Bayesian Estimation of Mutual Information with Missing Data

Bence Bolgár
Budapest University of Technology and Economics

bolgar@mit.bme.hu
June 27, 2021

Goal

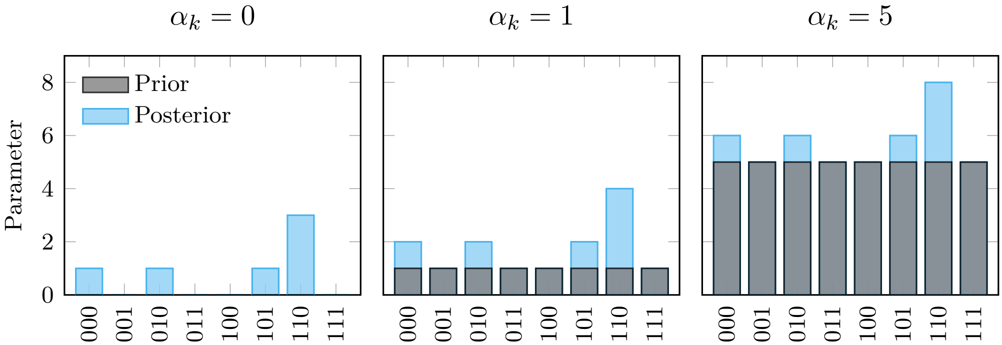
Estimate entropy and mutual information of partially observed, multivariate, binary variables.

$$
\text{Compounds}\begin{cases}
\boldsymbol{y}_1 = \\
\boldsymbol{y}_2 = \\
\boldsymbol{y}_3 = \\
\boldsymbol{y}_4 = \\
\boldsymbol{y}_5 =
\end{cases}
\underbrace{\begin{bmatrix} 0 & 1 & 0 \\ 1 & 1 & ? \\ ? & 0 & ? \\ 1 & 0 & ? \\ 0 & 1 & 0 \end{bmatrix}}_{d \text{ tasks}}
\quad \text{vs.} \quad
\begin{bmatrix} 1 & ? & 1 \\ 0 & 1 & ? \\ ? & 1 & 0 \\ ? & 0 & ? \\ 0 & 0 & ? \end{bmatrix}
$$

Problems:

- Handling a large number of empty bins?
- Handling missing observations?
- The number of bins grows as $\mathcal{O}\left(2^d\right)$.

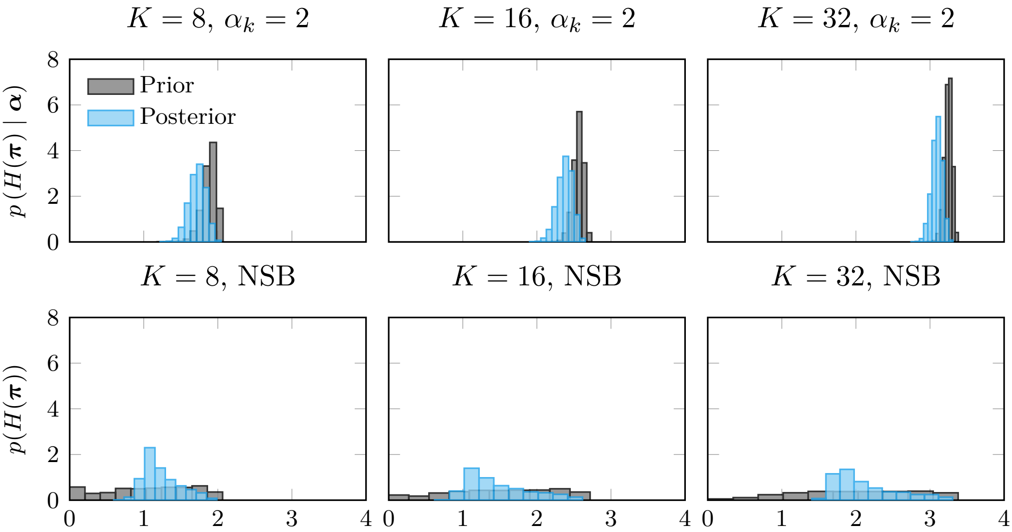
Dirichlet priors (Laplace smoothing)



