i=1}^{j} \psi(\gamma_{i1}+\gamma_{i2}) - \sum_{i=1}^{k} \tilde{q}_i \log \tilde{q}_i. \quad (61)$$

Functional form: From subproblem (58) and the general MED solution under mean-field assumptions (33), we may easily obtain the functional form of $p(Z)$ as

$$\begin{aligned} p(Z) &\propto \exp\left\{ \mathbb{E}_{\nu}[\log p_0(Z|\nu)] + \sum_{ij\in\mathcal{I}} \omega_{ij} Y_{ij} Z_i \Lambda_j^{\top} \right\} \\ &\propto \prod_{i=1}^{N}\prod_{k=1}^{K} \exp\left\{\zeta_{ik} Z_{ik}\right\}, \end{aligned} \quad (62)$$

where

$$\zeta_{ik} = \sum_{j=1}^{k} \mathbb{E}_{\nu}[\log \nu_j] - \mathbb{E}_{\nu}[\log(1 - \prod_{j=1}^{k} \nu_j)] + \sum_{j|ij\in\mathcal{I}} \omega_{ij} Y_{ij} \Lambda_{jk}$$

is a constant, and hence $p(Z)$ is fully factorized. Furthermore, the fact that $Z \in \{0,1\}^{N\times K}$ is a binary matrix naturally suggests a Bernoulli parametrization for its entries, which justifies our pretreatment of $p(Z)$ in Sec. 4.

D.1.4 Solving for $p(\theta)$

Subproblem:

$$\min_{p(\theta)} \quad \mathrm{KL}(p(\theta)\|p_0(\theta)) + C\sum_{ij\in\mathcal{I}}\sum_{r=1}^{L-1} h_{\ell}\left(T_{ij}^{r}(\mathbb{E}_p[\theta_{ir}] - \psi_i\Lambda_j^{\top})\right) \quad (63)$$

Solution:

$$p(\theta) \propto p_0(\theta)\exp\left\{\sum_{ij\in\mathcal{I}}\sum_{r=1}^{L-1} \omega_{ij}^{r} T_{ij}^{r} \theta_{ir}\right\}. \quad (64)$$

Note that $p(\theta)$ remains an isotropic Gaussian

$$p(\theta) = \prod_{i=1}^{N}\prod_{r=1}^{L-1} \mathcal{N}(\theta_{ir}|\varrho_{ir}, \varsigma^2), \quad (65)$$

where $\varrho_{ir} = \rho_r - \varsigma^2 \sum_{j|ij\in\mathcal{I}} \omega_{ij}^{r} T_{ij}^{r}$.

Equivalent Primal: As a result, we may decompose the original subproblem (63) into $N \times (L-1)$ independent sub-subproblems, one for θ_{ir} each, as follows,

$$\min_{\varrho_{ir}} \quad \frac{1}{2\varsigma^2}(\varrho_{ir} - \rho_r)^2 + C\sum_{j|ij\in\mathcal{I}} h_{\ell}\left(T_{ij}^{r}(\varrho_{ir} - \psi_i\Lambda_j^{\top})\right). \quad (66)$$

Note that as $\varsigma \to +\infty$, the Gaussian distribution regresses to a uniform distribution and problem (66) reduces accordingly to the non-probabilistically-formulated subproblem as is the case of M^{3}F (B.3). It's clear that the binary search algorithm introduced in Appendix C directly applies here.

