Thurstonian Boltzmann Machines: Learning from Multiple Inequalities

Supplementary Material

Truyen Tran, Dinh Phung and Svetha Venkatesh

For readability, let us first recall the definition of the joint distribution

$$P(\boldsymbol{x}, \boldsymbol{h}) = \frac{1}{Z} \exp\left\{-E(\boldsymbol{x}, \boldsymbol{h})\right\} \tag{1}$$

$$E(\boldsymbol{x}, \boldsymbol{h}) = \sum_i \left(\frac{x_i^2}{2} - (\alpha_i + W_{i\bullet}\boldsymbol{h})\, x_i \right) - \boldsymbol{\gamma}' \boldsymbol{h} \tag{2}$$

where $Z = \sum_{\boldsymbol{h}} \int \exp\left\{-E(\boldsymbol{x}, \boldsymbol{h})\right\} d\boldsymbol{x}$ is the normalising constant and $E(\boldsymbol{x}, \boldsymbol{h})$ is the model energy, and $\{\alpha_i\}_{i=1}^N, W = \{w_{ik}\}, \boldsymbol{\gamma} = \{\gamma_k\}$ are free parameters.

1 Inference

1.1 Estimating the Partition Function

For convenience, let us re-parameterise the distribution as follows

$$\begin{aligned} \phi_i(x_i) &= \exp\left\{-\frac{x_i^2}{2} + \alpha_i x_i\right\} \\ \psi_{ik}(x_i, h_k) &= \exp\left\{W_{ik} x_i h_k\right\} \\ \phi_k(h_k) &= \exp\left\{\gamma_k h_k\right\} \end{aligned} \tag{3}$$

The model potential is then the product of all local potentials

$$\Psi(\boldsymbol{x}, \boldsymbol{h}) = \left[\prod_i \phi_i(x_i)\right] \left[\prod_{ik} \psi_{ik}(x_i, h_k)\right] \left[\prod_k \phi_k(h_k)\right] \tag{4}$$

The partition function can be rewritten as

$$\begin{aligned} Z &= \sum_{\boldsymbol{h}} \int_{\boldsymbol{x}} \Psi(\boldsymbol{x}, \boldsymbol{h}) d\boldsymbol{x} \\ &= \sum_{\boldsymbol{h}} \Omega(\boldsymbol{h}) \end{aligned}$$

where $\Omega(\boldsymbol{h}) = \int_{\boldsymbol{x}} \Psi(\boldsymbol{x}, \boldsymbol{h}) d\boldsymbol{x}$. We now proceed to compute $\Omega(\boldsymbol{h})$:

$$\begin{aligned}
\Omega(\boldsymbol{h}) &= \left[\prod_k \phi_k(h_k)\right] \int_{\boldsymbol{x}} \left[\prod_i \phi_i(x_i)\right] \left[\prod_{ik} \psi_{ik}(x_i, h_k)\right] d\boldsymbol{x} \\
&= \left[\prod_k \phi_k(h_k)\right] \prod_i \int_{x_i} \exp\left\{-\frac{x_i^2}{2} + (\alpha_i + \sum_k W_{ik} h_k) x_i\right\} dx_i \\
&= \left[\prod_k \phi_k(h_k)\right] \prod_i C_i \int_{x_i} \exp\left\{-\frac{(x_i - \mu_i(\boldsymbol{h}))^2}{2}\right\} dx_i \\
&= \left[\prod_k \phi_k(h_k)\right] \prod_i C_i \sqrt{2\pi\sigma_i^2}
\end{aligned}$$

where

$$\begin{aligned}
\mu_i(\boldsymbol{h}) &= \alpha_{id} + \sum_{k=1}^{K} W_{ik} h_k \\
C_i &= \exp\left\{\frac{1}{2}\left(\frac{\mu_i(\boldsymbol{h})}{\sigma_i}\right)^2\right\}
\end{aligned}$$

Now we can define the distribution over the hidden layer as follows

$$P(\boldsymbol{h}) = \frac{1}{Z}\Omega(\boldsymbol{h})$$

Now we apply the Annealed Importance Sampling (AIS) Neal [2001]. The idea is to introduces the notion of inverse-temperature τ into the model, i.e., $P(\boldsymbol{h}|\tau) \propto \Omega(\boldsymbol{h})^{\tau}$.