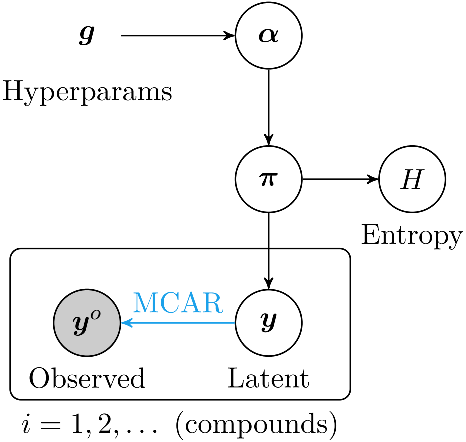
Entropy behaves badly under this smoothing ⇒ we have to integrate over $\boldsymbol{\alpha}$, *i.e.* use an infinite mixture of Dirichlet priors[1].

[1] Ilya Nemenman, F. Shafee, and William Bialek. “Entropy and Inference, Revisited”. In: *NIPS*. 2001, pp. 471–478.

Entropy distribution



Probabilistic model



Let $\boldsymbol{y} \in \{0, 1\}^d$ denote a row of the DTI matrix, which is partially observed. Our model is

$$
\begin{aligned}
p\left(\boldsymbol{\pi} \mid \boldsymbol{\alpha}\right) &= Dir\left(\boldsymbol{\pi} \mid \boldsymbol{\alpha}\right), \\
p\left(\boldsymbol{y} \mid \boldsymbol{\pi}\right) &= Mult\left(\boldsymbol{y} \mid \boldsymbol{\pi}\right), \\
p\left(\boldsymbol{y}^o \mid \boldsymbol{y}\right) &= MCAR\left(\boldsymbol{y}^o \mid \boldsymbol{y}\right),
\end{aligned}
$$

where $\boldsymbol{\alpha}, \boldsymbol{\pi} \in \mathbb{R}_+^K$, $K = 2^d$. The first goal is to compute the expected entropy

$$E\left[H(\boldsymbol{\pi}) \mid \boldsymbol{y}^o, \boldsymbol{g}\right],$$

which can be utilized to compute mutual information.

Expected entropy

The expected entropy can be written as

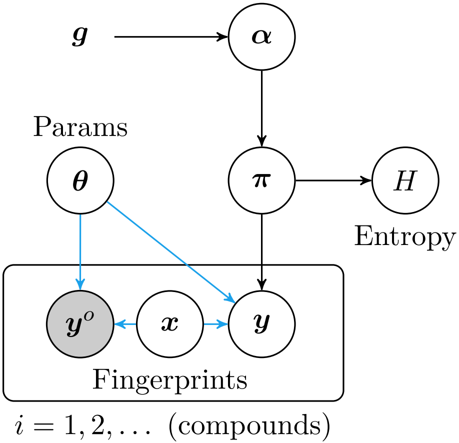
$$E\left[H(\boldsymbol{\pi}) \mid \boldsymbol{y}^o, \boldsymbol{g}\right] = \underbrace{\int \underbrace{\int \underbrace{\int \left[-\sum_{k=1}^{K} \pi_k \ln \pi_k\right] p\left(\boldsymbol{\pi} \mid \boldsymbol{y}, \boldsymbol{\alpha}\right) d\boldsymbol{\pi}}_{\text{analytical}} \, p\left(\boldsymbol{\alpha} \mid \boldsymbol{g}\right) d\boldsymbol{\alpha}}_{\text{Gaussian quadrature}} \, p\left(\boldsymbol{y} \mid \boldsymbol{y}^o\right) d\boldsymbol{y}}_{\text{Monte Carlo integration}} \,.$$

We utilize a different strategy to deal with each of the integrals:

1. Due to conjugacy, $p\left(\boldsymbol{\pi} \mid \boldsymbol{y}, \boldsymbol{\alpha}\right)$ is Dirichlet and its expected entropy can be calculated analytically,
2. We parameterize $\boldsymbol{\alpha}$ as in a previous work[2] and use Gaussian quadrature,
3. We use a variational strategy to sample from $p\left(\boldsymbol{y} \mid \boldsymbol{y}^o\right)$, implemented via a Bayesian neural network and proceed by Monte Carlo integration.