D.2 Inference in the iBPM^{3}F model

D.2.1 Solving for $p(V)$

Subproblem:

$$\min_{p(V)} \quad \mathrm{KL}(p(V)p(\mu,\Omega)\|p_0(V,\mu,\Omega)) + \sum_{ij\in\mathcal{I}} h_\ell\left(Y_{ij}\psi_i\mathbb{E}_p[V_j]^\top\right) \tag{67}$$

Solution:

$$\begin{aligned} p(V) &\propto \exp\left\{\sum_{j=1}^{M}\mathbb{E}_{\mu,\Omega}[\log p_0(V_j|\mu,\Omega)] + \sum_{ij\in\mathcal{I}}\omega_{ij}Y_{ij}\psi_i V_j^\top\right\} \\ &\propto \prod_{j=1}^{M}\exp\left\{-\frac{1}{2}V_j\mathbb{E}[\Omega]V_j^\top + \mathbb{E}[\mu\Omega]V_j^\top + \sum_{i|ij\in\mathcal{I}}\omega_{ij}Y_{ij}\psi_i V_j^\top\right\} \end{aligned} \tag{68}$$

factorizes into M Gaussian distributions, one on V_j each, parameterized as $V_j \sim \mathcal{N}(\Lambda_j, \Xi^{-1})$, where

$$\Lambda_j = \tilde{\mu} + \left(\sum_{i|ij\in\mathcal{I}}\omega_{ij}Y_{ij}\psi_i\right)\Xi^{-1},\ \tilde{\mu} = \mathbb{E}[\mu],\ \Xi = \mathbb{E}[\Omega]. \tag{69}$$

Note that we've made use of the equation $\mathbb{E}[\mu\Omega] = \mathbb{E}[\mu]\mathbb{E}[\Omega]$ due to the fact that $\mathbb{E}[\mu|\Omega] = \tilde{\mu}$ does not depend on Ω, which will soon be shown in D.2.2 as $p(\mu,\Omega)$ remains a Gaussian-Wishart distribution.

Dual: Rewrite the decomposed primal of each $p(V_j)$ as

$$\min_{p(V_j)} \quad \mathrm{KL}\left(p(V_j)\|\tilde{p}(V_j)\right) + \sum_{i|ij\in\mathcal{I}} h_\ell\left(Y_{ij}\psi_i\mathbb{E}_p[V_j]^\top\right), \tag{70}$$

where $\tilde{p}(V_j) \propto \exp\left\{\mathbb{E}[\log p_0(V_j|\mu,\Omega)]\right\} \propto \mathcal{N}(\tilde{\mu}, \Xi^{-1})$. Then it's obvious from the general dual form (30) that the dual takes the form of

$$\begin{aligned} \max_{\boldsymbol{\omega}_j} \quad & \ell\sum_{i|ij\in\mathcal{I}}\omega_{ij} - \frac{1}{2}\mathcal{Z}_j(\boldsymbol{\omega}_j)\Xi^{-1}\mathcal{Z}_j(\boldsymbol{\omega}_j)^\top - \tilde{\mu}\mathcal{Z}_j(\boldsymbol{\omega}_j)^\top \\ \text{s.t.} \quad & \forall i|ij\in\mathcal{I} : 0 \le \omega_{ij} \le 1, \end{aligned} \tag{71}$$

where $\mathcal{Z}_j(\boldsymbol{\omega}_j) = \sum_{i|ij\in\mathcal{I}}\omega_{ij}Y_{ij}\psi_i$ and the dual is again a box-constrained quadratic programming.

Equivalent primal: We rewrite the primal (70) by replacing $p(V_j)$ and $\tilde{p}(V_j)$ with their respective parameterized Gaussian density, thus yielding

$$\min_{\Lambda_j} \quad \frac{1}{2}(\Lambda_j - \tilde{\mu})\Xi(\Lambda_j - \tilde{\mu})^\top + \sum_{i|ij\in\mathcal{I}} h_\ell\left(Y_{ij}\psi_i\Lambda_j^\top\right). \tag{72}$$

Now suppose $\Xi = \mathrm{P}\mathrm{P}^\top$ $(\mathrm{P} \succ 0)$ and let $\Lambda_j' = (\Lambda_j - \tilde{\mu})\mathrm{P}$, and we have

$$\min_{\Lambda_j'} \quad \frac{1}{2}\|\Lambda_j'\|_2^2 + \sum_{i|ij\in\mathcal{I}} h_{\ell_{ij}}\left(Y_{ij}\Lambda_j'\mathrm{P}^{-1}\psi_i^\top\right), \tag{73}$$

where $\ell_{ij} = \ell - Y_{ij}\tilde{\mu}\psi_i^\top$ becomes the sample-specific margin. Now we may solve for Λ_j' via a slightly changed *SVM^{struct}* and get $\Lambda_j = \Lambda_j' P^{-1} + \tilde{\mu}$.