Let $\{\tau_s\}_{s=0}^{S}$ be the (slowly) increasing sequence of temperature, where $\tau_0 = 0$ and $\tau_S = 1$, that is $\tau_0 < \tau_1 ... < \tau_S$. At $\tau_0 = 0$, we have a uniform distribution, and at $\tau_S = 1$, we obtain the desired distribution. At each step s, we draw a sample $\boldsymbol{h}^s$ from the distribution $P(\boldsymbol{h}|\tau_{s-1})$ (e.g. using some Metropolis-Hastings procedure). Let $P^*(\boldsymbol{h}|\tau)$ be the unnormalised distribution of $P(\boldsymbol{h}|\tau)$, that is $P(\boldsymbol{h}|\tau) = P^*(\boldsymbol{h}|\tau)/Z(\tau)$. The final weight after the annealing process is computed as

$$\omega = \frac{P^*(\boldsymbol{h}^1|\tau_1)}{P^*(\boldsymbol{h}^1|\tau_0)} \frac{P^*(\boldsymbol{h}^2|\tau_2)}{P^*(\boldsymbol{h}^2|\tau_1)} ... \frac{P^*(\boldsymbol{h}^S|\tau_S)}{P^*(\boldsymbol{h}^S|\tau_{S-1})}$$

The above procedure is repeated T times. Finally, the normalisation constant at $\tau = 1$ is computed as $Z(1) \approx Z(0)\left(\sum_{t=1}^{T} \omega^{(t)}/T\right)$ where $Z(0) = 2^K$, which is the number of configurations of the hidden variables $\boldsymbol{h}$.

1.2 Estimating Posteriors using Mean-field

Recall that for evidence $\boldsymbol{e}$ we want to estimate posteriors $P(\boldsymbol{h} \mid \boldsymbol{e}) = \sum_{\boldsymbol{x} \in \Omega(\boldsymbol{e})} P_{\Omega(\boldsymbol{e})}(\boldsymbol{h}, \boldsymbol{x} \mid \boldsymbol{e})$. Assume that the evidences can be expressed in term of boxed constraints, which lead to the following factorisation

$$P(\boldsymbol{x} \mid \boldsymbol{e}, \boldsymbol{h}) = \prod_i P(x_i \mid \boldsymbol{e}, \boldsymbol{h})$$

This factorisation is critical because it ensures that there are no deterministic constraints among $\{x_i\}_{i=1}^n$, which are the conditions that variational methods such as mean-fields would work well. This is because mean-field solution will generally not satisfy deterministic constraints, and thus may assign non-zeros probability to improbably areas.

To be more concrete, the mean-field approximation would be $Q(\boldsymbol{h}, \boldsymbol{x}) \approx P(\boldsymbol{h}, \boldsymbol{x} \mid \boldsymbol{e})$

$$\begin{aligned} Q(\boldsymbol{h}, \boldsymbol{x}) &= \prod_k Q_k(h_k) \prod_i Q_i(x_i) \\ \text{s.t.} \quad \boldsymbol{x} &\in \Omega(\boldsymbol{e}) \end{aligned}$$

The best mean-field approximation will be the minimiser of the Kullback-Leibler divergence

$$\begin{aligned} \mathcal{D}\left(Q||P\right) &= \sum_{\boldsymbol{h}} \sum_{\boldsymbol{x} \in \Omega(\boldsymbol{e})} Q(\boldsymbol{h}, \boldsymbol{x}) \log \frac{Q(\boldsymbol{h}, \boldsymbol{x})}{P(\boldsymbol{h}, \boldsymbol{x} \mid \boldsymbol{e})} \\ &= -H\left[Q(\boldsymbol{h}, \boldsymbol{x})\right] - \sum_{\boldsymbol{h}} \sum_{\boldsymbol{x} \in \boldsymbol{\Omega}(\boldsymbol{e})} Q(\boldsymbol{h}, \boldsymbol{x}) \log P(\boldsymbol{h}, \boldsymbol{x} \mid \boldsymbol{e}) \end{aligned} \tag{5}$$

where $H\left[Q(\boldsymbol{h}, \boldsymbol{x})\right]$ is the entropy function. Now first, exploit the fact that Q is factorisable, and thus its entropy is decomposable, i.e.,

$$H\left[Q(\boldsymbol{h}, \boldsymbol{x})\right] = \sum_k H\left[Q_k(h_k)\right] + \sum_i H\left[Q_i(x_i)\right] \tag{6}$$