[2] Evan W Archer, Il Memming Park, and Jonathan W Pillow. "Bayesian entropy estimation for binary spike train data using parametric prior knowledge". In: *Advances in Neural Information Processing Systems*. Ed. by C. J. C. Burges et al. Vol. 26. Curran Associates, Inc., 2013, pp. 1700–1708.

Missing data



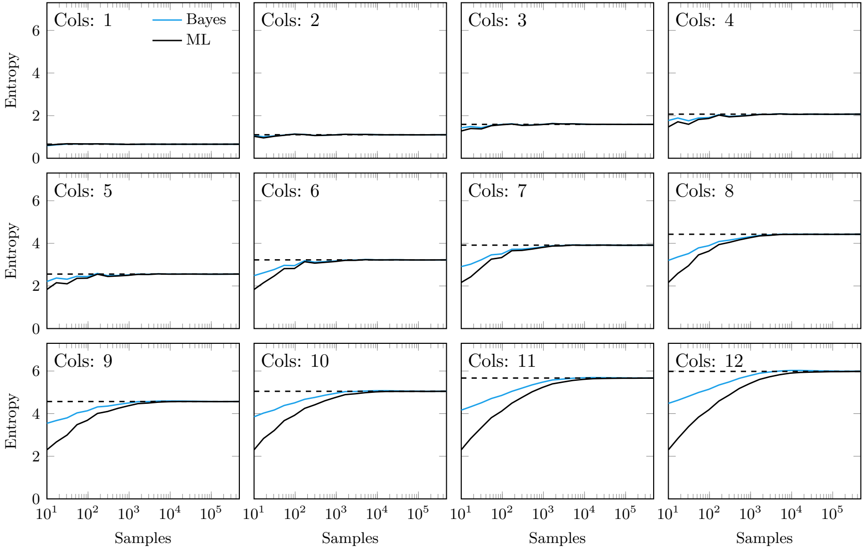
We still need to estimate $p\left(\boldsymbol{y} \mid \boldsymbol{y}^o\right)$, which we do by introducing the fingerprints $\boldsymbol{x}$ and NN parameters $\boldsymbol{\theta}$:

$$p\left(\boldsymbol{y} \mid \boldsymbol{y}^o\right) = \int p\left(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{x}\right) p\left(\boldsymbol{\theta}, \boldsymbol{x} \mid \boldsymbol{y}^o\right) d\boldsymbol{x} d\boldsymbol{\theta}.$$

Options are:

- Modelling $p\left(\boldsymbol{\theta}, \boldsymbol{x} \mid \boldsymbol{y}^o\right)$ (*e.g.* VAEs),
- Conditioning on $\boldsymbol{x}$ and modelling $p\left(\boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{y}^o\right)$ (*e.g.* Bayesian NNs),
- Conditioning on $(\boldsymbol{x}, \boldsymbol{\theta})$, *i.e.* obtaining $\boldsymbol{\theta}$ in a separate training (*e.g.* NNs).

Estimator bias



Mutual information

Mutual information can be estimated either by

- Using the entropy estimators for $\boldsymbol{y}^{(1)}$, $\boldsymbol{y}^{(2)}$ and $\boldsymbol{y}^{(1,2)}$ as

$$E\left[H \mid \boldsymbol{y}^{(1),o}, \boldsymbol{g}\right] + E\left[H \mid \boldsymbol{y}^{(2),o}, \boldsymbol{g}\right] - E\left[H \mid \boldsymbol{y}^{(1,2),o}, \boldsymbol{g}\right],$$

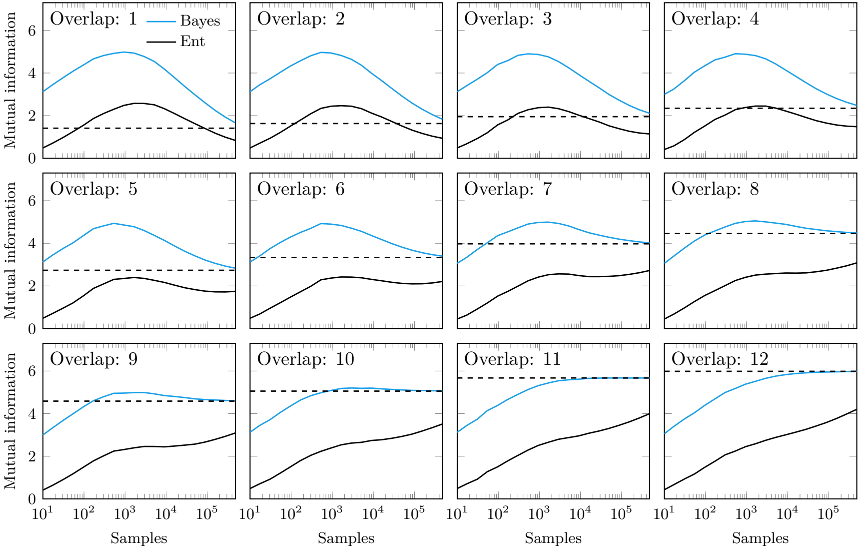
- Or in a "more Bayesian" manner as[3]