D.2.2 Solving for $p(\mu,\Omega)$

Like solving for ν in iPM^{3}F, this part is independent of the loss so we work on the KL-divergence directly.

Subproblem:

$$\min_{p(\mu,\Omega)} \quad \mathrm{KL}(p(\mu,\Omega)p(V)\|p_0(\mu,\Omega,V)) \tag{74}$$

Solution:

$$p(\mu,\Omega) \propto p_0(\mu,\Omega)\exp\left\{\sum_{j=1}^{M}\mathbb{E}_{V_j}[\log p_0(V_j|\mu,\Omega)]\right\}, \tag{75}$$

where

$$\begin{aligned}
\mathbb{E}_{V_j}[\log p_0(V_j|\mu,\Omega)] &= -\frac{1}{2}\left(\mathbb{E}_{V_j}[(V_j-\mu)\Omega(V_j-\mu)^\top] - \log|\Omega|\right) + \text{const.}\\
&= -\frac{1}{2}\left(\mathbb{E}_{V_j}[\mathrm{tr}\left((V_j-\mu)^\top(V_j-\mu)\Omega\right)] - \log|\Omega|\right) + \text{const.}\\
&= -\frac{1}{2}\,\mathrm{tr}\left(\mathbb{E}_{V_j}[(V_j-\Lambda_j+\Lambda_j-\mu)^\top(V_j-\Lambda_j+\Lambda_j-\mu)]\Omega\right)\\
&\quad + \frac{1}{2}\log|\Omega| + \text{const.} \qquad (\text{recall that } V_j \sim \mathcal{N}(\Lambda_j,\Xi^{-1}))\\
&= -\frac{1}{2}\left[\mathrm{tr}\left(\left[\Xi^{-1}+(\Lambda_j-\mu)^\top(\Lambda_j-\mu)\right]\Omega\right) - \log|\Omega|\right] + \text{const.}\\
&= -\frac{1}{2}\left[(\Lambda_j-\mu)\Omega(\Lambda_j-\mu)^\top + \mathrm{tr}\left(\Xi^{-1}\Omega\right) - \log|\Omega|\right] + \text{const.}
\end{aligned} \tag{76}$$

Recall the functional form of a Gaussian-Wishart density is

$$\begin{aligned}
\mathcal{GW}(\mu,\Omega|\tilde{\mu},\tilde{\beta},\tilde{W},\tilde{\tau}) &= \mathcal{N}(\mu|\tilde{\mu},(\tilde{\beta}\Omega)^{-1})\mathcal{W}(\Omega|\tilde{W},\tilde{\tau})\\
&\propto \exp\left\{-\frac{1}{2}(\mu-\tilde{\mu})\tilde{\beta}\Omega(\mu-\tilde{\mu})^\top\right\}|\Omega|^{\frac{\tilde{\tau}-K-1}{2}}\exp\left\{-\frac{1}{2}\,\mathrm{tr}\left(\tilde{W}^{-1}\Omega\right)\right\}\\
&= |\Omega|^{\frac{\tilde{\tau}-K-1}{2}}\exp\left\{-\frac{1}{2}\tilde{\beta}\mu\Omega\mu^\top + \tilde{\beta}\tilde{\mu}\Omega\mu^\top - \frac{1}{2}\,\mathrm{tr}\left(\left(\tilde{\beta}\tilde{\mu}^\top\tilde{\mu}+\tilde{W}^{-1}\right)\Omega\right)\right\}.
\end{aligned}$$