Second recall from Eq. (11) that

$$P(\boldsymbol{h}, \boldsymbol{x} \mid \boldsymbol{e}) = \frac{1}{Z(\boldsymbol{e})} \exp\left\{-E(\boldsymbol{x}, \boldsymbol{h})\right\}$$

and thus

$$\sum_{\boldsymbol{h}} \sum_{\boldsymbol{x} \in \boldsymbol{\Omega}(\boldsymbol{e})} Q(\boldsymbol{h}, \boldsymbol{x}) \log P(\boldsymbol{h}, \boldsymbol{x} \mid \boldsymbol{e}) = -\sum_{\boldsymbol{h}} \sum_{\boldsymbol{x} \in \boldsymbol{\Omega}(\boldsymbol{e})} Q(\boldsymbol{h}, \boldsymbol{x}) E(\boldsymbol{x}, \boldsymbol{h}) - \log Z(\boldsymbol{e})$$

Since $\log Z(\boldsymbol{e})$ is a constraint w.r.t. $Q(\boldsymbol{h}, \boldsymbol{x})$, we can safely ignore it here.

Now since $E(\boldsymbol{x}, \boldsymbol{h})$ is decomposable (see Eq. (2)), we have

$$\begin{aligned} \sum_{\boldsymbol{h}} \sum_{\boldsymbol{x} \in \boldsymbol{\Omega}(\boldsymbol{e})} Q(\boldsymbol{h}, \boldsymbol{x}) E(\boldsymbol{x}, \boldsymbol{h}) &= \left(\sum_i \sum_{x_i \in \Omega(e_i)} Q_i(x_i) E_i(x_i) \right) + \left(\sum_k \sum_{h_k} Q_k(h_k) E_k(h_k) \right) + \\ &\quad + \left(\sum_i \sum_k \sum_{x_i \in \Omega(e_i)} \sum_{h_k} Q_i(x_i) Q_k(h_k) E_{ik}(x_i, h_k) \right) \end{aligned}$$

where

$$\begin{aligned} E_i(x_i) &= \frac{x_i^2}{2} - \alpha_i x_i \\ E_k(h_k) &= -\gamma_k h_k \\ E_{ik}(x_i, h_k) &= -W_{ik} x_i h_k \end{aligned}$$

Combining this decomposition and Eq. (6), we have completely decomposed the Kullback-Leibler divergence in Eq. (5) into local terms:

$$\begin{aligned} \mathcal{D}\left(Q||P\right) &= \sum_i \sum_{x_i \in \Omega(e_i)} Q_i(x_i) E_i(x_i) + \sum_k \sum_{h_k} Q_k(h_k) E_k(h_k) + \\ &+ \left(\sum_i \sum_k \sum_{x_i \in \Omega(e_i)} \sum_{h_k} Q_i(x_i) Q_k(h_k) E_{ik}(x_i, h_k) \right) - \sum_i H\left[Q_i(x_i)\right] - \sum_k H\left[Q_k(h_k)\right] \end{aligned}$$

Now we wish to minimise the divergence with respect to the local distributions $\{Q_i(x_i), Q_k(h_k)\}$ for $i = 1, 2, ..., N$ and$k = 1, 2, ..., K$ knowing the proper distribution constraints

$$\begin{aligned} \int_{x_i \in \Omega(e_i)} Q_i(x_i) &= 1 \\ \sum_{h_k} Q_k(h_k) &= 1 \end{aligned}$$

By the method of Lagrangian multiplier, we have

$$L(\lambda) = \mathcal{D}\left(Q||P\right) + \sum_i \lambda_i \left(\int_{x_i \in \Omega(e_i)} Q_i(x_i) - 1 \right) + \sum_k \kappa_k \left(\sum_{h_k} Q_k(h_k) - 1 \right)$$