$$\int\int E\left[MI \mid \boldsymbol{y}^{(1,2),o}, \boldsymbol{g}, \alpha\right] p\left(\alpha \mid \boldsymbol{g}\right) p\left(\boldsymbol{y} \mid \boldsymbol{y}^{(1,2),o}\right) d\alpha d\boldsymbol{y},$$

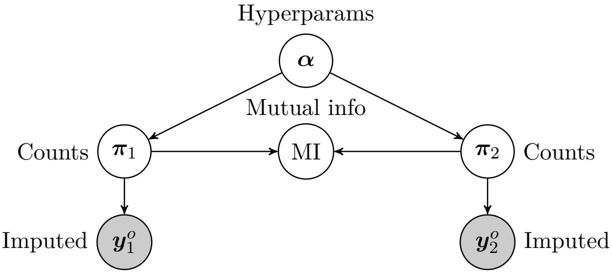
  where MI can be computed analytically for fixed values of α with a suitable prior $p\left(\alpha \mid \boldsymbol{g}\right)$, derived similarly to the NSB prior.

[3] Evan Archer, Il Park, and Jonathan Pillow. "Bayesian and Quasi-Bayesian Estimators for Mutual Information from Discrete Data". In: *Entropy* 15 (May 2013), pp. 1738–1755. DOI: 10.3390/e15051738.