Then after substituting Eq. (76) into (75), we conclude, by observation, that

$$p(\mu,\Omega) = \mathcal{GW}(\tilde{\mu},\tilde{\beta},\tilde{W},\tilde{\tau})$$

where

$$\tilde{\tau} = \tau_0 + M \tag{77}$$

$$\tilde{\beta} = \beta_0 + M \tag{78}$$

$$\tilde{\mu} = \frac{1}{\tilde{\beta}}\left(\beta_0\mu_0 + \sum_{j=1}^{M}\Lambda_j\right) \tag{79}$$

$$\tilde{W}^{-1} = -\tilde{\beta}\tilde{\mu}^\top\tilde{\mu} + \beta_0\mu_0^\top\mu_0 + W_0^{-1} + M\Xi^{-1} + \sum_{j=1}^{M}\Lambda_j^\top\Lambda_j. \tag{80}$$

Incidentally, $\Xi = \mathbb{E}[\Omega] = \tilde{\tau}\tilde{W}$.

D.2.3 KL-divergence

By use of the decomposition (34), we have

$$\begin{aligned}
\mathrm{KL}(p(V)p(\mu,\Omega)\|p_0(V,\mu,\Omega)) &= \mathrm{KL}(p(V)\|\tilde{p}(V)) + \mathrm{KL}(p(\mu,\Omega)\|p_0(\mu,\Omega))\\
&\quad - \log\int\exp\left\{\mathbb{E}_{\mu,\Omega}[\log p_0(V|\mu,\Omega)]\right\}dV,
\end{aligned} \tag{81}$$

where $\mathrm{KL}(p(V)\|\tilde{p}(V))$ is the KL-divergence between Gaussian distributions:

$$\begin{aligned}\mathrm{KL}(p(V)\|\tilde{p}(V)) &= \sum_{j=1}^{M}\mathrm{KL}(\mathcal{N}(\Lambda_j,\Xi^{-1})\|\mathcal{N}(\tilde{\mu},(\tilde{\tau}\tilde{W})^{-1}))\\ &= \frac{M}{2}\left(\tilde{\tau}\,\mathrm{tr}(\tilde{W}\Xi^{-1}) - \log\frac{\tilde{\tau}^K|\tilde{W}|}{|\Xi|} - K\right) + \frac{\tilde{\tau}}{2}\sum_{j=1}^{M}(\Lambda_j-\tilde{\mu})\tilde{W}(\Lambda_j-\tilde{\mu})^\top;\end{aligned} \tag{82}$$

$\mathrm{KL}(p(\mu,\Omega)\|p_0(\mu,\Omega))$ is the KL-divergence between Gaussian-Wishart distributions:

$$\begin{aligned}\mathrm{KL}(p(\mu,\Omega)\|p_0(\mu,\Omega)) &= \mathrm{KL}(\mathcal{GW}(\tilde{\mu},\tilde{\beta},\tilde{W},\tilde{\tau})\|\mathcal{GW}(\mu_0,\beta_0,W_0,\tau_0))\\ &= \mathrm{KL}(p(\Omega)\|p_0(\Omega)) + \mathbb{E}_{p(\Omega)}[\mathrm{KL}(p(\mu|\Omega)\|p_0(\mu|\Omega))]\\ &= \mathrm{KL}(\mathcal{W}(\tilde{W},\tilde{\tau})\|\mathcal{W}(W_0,\tau_0))\\ &\quad + \mathbb{E}_{p(\Omega)}[\mathrm{KL}\left(\mathcal{N}(\tilde{\mu},(\tilde{\beta}\Omega)^{-1})\|\mathcal{N}(\mu_0,(\beta_0\Omega)^{-1})\right)]\end{aligned}$$