- Let us compute the partial derivative w.r.t. $Q_i(x_i)$:

$$\partial_{Q_i(x_i)} L(\lambda) = \log Q_i(x_i) + 1 + E_i(x_i) + \sum_k Q_k(h_k^1) E_{ik}(x_i, h_k^1) + \lambda_i$$

  where h_k^1 is a short hand for $h_k = 1$ and we have made use of the fact that $E_{ik}(x_i, h_k = 0) = 0$. Setting this gradient to zero yields

$$\begin{aligned} Q_i(x_i) &= \exp\left\{ -\left(E_i(x_i) + \sum_k Q_k(h_k^1) E_{ik}(x_i, h_k^1) \right) - 1 - \lambda_i \right\} \\ &= \exp\left\{ -\frac{1}{2}\left(x_i - \left(\alpha_i + \sum_k Q_k(h_k^1) W_{ik} \right) \right)^2 - 1 - \lambda_i \right\} \end{aligned} \tag{7}$$

  for $x_i \in \Omega(e_i)$. Normalising this distribution would lead to the truncated form of the normal distribution those the mean is

$$\mu_i = \alpha_i + \sum_k Q_k(h_k^1) W_{ik} h_k \tag{8}$$

- In a similar way, the partial derivative w.r.t. $Q_k(h_k)$ would be

$$\partial_{Q_k(h_k)} L(\lambda) = \log Q_k(h_k) - h_k \left(\gamma_k + \sum_i W_{ik} \sum_{x_i \in \Omega(e_i)} Q_i(x_i) x_i \right) + 1 + \kappa_k$$

Equating the gradient to zero, we have

$$Q_k(h_k) \propto \exp \left\{ h_k \left(\gamma_k + \sum_i W_{ik} \hat{\mu}_i \right) \right\}$$

where $\hat{\mu}_i$ is the mean of the truncated normal distribution

$$\hat{\mu}_i = \sum_{x_i \in \Omega(e_i)} Q_i(x_i) x_i$$

Normalising $Q_k(h_k)$ would lead to

$$Q_k(h_k^1) = \left[1 + \exp \left\{ -\gamma_k - \sum_i W_{ik} \hat{\mu}_i \right\} \right]^{-1} \tag{9}$$

- Finally, combining these findings in Eqs. (7,8,9), and letting $\Omega(e_i) = [b_i, c_i]$ be the boxed constraint, and using the fact that the mean of the truncated distribution is

$$\hat{\mu}_i = \mu_i + \frac{\phi(b_i - \mu_i) - \phi(c_i - \mu_i)}{\Phi(c_i - \mu_i) - \Phi(b_i - \mu_i)}$$

we would arrive at the three recursive equations

$$\begin{aligned} q_k &\leftarrow \frac{1}{1 + \exp\{-\gamma_k - \sum_i W_{ik} \hat{\mu}_i\}} \\ \mu_i &\leftarrow \alpha_i + \sum_k W_{ik} q_k \\ \hat{\mu}_i &\leftarrow \mu_i + \frac{\phi(b_i - \mu_i) - \phi(c_i - \mu_i)}{\Phi(c_i - \mu_i) - \Phi(b_i - \mu_i)} \end{aligned}$$

where q_k is a short hand for $Q_k(h_k^1)$ and$\phi(z)$ is the normal probability density function, and $\Phi(z)$ is the cumulative distribution function ♣

1.3 Seeking Modes and Generating Representative Samples

Once the model has been learned, samples can be generated straightforwardly by first sampling the underlying Gaussian RBM and then collect the true samples that satisfy the inequalities of interest. For example, for binary samples, if the generated Gaussian value for a visible unit is larger than the

threshold, then we have an active sample. Likewise, rank samples, we only need to rank the sampled Gaussian values.

However, this may suffer from the poor mixing if we use standard Gibbs sampling, that is the Markov chain may get stuck in some energy traps. To jump out of the trap we propose to periodically raise the temperature to a certain level (e.g., 10) and then slowly cool down to the original temperature (which is 1). In our experiment, the cooling is scheduled as follows

$$T \leftarrow \eta T$$

where $\eta \in (0, 1)$ is estimated so that for n steps, the temperature will drop from T_{max} to T_{min}. That is, $T_{min} = \eta^n T_{max}$, leading to $\eta = (T_{min}/T_{max})^{1/n}$.

To locate a basis of attraction, we can lower the temperature further (e.g., to 0.1) to trap the particles there. Then we collect k successive samples and take the average to be the representative sample. In our experiments, $k = 50$.

