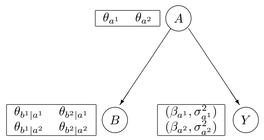
Improved Mean and Variance Approximations for Belief Net Responses via Network Doubling

Peter Hooper[1], Yasin Abbasi-Yadkori[2], Russ Greiner[2], and Bret Hoehn[2]
[1]Department of Mathematical & Statistical Sciences, [2]Department of Computing Science, University of Alberta

Introduction

Query: what is probability of $A = a$, given $(B = b, Y = y)$?



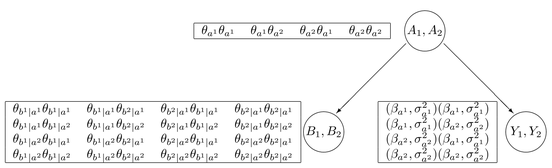
- A, B binary, Y Gaussian, naïve Bayes, $B \perp Y \mid A$
- $\theta_a = P(A = a)$, $\theta_{b|a} = P(B = b \mid A = a)$, $Y \mid A = a \sim N(\beta_a, \sigma^2_a)$
- Θ = vector containing all parameters $\theta_a, \theta_{b|a}, \beta_a, \sigma_a$

$$q(\Theta) = P(A = a \mid B = b, Y = y, \Theta) = \frac{\theta_a \theta_{b|a} p(y \mid \beta_a, \sigma_a)}{\sum_{a_1} \theta_{a_1} \theta_{b|a_1} p(y \mid \beta_{a_1}, \sigma_{a_1})}$$

Bayesian paradigm: Θ random, want posterior distribution of $q(\Theta)$

- Approx mean: replace $\theta_a, \theta_{b|a}$ by means; density by Student's t
- Approx variance: delta method [2]
  - uses approx local linearity of function q
- Approx distribution of $q(\Theta)$: Beta

Novel approach to approximations via network doubling:



- For replicates $(A_1, B_1, Y_1) \perp (A_2, B_2, Y_2) \mid \Theta$

$$q(\Theta)^2 = P(A_1 = A_2 = a \mid B_1 = B_2 = b, Y_1 = Y_2 = y, \Theta)$$
$$= \text{a query in doubled network}$$

- Compute approx mean of $q(\Theta)^2$
- Variance = $E\{q(\Theta)^2\} - [E\{q(\Theta)\}]^2$
- Doubling method uses minor changes in conditioning events
- Doubling works for discrete, continuous, and hybrid networks
- Our paper emphasizes implementation for discrete networks:
  - asymptotic analysis and adjustments to improve accuracy
  - numerical results and computational issues.

Assumptions and Notation

- Discrete network
- Θ = vector of all CPtable parameters
  $\theta_{b|\boldsymbol{a}} = P\{B = b \mid \boldsymbol{A} = \boldsymbol{a}, \Theta\}$, where $\boldsymbol{A}$ is vector of parents of B
- Conjugate prior: global and local independence, Dirichlet
- Complete data $\mathcal{D}$
- $q(\Theta) = E\{w(\boldsymbol{H}) \mid \boldsymbol{E} = \boldsymbol{e}, \Theta\}$
  - $\boldsymbol{H}, \boldsymbol{E}$ are vectors of hypothesis and evidence variables
  - w is an indicator function, e.g., $I_{\boldsymbol{h}}(\boldsymbol{H})$

Important Formula

Plug-in approximation of posterior mean $E\{q(\Theta) \mid \mathcal{D}\}$ [1]

$$E\{q(\Theta) \mid \mathcal{D}, \boldsymbol{e}\} = E\{w(\boldsymbol{H}) \mid \boldsymbol{E} = \boldsymbol{e}, \mathcal{D}\} = q(E\{\Theta \mid \mathcal{D}\}) \quad (1)$$

- $(\mathcal{D}, \boldsymbol{e})$ = augmented data, $\mathcal{D}$ plus partial observation $\boldsymbol{E}$
- Global independence needed, not local independence or Dirichlet

Network Doubling

Apply (1) to doubled network:

- Replicates $(B_1, \boldsymbol{A}_1, \boldsymbol{H}_1, \boldsymbol{E}_1) \perp (B_2, \boldsymbol{A}_2, \boldsymbol{H}_2, \boldsymbol{E}_2) \mid \Theta$
- Doubled CPtable entries: $\theta^*_{b_1 b_2|\boldsymbol{a}_1\boldsymbol{a}_2} = \theta_{b_1|\boldsymbol{a}_1}\theta_{b_2|\boldsymbol{a}_2}$
- $\boldsymbol{H}^* = (\boldsymbol{H}_1, \boldsymbol{H}_2)$, $\boldsymbol{E}^* = (\boldsymbol{E}_1, \boldsymbol{E}_2)$, $\boldsymbol{e}^* = (\boldsymbol{e}, \boldsymbol{e})$
- $w^*(\boldsymbol{H}^*) = w(\boldsymbol{H}_1)w(\boldsymbol{H}_2)$
- Result:

$$q(\Theta)^2 = E\{w^*(\boldsymbol{H}^*) \mid \boldsymbol{E}^* = \boldsymbol{e}^*, \Theta\}$$
$$q(\Theta) = E\{w(H_1) \mid \boldsymbol{E}^* = \boldsymbol{e}^*, \Theta\}$$