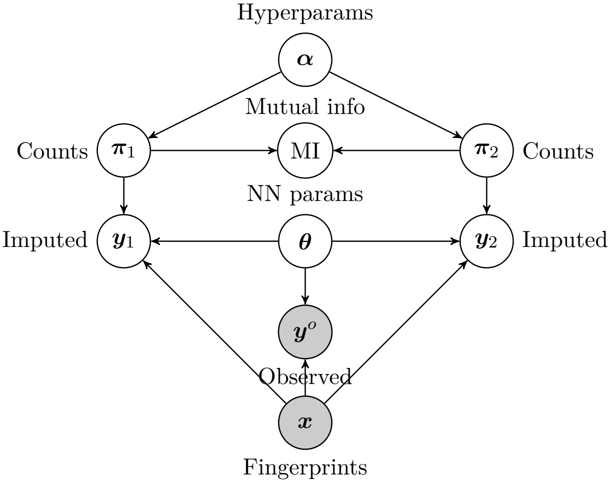
Estimator bias



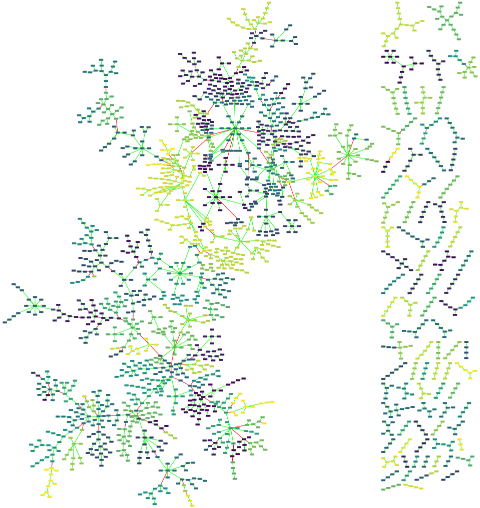
Missing data



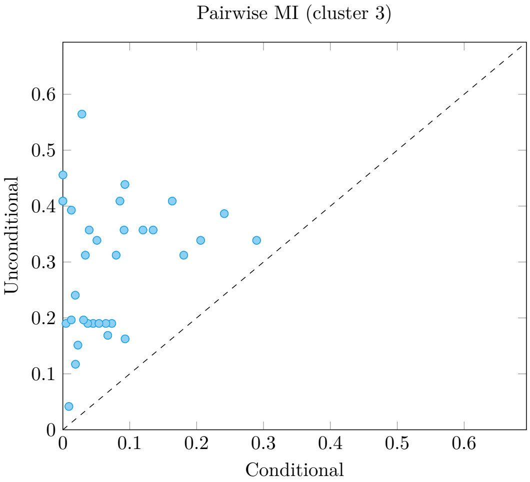
Missing data



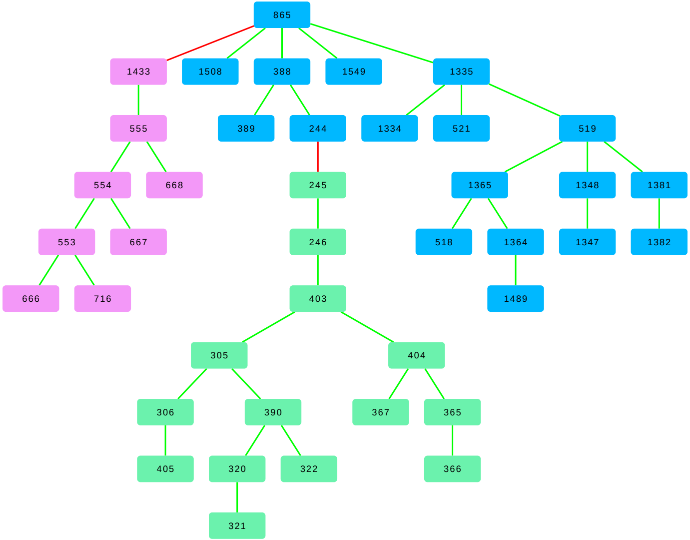
Task landscape (pairwise, unconditional)



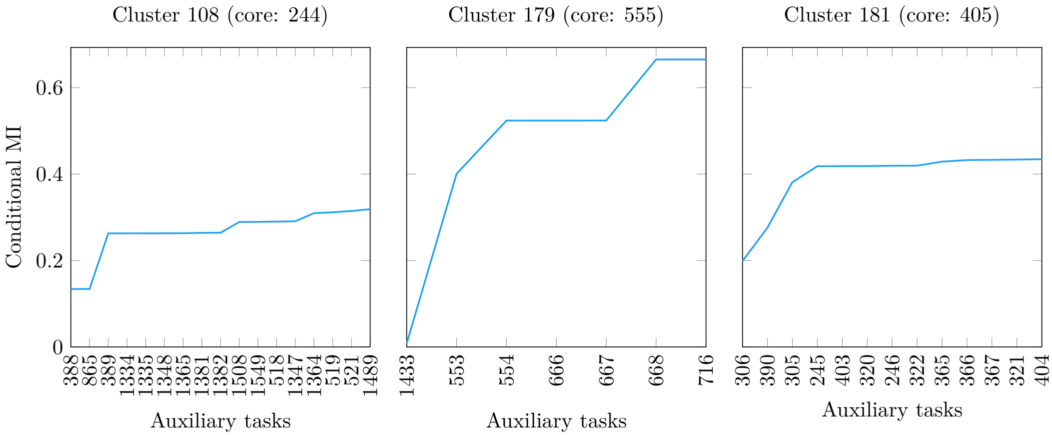
Conditional vs. unconditional MI



Clusters


865
1433
1508
388
1549
1335
555
389
244
1334
521
519
554
668
245
1365
1348
1381
553
667
246
518
1364
1347
1382
666
716
403
1489
305
404
306
390
367
365
405
320
322
366
321