2 Learning

2.1 Gradient of the Likelihood

The log-likelihood of an evidence is

$$\begin{aligned}
\mathcal{L} &= \log P(\boldsymbol{e}) = \log \sum_{\boldsymbol{h}} \int_{\boldsymbol{\Omega}(\boldsymbol{e})} P(\boldsymbol{h}, \boldsymbol{x}) d\boldsymbol{x} \\
&= \log \sum_{\boldsymbol{h}} \int_{\boldsymbol{\Omega}(\boldsymbol{e})} \exp\left\{-E(\boldsymbol{x}, \boldsymbol{h})\right\} d\boldsymbol{x} - \log Z \\
&= \log Z(\boldsymbol{e}) - \log Z
\end{aligned}$$

where $Z(\boldsymbol{e}) = \sum_{\boldsymbol{h}} \int_{\boldsymbol{\Omega}(\boldsymbol{e})} \exp\left\{-E(\boldsymbol{x}, \boldsymbol{h})\right\} d\boldsymbol{x}$. The gradient of $\log Z(\boldsymbol{e})$ w.r.t. the mapping parameter W_{ik} reads

$$\begin{aligned}
\partial_{W_{ik}} \log Z(\boldsymbol{e}) &= \frac{-1}{Z(\boldsymbol{e})} \sum_{\boldsymbol{h}} \int_{\boldsymbol{\Omega}(\boldsymbol{e})} \exp\left\{-E(\boldsymbol{x}, \boldsymbol{h})\right\} \partial_{W_{ik}} E(\boldsymbol{x}, \boldsymbol{h}) d\boldsymbol{x} \\
&= -\sum_{\boldsymbol{h}} \int_{\boldsymbol{\Omega}(\boldsymbol{e})} P(\boldsymbol{x}, \boldsymbol{h} \mid \boldsymbol{e}) \partial_{W_{ik}} E(\boldsymbol{x}, \boldsymbol{h}) d\boldsymbol{x} \qquad (10)
\end{aligned}$$

where we have moved the constant $Z^{-1}(\boldsymbol{e})$ into the sum and integration and make use of the fact that

$$P(\boldsymbol{x}, \boldsymbol{h} \mid \boldsymbol{e}) = \frac{P(\boldsymbol{x}, \boldsymbol{h})}{P(\boldsymbol{e})} = \frac{1}{Z(\boldsymbol{e})} \exp\left\{-E(\boldsymbol{x}, \boldsymbol{h})\right\} \qquad (11)$$

where the domain of the Gaussian is constrained to $\boldsymbol{x} \in \boldsymbol{\Omega}(\boldsymbol{e})$.

From the definition of the energy function in Eq. (2), we know that the energy is decomposable, and thus the gradient w.r.t. W_{ik} only involves the pair (x_i, h_k). In particular

$$\partial_{W_{ik}} E(\boldsymbol{x}, \boldsymbol{h}) = -x_i h_k$$

This simplifies Eq. (10)

$$\begin{aligned} \partial_{W_{ik}} \log Z(\boldsymbol{e}) &= -\sum_{h_i} \int_{\Omega(e_i)} P(x_i, h_k \mid \boldsymbol{e}) \partial_{W_{ik}} E(\boldsymbol{x}, \boldsymbol{h}) dx_i \\ &= \mathbb{E}_{P(\boldsymbol{x}, \boldsymbol{h} \mid \boldsymbol{e})} [x_i h_k] \end{aligned}$$

A similar process would lead to

$$\partial_{W_{ik}} \log Z = \mathbb{E}_{P(x_i, h_k)} [x_i h_k]$$

and finally:

$$\partial_{W_{ik}} \mathcal{L} = \mathbb{E}_{P(x_i, h_k \mid \boldsymbol{e})} [x_i h_k] - \mathbb{E}_{P(x_i, h_k)} [x_i h_k] \clubsuit$$

2.2 Regularising the Markov Chains

One undesirable feature of the MCMC chains used in learning we have experiences so far is the tendency for the binary hidden states to get stuck, i.e., after some point they do not flip their assignments as learning progresses. We conjecture that this phenomenon may be due to the *saturation effect* inherent in the factor posterior:

$$P(h_k = 1 \mid \boldsymbol{x}) = \frac{1}{1 + \exp\left(-\gamma_k - \sum_i W_{ik} x_i\right)}$$

i.e., once the collected value to a node $(\gamma_k + \sum_i W_{ik} x_i)$ is too high or too low, it is very hard to over turn.