- Approx $\mathrm{Var}\{q(\Theta) \mid \mathcal{D}\}$ by

$$\mathrm{Var}\{q(\Theta) \mid \mathcal{D}, \boldsymbol{e}^*\} = E\{q(\Theta)^2 \mid \mathcal{D}, \boldsymbol{e}^*\} - [E\{q(\Theta) \mid \mathcal{D}, \boldsymbol{e}^*\}]^2 \quad (2)$$

- Doubled network retains global independence, so (1) valid:

$$E\{q^*(\Theta^*) \mid \mathcal{D}, \boldsymbol{e}^*\} = q^*(E\{\Theta^* \mid \mathcal{D}\}) \quad (3)$$

- $(\mathcal{D}, \boldsymbol{e}^*)$ = augmented data, $\mathcal{D}$ plus partial observations $(\boldsymbol{E}_1, \boldsymbol{E}_2)$
- Compute $E\{\Theta^* \mid \mathcal{D}\}$ from Dirichlet means and covariances
- Compute (2) using (3) with $q^*(\Theta^*) = q(\Theta)^2$ and $q^*(\Theta^*) = q(\Theta)$

Adjustments

Adjusted mean and variance approximations:

Means	Variances
$\hat{q}_1 = E\{q(\Theta) \mid \mathcal{D}, \boldsymbol{e}\}$	$\hat{v}_1$ = delta method [2]
$\hat{q}_2 = E\{q(\Theta) \mid \mathcal{D}, \boldsymbol{e}, \boldsymbol{e}\}$	$\hat{v}_2 = \mathrm{Var}\{q(\Theta) \mid \mathcal{D}, \boldsymbol{e}, \boldsymbol{e}\}$ in expression (2)
$\hat{q}_3 = \hat{q}_1 - (\hat{q}_2 - \hat{q}_1)$	$\hat{v}_3$ = expression (5)
$\hat{q}_4 = \hat{q}_1 - \hat{\sigma}_{qr}/\mu_r$	$\hat{v}_4$ = expression (6)

- Simple adjustments $(\hat{q}_3, \hat{v}_3)$, more complex $(\hat{q}_4, \hat{v}_4)$
- $Q = q(\Theta)$ and $R = P(\boldsymbol{E} = \boldsymbol{e} \mid \Theta)$
- Formulae involve moments of posterior distribution of (Q, R)
- $\hat{q}_4$ and $\hat{v}_4$ use $\mathrm{Cov}(Q, R)$ approximation:

$$\hat{\sigma}_{qr} = \frac{(\hat{q}_2 - \hat{q}_1)\mu_r(\mu_r^2 + \sigma_{rr})\{\mu_r(1 - \mu_r) + \sigma_{rr}\}}{\mu_r^3(1 - \mu_r) + \mu_r(1 - 2\mu_r)\sigma_{rr} - \sigma_{rr}^2} \quad (4)$$

- Adjusted variance approximations obtained by iteratively solving

$$\hat{v}_3 := \frac{\hat{v}_2 + 2(\hat{q}_2 - \hat{q}_1)^2}{1 + 4(\hat{q}_2 - \hat{q}_1)(1 - 2\hat{q}_3)/\{\hat{q}_3(1 - \hat{q}_3) + \hat{v}_3\}} \quad (5)$$

$$\hat{v}_4 := \frac{(\mu_r^2 + \sigma_{rr})\{\hat{v}_2 + (\hat{q}_2 - \hat{q}_4)^2\} - 2\hat{\sigma}_{qr}^2}{\mu_r^2 + \sigma_{rr} + 4\mu_r\hat{\sigma}_{qr}(1 - 2\hat{q}_4)/\{\hat{q}_4(1 - \hat{q}_4) + \hat{v}_4\}} \quad (6)$$

Asymptotic Analysis

As effective sample size $m \to \infty$ with $E\{\Theta \mid \mathcal{D}\}$ fixed:

- query mean μ_q remains constant
- query variance $\sigma_{qq} = O(m^{-1})$
- errors $\hat{q}_j - \mu_q = O(m^{-1})$ for $j = 1, 2$ and $O(m^{-3/2})$ for $j = 3, 4$
- relative errors $(\hat{v}_j - \sigma_{qq})/\sigma_{qq} = O(m^{-1})$ for $j = 1, 2, 3, 4$