Aggregate MI



Integration w.r.t. $\boldsymbol{\pi}$ – conjugacy

Using the conjugacy of the Dirichlet–Multinomial model, the posterior is

$$p(\boldsymbol{\pi} \mid \boldsymbol{y}, \boldsymbol{\alpha}) = Dir(\boldsymbol{\pi} \mid \hat{\boldsymbol{\alpha}}) = \frac{1}{Z(\hat{\boldsymbol{\alpha}})} \prod_{k=1}^{K} \pi_k^{\hat{\alpha}_k - 1} = \exp\left\{ \sum_{k=1}^{K} (\hat{\alpha}_k - 1) \cdot \ln \pi_k - \ln Z(\hat{\boldsymbol{\alpha}}) \right\},$$

where

$$\hat{\alpha}_k = \alpha_k + n_k, \quad \sum_{k=1}^{K} \pi_k = 1, \quad \pi_k > 0,$$

where n_k is the number of the instances in the kth category in $\boldsymbol{y}$, and the partition function is

$$Z(\hat{\boldsymbol{\alpha}}) = \frac{\prod_{k=1}^{K} \Gamma(\hat{\alpha}_k)}{\Gamma\left(\sum_{k=1}^{K} \hat{\alpha}_k\right)}.$$

Integration w.r.t. $\boldsymbol{\pi}$ – cumulants

Using the fact that Dir is an exponential family distribution, we have

$$
\begin{aligned}
\int_{\mathcal{S}} \ln \pi_k Dir\left(\boldsymbol{\pi} \mid \hat{\boldsymbol{\alpha}}\right) d\boldsymbol{\pi} &= \frac{\partial \ln Z}{\partial \hat{\alpha}_k} \\
&= \frac{1}{Z(\hat{\boldsymbol{\alpha}})}\left[\frac{\Gamma'(\hat{\alpha}_k)\prod_{j\neq k}\Gamma(\hat{\alpha}_j)}{\Gamma(m)} - \frac{\Gamma'(m)\prod_{j=1}^{K}\Gamma(\hat{\alpha}_j)}{\Gamma^2(m)}\right] \\
&= \frac{1}{Z(\hat{\boldsymbol{\alpha}})}\left[\frac{\prod_{j=1}^{K}\Gamma(\hat{\alpha}_j)}{\Gamma(m)}\left(\Psi(\hat{\alpha}_k) - \Psi(m)\right)\right] \\
&= \Psi(\hat{\alpha}_k) - \Psi(m),
\end{aligned}
$$

where $\mathcal{S}$ denotes the simplex, Ψ is the digamma function and

$$
m = \sum_{k=1}^{K} \hat{\alpha}_k.
$$

Integration w.r.t. $\boldsymbol{\pi}$ – expected entropy

From the previous results, for the Dirichlet expected entropy we have

$$
\begin{aligned}
E\left[H(\boldsymbol{\pi}) \mid \boldsymbol{\alpha}, \boldsymbol{y}\right] &= \int_{\mathcal{S}} \left[-\sum_{k=1}^{K} \pi_k \ln \pi_k\right] p\left(\boldsymbol{\pi} \mid \boldsymbol{y}, \boldsymbol{\alpha}\right) d\boldsymbol{\pi} \\
&= -\sum_{k=1}^{K} \frac{\Gamma\left(m\right)}{\prod_{j=1}^{K} \Gamma(\hat{\alpha}_j)} \int_{\mathcal{S}} \pi_k \ln \pi_k \prod_{j=1}^{K} \pi_j^{\hat{\alpha}_j - 1} d\boldsymbol{\pi} \\
&= -\sum_{k=1}^{K} \frac{\frac{1}{m}\Gamma\left(m+1\right)}{\frac{1}{\hat{\alpha}_k} \prod_{j=1}^{K} \Gamma(\hat{\alpha}_j + \delta_{jk})} \int_{\mathcal{S}} \ln \pi_k \prod_{j=1}^{K} \pi_j^{\hat{\alpha}_j - 1 + \delta_{jk}} d\boldsymbol{\pi} \\
&= \sum_{k=1}^{K} \frac{\hat{\alpha}_k}{m} \left(\Psi(m+1) - \Psi(\hat{\alpha}_k + 1)\right) \\
&= \Psi(m+1) - \sum_{k=1}^{K} \frac{\hat{\alpha}_k}{m} \Psi(\hat{\alpha}_k + 1).
\end{aligned}
$$

Integration w.r.t. $\boldsymbol{\alpha}$ – parameterization

Now we turn to the integral

$$\int E\left[H(\boldsymbol{\pi}) \mid \boldsymbol{\alpha}, \boldsymbol{y}\right] p\left(\boldsymbol{\alpha} \mid \boldsymbol{g}\right) d\boldsymbol{\alpha} = \int \left[\Psi(m+1) - \sum_{k=1}^{K} \frac{\hat{\alpha}_k}{m} \Psi(\hat{\alpha}_k + 1)\right] p\left(\boldsymbol{\alpha} \mid \boldsymbol{g}\right) d\boldsymbol{\alpha}.$$