Fortunately, there is a known technique to regularise the chain: we enforce that at a time, there should be only a fraction ρ of nodes which are active, where $\rho \in (0, 1)$. One way is to maximise the following objective function

$$\mathcal{L}_2 = \mathcal{L} + \lambda \int \left[\sum_k \sum_{h_k} \rho(h_k) \log P(h_k \mid \boldsymbol{x}) \right] P(\boldsymbol{x} \mid \boldsymbol{e}) d\boldsymbol{x}$$

where $\lambda > 0$ is the weighting factor, $\rho(h_k) = \rho$ if $h_k = 1$ and $\rho(h_k) = 1 - \rho$ otherwise.

The gradient with respect to $g_k = (\gamma_k + \sum_i W_{ik} x_i)$ is then

$$
\begin{aligned}
\partial_{g_k}\mathcal{L}_2 &= \partial_{g_k}\mathcal{L} + \lambda \int \left[\partial_{g_k} \sum_{h_k} \rho(h_k) \log P(h_k \mid \boldsymbol{x})\right] P(\boldsymbol{x}|\boldsymbol{e}) d\boldsymbol{x} \\
&= \partial_{g_k}\mathcal{L} + \lambda \int \left[\rho - P(h_k^1 \mid \boldsymbol{x})\right] P(\boldsymbol{x}|\boldsymbol{e}) d\boldsymbol{x} \\
&\approx \partial_{g_k}\mathcal{L} + \frac{1}{S}\lambda \sum_s \left[\rho - P(h_k^1 \mid \boldsymbol{x}^{(s)})\right]
\end{aligned}
$$

where S is the number of samples and $P(h_k^1 \mid \boldsymbol{x})$ is a shorthand for $P(h_k = 1 \mid \boldsymbol{x})$. Using the chain rule, we have:

$$
\begin{aligned}
\partial_{\gamma_k}\mathcal{L}_2 &\approx \partial_{\gamma_k}\mathcal{L} + \frac{1}{S}\lambda \sum_s \left[\rho - P(h_k^1 \mid \boldsymbol{x}^{(s)})\right] \\
\partial_{W_{ik}}\mathcal{L}_2 &\approx \partial_{w_{ikd}}\mathcal{L} + \frac{1}{S}\lambda \sum_s x_i \left[\rho - P(h_k^1 \mid \boldsymbol{x}^{(s)})\right]
\end{aligned}
$$

2.3 Online Estimation of Posteriors

For tasks such as data completion (e.g., collaborative filtering) we need the posteriors $P(\boldsymbol{h} \mid \boldsymbol{e})$ for the prediction phase. One way is to run the Markov chain or doing mean-field from scratch. Here we suggest a simple way to obtain an approximation directly from the training phase without any further cost. The idea is to update the estimated posterior $\hat{\boldsymbol{h}}$ at each learning step t in an *exponential smoothing* fashion:

$$
\hat{\boldsymbol{h}}^{(t)} \leftarrow \eta \hat{\boldsymbol{h}}^{(t-1)} + (1-\eta)\hat{\boldsymbol{h}}^{(t)}
$$

for some *smoothing factor* $\eta \in (0, 1)$ and initial $\hat{\boldsymbol{h}}^{(0)}$, where $\bar{h}_k^{(t)} = P(h_k^1 \mid \boldsymbol{x}^{(t)}, \boldsymbol{e})$ and $\boldsymbol{x}^{(t)}$ is the sampled Gaussian at time t.

As learning progresses, early samples, which are from incorrect models, will be exponentially weighted down. Typically we choose η close to 1, e.g., $\eta = 0.9$.

2.4 Monitoring the Learning Progress

It is often of practical importance to track the learning progress, either by the reconstruction errors or by the data likelihood. The data likelihood can be estimated as

$$
P(\boldsymbol{e}) \approx \frac{1}{S} \sum_{s=1}^{S} \int_{\boldsymbol{\Omega}(\boldsymbol{e})} P(\boldsymbol{x} \mid \boldsymbol{h}^{(s)}) d\boldsymbol{x}
$$

where $\boldsymbol{h}^{(s)}$ are those samples collected as learning progressed in the data-independent phase, and the integration can be carried out using the technique described in the main text.

3 Extreme Value Distributions

Extreme value distributions are a class of distributions of extremal measurements Gumbel [1958]. Here we are concerned about the popular Gumbel's distribution.