Computational Issues

- Belief net inference is NP-complete
- Complexity of Variable Elimination (VE) Algorithm is $O(d^r)$
  - d = upper bound on variable domain size
  - r = upper bound on size of a factor generated by VE
- Doubling method uses technique for variance that is similar to evaluating a query, so has corresponding computational complexity
- Doubled CPtables are larger (squared number of rows and columns), so computational complexity of VE is $O(d^{2r})$
- Delta method retains $O(d^r)$ complexity, so faster in large networks
- Sometimes exploit network or query structure for polynomial time

Numerical Results

$\hat{q}_j$ and $\hat{v}_j$ compared with empirical estimates of μ_q and σ_{qq}:

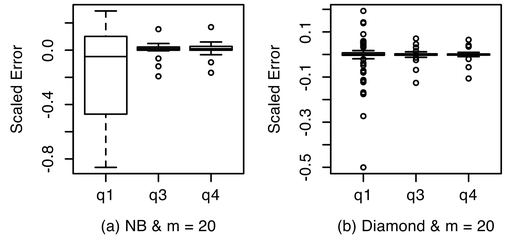
- simulate 10^6 replicates of Θ from posterior
- evaluate each $q(\Theta)$
- calculate sample mean and variance

Examples differ with respect to network structure, posterior, and query

- BDe posteriors, hyperparameters determined by m and $E\{\Theta \mid \mathcal{D}\}$
- Examples from three small networks, each with one vector $E\{\Theta \mid \mathcal{D}\}$ and $m \in \{20, 50, 100, 200, 500\}$:
  - Two naïve Bayes networks (NB-2 and NB-4 with 2 and 4 binary features plus the binary root variable H); $\boldsymbol{E}$ = all children of H, $\boldsymbol{e}$ varies over all combinations (2^2 for NB-2, 2^4 for NB-4)
  - Diamond network  , 4 binary variables, 108 distinct queries with one hypothesis variable

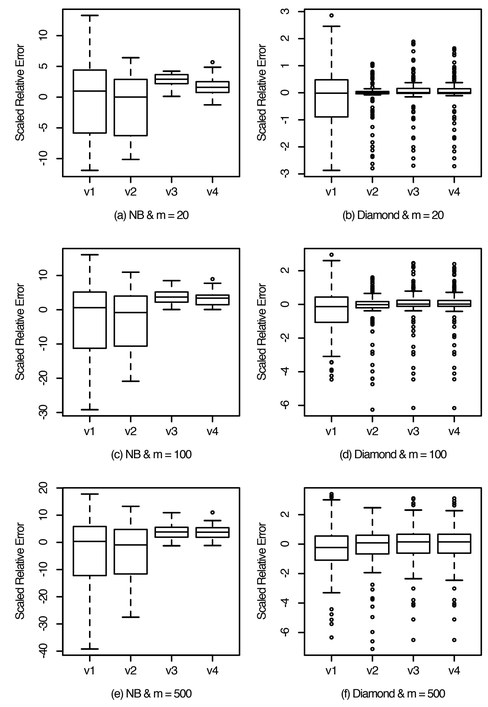
Results consistent with asymptotics.

- **Query mean:**
  - $\hat{q}_3 \approx \hat{q}_4$, both more accurate than $\hat{q}_1$



(a) NB & m = 20 (b) Diamond & m = 20

- **Query Variance:**
  - $\hat{v}_3 \approx \hat{v}_4$, both more accurate than $\hat{v}_1$
  - NB networks, $\hat{v}_2$ accuracy similar to $\hat{v}_1$
  - Diamond network, $\hat{v}_2$ accuracy similar to $\hat{v}_3$ and $\hat{v}_4$
  - Variation in empirical variances noticeable for $m = 500$



(a) NB & m = 20 (b) Diamond & m = 20 (c) NB & m = 100 (d) Diamond & m = 100 (e) NB & m = 500 (f) Diamond & m = 500

Continuous Variables

- Doubling applies to continuous and hybrid networks (parents of discrete variables must be discrete) but implementation more complex
- Continuous variables Gaussian, mean related linearly to continuous parents, coefficients and variance depending on discrete parents
- Conjugate prior is normal-(inverse chi-square)
- Predictive distribution substitutes means for CPtable parameters and t densities for normal (bivariate t for doubled network)
- Work in progress on computational issues

References

[1] G. Cooper and E. Herskovits (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning* **9**: 309–347.

[2] T. Van Allen, A. Singh, R. Greiner, and P. Hooper (2008). Quantifying the uncertainty of a belief net response: Bayesian error-bars for belief net inference. *Artificial Intelligence* **172**: 483–513.

UAI June 2009