In order to make it tractable, we parameterize $\boldsymbol{\alpha}$ as

$$\alpha_k := \alpha \cdot g_k$$

using a fixed parameter vector $\boldsymbol{g}$ with $\sum_{k=1}^{K} g_k := G$. The bracketed term now reads

$$\Psi\left(\alpha G + N + 1\right) - \sum_{k=1}^{K} \frac{\alpha g_k + n_k}{\alpha G + N} \Psi(\alpha g_k + n_k + 1),$$

where N is the number of instances in $\boldsymbol{y}$.

Integration w.r.t. $\boldsymbol{\alpha}$ – prior

To evaluate the integral, we also need a prior $p\,(\alpha \mid \boldsymbol{g})$. *A priori*, the Dirichlet expected entropy is

$$U_{\boldsymbol{g}}(\alpha) := \Psi\,(\alpha G + 1) - \sum_{k=1}^{K} \frac{g_k}{G} \Psi(\alpha g_k + 1).$$

Using the observation[4] that $p\,(H \mid \boldsymbol{\alpha})$ is "almost" a Dirac-δ at $U_{\boldsymbol{g}}(\alpha)$

$$p\,(\alpha \mid \boldsymbol{g}) = p\,(U_{\boldsymbol{g}}(\alpha)) \cdot \left|\frac{\partial U_{\boldsymbol{g}}}{\partial \alpha}\right| \approx p\,(H \mid \boldsymbol{\alpha}) \cdot \left|\frac{\partial U_{\boldsymbol{g}}}{\partial \alpha}\right|.$$

Since we want $p\,(H \mid \boldsymbol{\alpha})$ to be as uniform as possible,

$$p\,(\alpha \mid \boldsymbol{g}) \propto \left|\frac{\partial U_{\boldsymbol{g}}}{\partial \alpha}\right|.$$

[4] Nemenman, Shafee, and Bialek, "Entropy and Inference, Revisited".

Integration w.r.t. $\boldsymbol{\alpha}$ – prior

Thus, we specify the prior as

$$p\left(\alpha \mid \boldsymbol{g}\right) \propto \left|\frac{\partial U_{\boldsymbol{g}}}{\partial \alpha}\right| = G\Psi_1\left(\alpha G + 1\right) - \sum_{k=1}^{K} \frac{g_k^2}{G}\Psi_1(\alpha g_k + 1),$$

where Ψ_1 is the trigamma function, and the normalization constant is found to be

$$-\sum_{k=1}^{K} \frac{g_k}{G} \ln \frac{g_k}{G}.$$

Using these results, we can evaluate the integral w.r.t. α using Gaussian quadrature.

Mutual information

Given two task sets, an estimate of the mutual information can be computed from the expected entropies as

$$E\left[H(\boldsymbol{\pi}^{(1)}) \mid \boldsymbol{\alpha}^{(1)}, \boldsymbol{y}^{(1)}\right] + E\left[H(\boldsymbol{\pi}^{(2)}) \mid \boldsymbol{\alpha}^{(2)}, \boldsymbol{y}^{(2)}\right] - E\left[H(\boldsymbol{\pi}^{(1,2)}) \mid \boldsymbol{\alpha}^{(1,2)}, \boldsymbol{y}^{(1,2)}\right].$$