3.1 Gumbel Distribution for Categorical Choices

Let us start from the Gumbel density function

$$P(x) = \frac{1}{\sigma} \exp\left\{ -\left(\frac{x-\mu}{\sigma} + e^{-\frac{x-\mu}{\sigma}} \right) \right\}$$

where μ is the mode (location) and σ is the scale parameter.

Using Laplace's approximation (e.g., via Taylor's expansion of $\left(\frac{x-\mu}{\sigma} + e^{-\frac{x-\mu}{\sigma}} \right)$ using the second-order polynomial around μ), we have

$$P(x) \propto \frac{1}{e\sigma} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}$$

Renormalising this distribution, e.g., by replacing $e \approx 2.7183$ by $\sqrt{2\pi} \approx 2.5066$ we obtain the standard Gaussian distribution. Thus, *we can use the Gumbel as an approximation to the Gaussian distribution*.

Now we turn to the categorical model using Gumbel variables. We maintain one variable per category, which plays the role of the utility for the category. Assume that all utilities share the same scale parameter σ. The existing literature McFadden [1973] asserts that the probability of choosing the m-th category is

$$P(e = c_m) = \frac{e^{\mu_m/\sigma}}{\sum_l e^{\mu_l/\sigma}}$$

where μ_l is the location of the l-th utility.

When we choose a the m-th category we must ensure that $x_m > \max_{l \neq m} x_l$.

Let $y_l = \exp\left(-\frac{x_l - \mu_l}{\sigma}\right)$ or equivalently $x_l = \mu_l - \sigma \log y_l$. Thus $x_l < x_m$ means $\mu_l - \sigma \log y_l < \mu_m - \sigma \log y_m$, or $y_l > y_m \exp\left(\frac{\mu_l - \mu_m}{\sigma}\right)$.

The CDF of the l-th Gumbel distribution is

$$\begin{aligned} F_l(x_m) &= \exp\left(-e^{-\frac{x_m - \mu_l}{\sigma}} \right) \\ &= \exp\left(-e^{-\frac{x_m - \mu_m}{\sigma}} e^{\frac{\mu_l - \mu_m}{\sigma}} \right) \\ &= \exp\left(-y_m e^{\frac{\mu_l - \mu_m}{\sigma}} \right) \end{aligned}$$

Thus choosing category c_m would mean

$$\begin{aligned} P(e = c_m) &= \int P(x_m) \left(\prod_{l \neq m} \int^{xm} P(x_l) dx_l \right) dx_m \\ &= \int P(x_m) \prod_{l \neq m} F_l(x_m) dx_m \end{aligned}$$

We rewrite the Gumbel density function by changing variable from x_m to y_m:

$$P(y_m) = \frac{1}{\sigma} y_m \exp\{-y_m\}$$

for $y_m \geq 0$. Thus

$$P(x_m) \prod_{l \neq m} F_l(x_m) = \frac{1}{\sigma} y_m \exp\left(-y_m \left\{ 1 + \sum_{l \neq m} e^{\frac{\mu_l - \mu_m}{\sigma}} \right\} \right)$$

Now, by changing variable under the integration from x_m to y_m, we have

$$\begin{aligned} P(e = c_m) &= \sigma \int_0^\infty P(y_m) \prod_{l \neq m} F_l(y_m) \frac{1}{y_l} dy_m \\ &= \int_0^\infty \exp\left(-y_m \left\{ 1 + \sum_{l \neq m} e^{\frac{\mu_l - \mu_m}{\sigma}} \right\} \right) dy_m \\ &= \frac{1}{1 + \sum_{l \neq m} e^{\frac{\mu_l - \mu_m}{\sigma}}} \\ &= \frac{e^{\mu_m/\sigma}}{\sum_l e^{\mu_l/\sigma}} \end{aligned}$$

3.2 Gumbel Distribution for Rank

We now extend the case of categorical evidences rank evidences. Again we maintain one Gaussian variable per category. Without loss of generality, for a particular rank π we assume that we must ensure that $x_1 > x_2 > ... > x_D$. This is equivalent to

$$\left[x_1 > \max_{l>1} x_l \right] \cap \left[x_2 > \max_{l>2} x_l \right] \cap ... \cap [x_{D-1} > x_D]$$

The probability of this is essentially

$$P\left(\{e_m = m\}_{m=1}^D \right) = P(e_1 = 1) \prod_{m>2} P\left(e_l = m \mid \{e_d = l\}_{l=1}^{m-1} \right)$$

In words, this offers a stagewise process to rank categories: first we pick the best category, the pick the second best from the remaining categories and so on (see also Fligner and Verducci [1988]).