$$\Big\Downarrow \quad \text{breaking into parts}$$

$$\begin{aligned}\mathrm{KL}(\mathcal{W}(\tilde{W},\tilde{\tau})\|\mathcal{W}(W_0,\tau_0)) &= \frac{\tilde{\tau}-\tau_0}{2}\mathbb{E}_{\mathcal{W}(\Omega|\tilde{W},\tilde{\tau})}[\log|\Omega|] + \log\frac{B(\tilde{W},\tilde{\tau})}{B(W_0,\tau_0)}\\ &\quad - \frac{\tilde{\tau}K}{2} + \frac{\tilde{\tau}}{2}\,\mathrm{tr}\left(W_0^{-1}\tilde{W}\right)\\ \text{partition function:}\quad B(W,\tau) &= \left(|W|^{\tau/2}2^{\tau K/2}\Gamma_K(\tau/2)\right)^{-1}\\ \text{multivariate Gamma function:}\quad \Gamma_K(\tau/2) &= \pi^{K(K-1)/4}\prod_{k=1}^{K}\Gamma\left(\frac{\tau+1-k}{2}\right)\\ \mathbb{E}_{\mathcal{W}(\Omega|\tilde{W},\tilde{\tau})}[\log|\Omega|] &= \sum_{k=1}^{K}\psi\left(\frac{\tilde{\tau}+1-k}{2}\right) + K\log 2 + \log|\tilde{W}|\\ \mathrm{KL}\left(\mathcal{N}(\tilde{\mu},(\tilde{\beta}\Omega)^{-1})\|\mathcal{N}(\mu_0,(\beta_0\Omega)^{-1})\right) &= \frac{1}{2}\Big[\mathrm{tr}\left(\beta_0\Omega(\tilde{\beta}\Omega)^{-1}\right) + (\tilde{\mu}-\mu_0)\beta_0\Omega(\tilde{\mu}-\mu_0)^\top\\ &\quad - \log\left(|\beta_0\Omega|/|\tilde{\beta}\Omega|\right) - K\Big]\\ &= \frac{1}{2}\Big[K\beta_0/\tilde{\beta} + (\tilde{\mu}-\mu_0)\beta_0\Omega(\tilde{\mu}-\mu_0)^\top\\ &\quad -K\log\left(\beta_0/\tilde{\beta}\right) - K\Big]\end{aligned}$$

$$\Big\Downarrow \quad \text{putting back together}$$

$$\begin{aligned}\mathrm{KL}(p(\mu,\Omega)\|p_0(\mu,\Omega)) &= \frac{\tilde{\tau}}{2}\,\mathrm{tr}\left(W_0^{-1}\tilde{W}\right) - \frac{\tau_0}{2}\log\left(|\tilde{W}|/|W_0|\right)\\ &\quad + \frac{1}{2}(\tilde{\mu}-\mu_0)\beta_0\tilde{\tau}\tilde{W}(\tilde{\mu}-\mu_0)^\top\\ &\quad + \mathrm{T}\left(\tau_0,\tilde{\tau}\right) + \mathrm{B}\left(\beta_0,\tilde{\beta}\right)\\ \mathrm{T}\left(\tau_0,\tilde{\tau}\right) &= \sum_{k=1}^{K}\left[\log\Gamma\left(\frac{\tau_0+1-k}{2}\right) - \log\Gamma\left(\frac{\tilde{\tau}+1-k}{2}\right)\right]\\ &\quad + \frac{\tilde{\tau}-\tau_0}{2}\sum_{k=1}^{K}\psi\left(\frac{\tilde{\tau}+1-k}{2}\right) - \frac{\tilde{\tau}K}{2}\\ \mathrm{B}\left(\beta_0,\tilde{\beta}\right) &= \frac{K}{2}\left(\beta_0/\tilde{\beta} - \log\left(\beta_0/\tilde{\beta}\right) - 1\right),\end{aligned} \tag{83}$$

where T and B remain constant after the first update of $(\tilde{\mu}, \tilde{\beta}, \tilde{W}, \tilde{\tau})$ according to Eq. (77) and (78) and is thus dispensable given that we calculate the KL-divergence, a term in the objective, only as a guidance to the convergence of the variational inference;