A Bayesian version of the former uses a suitable prior on α to compute

$$\begin{aligned}
\int \Bigg[&\Psi\left(\alpha G^{(1)} + N^{(1)} + 1\right) - \sum_{k=1}^{K^{(1)}} \frac{\alpha g_k^{(1)} + n_k^{(1)}}{\alpha G^{(1)} + N^{(1)}} \Psi(\alpha g_k^{(1)} + n_k^{(1)} + 1) \\
&+ \Psi\left(\alpha G^{(2)} + N^{(2)} + 1\right) - \sum_{k=1}^{K^{(2)}} \frac{\alpha g_k^{(2)} + n_k^{(2)}}{\alpha G^{(2)} + N^{(2)}} \Psi(\alpha g_k^{(2)} + n_k^{(2)} + 1) \\
&- \Psi\left(\alpha G^{(1,2)} + N^{(1,2)} + 1\right) + \sum_{k=1}^{K^{(1)} \cdot K^{(2)}} \frac{\alpha g_k^{(1,2)} + n_k^{(1,2)}}{\alpha G^{(1,2)} + N^{(1,2)}} \Psi(\alpha g_k^{(1,2)} + n_k^{(1,2)} + 1) \Bigg] \\
&\times p\left(\alpha \mid \boldsymbol{g}^{(1)}, \boldsymbol{g}^{(2)}, \boldsymbol{g}^{(1,2)}\right) d\alpha.
\end{aligned}$$

Mutual information prior

Let

$$V_{\boldsymbol{g}}(\alpha) := \Psi\left(\alpha G^{(1)} + 1\right) - \sum_{k=1}^{K^{(1)}} \frac{g_k^{(1)}}{G^{(1)}} \Psi(\alpha g_k^{(1)} + 1) + \Psi\left(\alpha G^{(2)} + 1\right) - \sum_{k=1}^{K^{(2)}} \frac{g_k^{(2)}}{G^{(2)}} \Psi(\alpha g_k^{(2)} + 1)$$
$$- \Psi\left(\alpha G^{(1,2)} + 1\right) + \sum_{k=1}^{K^{(1)} \cdot K^{(2)}} \frac{g_k^{(1,2)}}{G^{(1,2)}} \Psi(\alpha g_k^{(1,2)} + 1).$$

Similarly to the NSB estimator, we choose[5]

$$p\left(\alpha \mid \boldsymbol{g}^{(1)}, \boldsymbol{g}^{(2)}, \boldsymbol{g}^{(1,2)}\right) \propto \left|\frac{\partial V_{\boldsymbol{g}}}{\partial \alpha}\right|,$$

which leads to a bimodal prior with a nontrivial zero α_0. The normalizing constant can be found by first finding the zero numerically, then the function can be intgrated analytically on $[0, \alpha_0]$ and $[\alpha_0, \infty]$.

[5] Archer, Park, and Pillow, "Bayesian and Quasi-Bayesian Estimators for Mutual Information from Discrete Data".

Integration w.r.t. $\boldsymbol{y}$ – approximation

Opting for the Bayesian NN solution, we have

$$p\left(\boldsymbol{y} \mid \boldsymbol{y}^{o}, \boldsymbol{x}\right) = \int p\left(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{x}\right) p\left(\boldsymbol{\theta} \mid \boldsymbol{y}^{o}, \boldsymbol{x}\right) d\boldsymbol{\theta}$$

and the full expected quantities are

$$E\left[H \text{ or } MI \mid \boldsymbol{y}^{o}, \boldsymbol{x}, \boldsymbol{g}\right] = \int\int [*] \cdot p\left(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{x}\right) p\left(\boldsymbol{\theta} \mid \boldsymbol{y}^{o}, \boldsymbol{x}\right) d\boldsymbol{\theta} d\boldsymbol{y},$$

where $[*]$ stands for the previous Bayesian estimation of entropy or mutual information. We approximate the outermost two integrals by

1. Obtaining $p\left(\boldsymbol{\theta} \mid \boldsymbol{y}^{o}, \boldsymbol{x}\right)$ via standard variational inference,
2. Monte Carlo sampling of $\boldsymbol{\theta}$ and $\boldsymbol{y}$, which translates to stochastic forward passes in the Bayesian NN.

References

Archer, Evan, Il Park, and Jonathan Pillow. "Bayesian and Quasi-Bayesian Estimators for Mutual Information from Discrete Data". In: *Entropy* 15 (May 2013), pp. 1738–1755. DOI: 10.3390/e15051738.

Archer, Evan W, Il Memming Park, and Jonathan W Pillow. "Bayesian entropy estimation for binary spike train data using parametric prior knowledge". In: *Advances in Neural Information Processing Systems*. Ed. by C. J. C. Burges et al. Vol. 26. Curran Associates, Inc., 2013, pp. 1700–1708.

Nemenman, Ilya, F. Shafee, and William Bialek. "Entropy and Inference, Revisited". In: *NIPS*. 2001, pp. 471–478.