The probability of picking the best category out of a subset is already given in Appendix 3.1:

$$\begin{aligned} P(e_1 = 1) &= \frac{e^{\mu_1/\sigma}}{\sum_{l\geq 1} e^{\mu_l/\sigma}} \\ P\left(e_l = m \mid \{e_d = l\}_{l=1}^{m-1}\right) &= \frac{e^{\mu_m/\sigma}}{\sum_{l\geq m} e^{\mu_l/\sigma}} \end{aligned}$$

This gives us the Plackett-Luce model Luce [1959], Plackett [1975] as mentioned in Stern [1990].

4 Global Attitude: Sample Questions

- **Q4** (*Ordinal*): [...] how would you describe the current economic situation in (survey country) – {*very good, somewhat good, somewhat bad, or very bad*}?
- **Q11a** (*Binary*): How do you think people in other countries of the world feel about China? – {*like, disliked*}?
- **Q35,35a** (*Category-ranking*): Which one of the following, if any, is hurting the world's environment the most/second-most {*India, Germany, China, Brazil, Japan, United States, Russia, Other*}?
- **Q76** (*Continuous*): How old were you at your last birthday?
- **Q85** (*Categorical*): What is your current employment situation {*A list of employment categories*}?

5 Other Supporting Materials

5.1 Laplace Approximation

Laplace approximation is the technique using a Gaussian distribution to approximate another distribution. For the univariate case, assume that the original density distribution has the form

$$P(x) \propto \exp\{-f(x)\}$$

First we find the mode μ of $P(x)$ or equivalently the minimiser of $f(x)$ given it exists. Then we apply Taylor's expansion

$$f(x) \approx f(\mu) + f''(\mu)\frac{(x-\mu)^2}{2}$$

The Gaussian approximation has the form

$$P^*(x) \propto \exp\left\{-f''(\mu)\frac{(x-\mu)^2}{2}\right\}$$

where $1/f''(\mu)$ is the new variance.

5.2 Some Properties of the Truncated Normal Distribution

For a normal distribution $P(x|\mu,\sigma)$ of mean μ and standard deviation σ truncated from both sides, i.e., $\alpha < x < \beta$, the new density reads

$$\bar{P}_{[\alpha,\beta]}(x\ |,\mu,\sigma) = \frac{Q(x^*)}{\sigma\left[\Phi(\beta^*) - \Phi(\alpha^*)\right]}$$

where

$$x^* \quad = \quad \frac{x-\mu}{\sigma}; \quad \alpha^* = \frac{\alpha-\mu}{\sigma}; \quad \beta^* = \frac{\beta-\mu}{\sigma}$$

and $Q(\cdot)$ and $\Phi(\cdot)$ are the probability density function and the cumulative distribution of the standard normal distribution, respectively. In particular, we are interested in the mean of the distribution $\bar{P}_{[\alpha,\beta]}$:

$$\bar{\mu} = \mu_i + \sigma\frac{Q(\alpha^*) - Q(\beta^*)}{\Phi(\beta^*) - \Phi(\alpha^*)}$$

Some special cases:

- When $\alpha = \beta$, this distribution reduces to the Dirac's delta.
- When $\alpha = -\infty$, we have a one-sided truncation from above since $\Phi(\alpha^*) = 0$.
- When $\beta = +\infty$, we obtain a one-sided truncation form below since $\Phi(\beta^*) = 0$.

References

M.A. Fligner and J.S. Verducci. Multistage ranking models. *Journal of the American Statistical Association*, 83(403):892–901, 1988.

EJ Gumbel. *Statistical of extremes*. Columbia University Press, New York, 1958.

R.D. Luce. *Individual choice behavior*. Wiley New York, 1959.

D. McFadden. Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics*, pages 105–142, 1973.

R.M. Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001.

R.L. Plackett. The analysis of permutations. *Applied Statistics*, pages 193–202, 1975.

H. Stern. Models for distributions on permutations. *Journal of the American Statistical Association*, 85(410):558–564, 1990.