And we calculate the last log-integral term in Eq. (81) as follows:

$$\begin{aligned}
\log \int \exp\left\{\mathbb{E}_{\mu,\Omega}[\log p_0(V|\mu,\Omega)]\right\} dV &= \log \int \exp\left\{\sum_{j=1}^{M} \mathbb{E}_{\mu,\Omega}[\log p_0(V_j|\mu,\Omega)]\right\} dV \\
&= \sum_{j=1}^{M} \log \int \exp\{\mathbb{E}_{\mu,\Omega}[\log p_0(V_j|\mu,\Omega)]\} dV_j
\end{aligned}$$

$$\Downarrow \quad \text{breaking into parts}$$

$$\begin{aligned}
\mathbb{E}_{\mu,\Omega}[\log p_0(V_j|\mu,\Omega)] &= -\frac{1}{2}\left\{\mathbb{E}_{\mu,\Omega}[(V_j-\mu)\Omega(V_j-\mu)^\top] - \mathbb{E}_{\Omega}[\log|\Omega|] + K\log(2\pi)\right\} \\
\mathbb{E}_{\mu,\Omega}[(V_j-\mu)\Omega(V_j-\mu)^\top] &= V_j\mathbb{E}[\Omega]V_j^\top - 2\mathbb{E}[\mu]\mathbb{E}[\Omega]V_j^\top + \mathbb{E}_{\Omega}[\mathbb{E}_{\mu|\Omega}[\mathrm{tr}(\mu^\top\mu\Omega)]] \\
&= \tilde{\tau}V_j\tilde{W}V_j^\top - 2\tilde{\tau}\tilde{\mu}\tilde{W}V_j^\top + \mathbb{E}_{\Omega}[\tilde{\mu}\Omega\tilde{\mu}^\top + K\tilde{\beta}^{-1}] \\
&= \tilde{\tau}(V_j-\tilde{\mu})\tilde{W}(V_j-\tilde{\mu})^\top + K\tilde{\beta}^{-1} \\
\mathbb{E}_{\Omega}[\log|\Omega|] &= \sum_{k=1}^{K}\psi\left(\frac{\tilde{\tau}+1-k}{2}\right) + K\log 2 + \log|\tilde{W}|
\end{aligned}$$

$$\Downarrow \quad \text{putting back together}$$

$$\begin{aligned}
\int \exp\{\mathbb{E}_{\mu,\Omega}[\log p_0(V_j|\mu,\Omega)]\} dV_j &= \frac{(2\pi)^{\frac{K}{2}}}{|\tilde{\tau}\tilde{W}|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}\left(K\tilde{\beta}^{-1} - \mathbb{E}_{\Omega}[\log|\Omega|] + K\log(2\pi)\right)\right\} \\
\log\int \exp\{\mathrm{E}_{\mu,\Omega}[\log p_0(V|\mu,\Omega)]\} dV &= -\frac{M}{2}\left(K\tilde{\beta}^{-1} - \mathrm{E}_{\Omega}[\log|\Omega|] + \log(\tilde{\tau}^K|\tilde{W}|)\right) \\
&= -\frac{M}{2}\left(K\tilde{\beta}^{-1} - \sum_{k=1}^{K}\psi\left(\frac{\tilde{\tau}+1-k}{2}\right) + K\log\frac{\tilde{\tau}}{2}\right).
\end{aligned} \tag{84}$$

Note that the update rule of $\tilde{W}$ (80) when solving for $p(\mu, \Omega)$ can also be derived by taking partial derivative of the KL-divergence term (81) with respect to $\tilde{W}$ and setting it to zero (assuming Eq. (77)$\sim$(79) are already at hand).