SUPPLEMENTARY MATERIAL

THE POP-EB FULL BAYES APPROXIMATION

Recall the form of the population empirical Bayes (POP-EB) predictive density,

$$p(\boldsymbol{x}_{\text{new}} \mid \mathbf{X}) = \int p(\boldsymbol{x}_{\text{new}} \mid \mathbf{Z})\ p(\mathbf{Z} \mid \mathbf{X})\, \mathrm{d}\mathbf{Z}.$$

Expanding the conditional density at the end gives,

$$p(\boldsymbol{x}_{\text{new}} \mid \mathbf{X}) = \int p(\boldsymbol{x}_{\text{new}} \mid \mathbf{Z})\, \frac{p(\mathbf{X} \mid \mathbf{Z})\ F(\mathbf{Z})}{\int p(\mathbf{X} \mid \mathbf{Z}')\ F(\mathbf{Z}')\, \mathrm{d}\mathbf{Z}'}\, \mathrm{d}\mathbf{Z}.$$

The plug-in principle replaces F with the empirical distribution of data $\widehat{F}$. We then approximate the empirical distribution using the bootstrap by replacing $\widehat{F}$ with $\widehat{G}$. This leads to the approximation

$$p(\boldsymbol{x}_{\text{new}} \mid \mathbf{X}) \approx \sum_{\mathbf{Z}} p(\boldsymbol{x}_{\text{new}} \mid \mathbf{Z})\, \frac{p(\mathbf{X} \mid \mathbf{Z})\ \widehat{G}(\mathbf{Z})}{\sum_{\mathbf{Z}'} p(\mathbf{X} \mid \mathbf{Z}')\ \widehat{G}(\mathbf{Z}')},$$

because $\widehat{G}$ is a discrete distribution.

Specifically, $\widehat{G}$ is uniform over a set of $b = 1, \ldots, B$ bootstrapped datasets, which gives

$$p(\boldsymbol{x}_{\text{new}} \mid \mathbf{X}) \approx \sum_{b=1}^{B} p(\boldsymbol{x}_{\text{new}} \mid \mathbf{Z}^{(b)})\, \frac{p(\mathbf{X} \mid \mathbf{Z}^{(b)})\ \frac{1}{B}}{\sum_{b=1}^{B} p(\mathbf{X} \mid \mathbf{Z}^{(b)})\ \frac{1}{B}}.$$

The POP-EB full Bayes (FB) is then

$$p_{\text{FB}}(\boldsymbol{x}_{\text{new}} \mid \mathbf{X}) = \sum_{b=1}^{B} w_b\ p(\boldsymbol{x}_{\text{new}} \mid \mathbf{Z}^{(i)}),$$

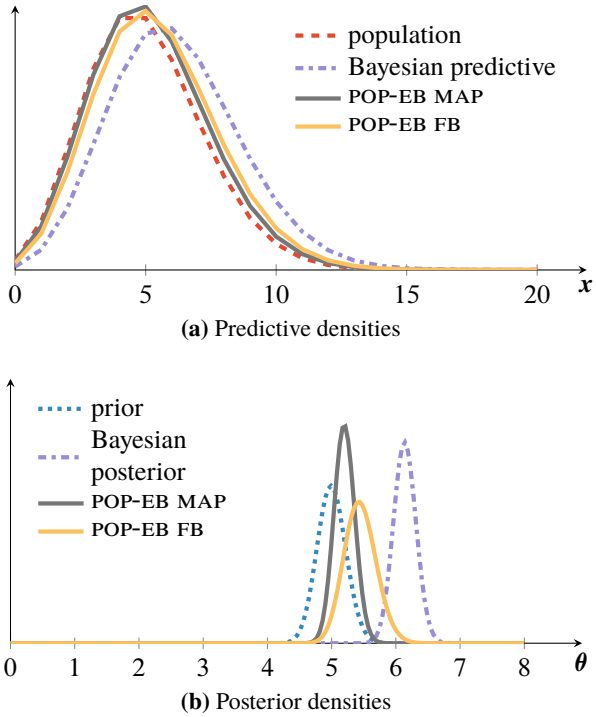
with weights

$$w_b = \frac{p(\mathbf{X} \mid \mathbf{Z}^{(b)})}{\sum_{b=1}^{B} p(\mathbf{X} \mid \mathbf{Z}^{(b)})}.$$

SIMULATION RESULTS WITH A SHARP PRIOR

Consider the model from the toy example in the paper. If we truly believe network failures are rare, we could posit an informative prior density. However, this has little effect in addressing model mismatch. Figure 1 shows the results of the same study under a very sharp prior centered at 5, $p(\boldsymbol{\theta}) = \text{Gam}(\alpha = 500,\ \beta = 100)$.

The Bayesian posterior is more accurate than before; it shifts closer to the dominant rate of $\boldsymbol{\theta} = 5$. This also moves the Bayesian predictive closer to the population. However, both POP-EB maximum a posteriori (MAP) and POP-EB full Bayes (FB) predictive densities still provide a better match to the true population.



(a) Predictive densities

(b) Posterior densities

Figure 1: Gamma-Poisson model with a sharp prior density centered at 5. The population in subpanel **(a)** has an additional small bump at 50 (not shown).

SIMULATION RESULTS WITH AN EMPIRICAL BAYES PRIOR

Empirical Bayes (EB) estimates the parameters of the prior density from the data. For simplicity, assume the prior is a Gamma distribution. One way to estimate the shape α and rate β parameters is to match the mean and variance of the Gamma distribution to that of the data.

The Gamma distribution has mean $= \alpha/\beta$ and variance $= \alpha/\beta^2$. This leads to the following pair of equations

$$\frac{\alpha}{\beta} = \text{Mean}(\mathbf{X}) = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n$$

$$\frac{\alpha}{\beta^2} = \text{Var}(\mathbf{X}) = \frac{1}{N}\sum_{n=1}^{N} (\boldsymbol{x}_n - \text{Mean}(\mathbf{X}))^2.$$

The solution is

$$\alpha = \frac{[\text{Mean}(\mathbf{X})]^2}{\text{Var}(\mathbf{X})} \text{ and } \beta = \frac{\text{Mean}(\mathbf{X})}{\text{Var}(\mathbf{X})}.$$

In our simulation study, these lead to estimates around $\alpha \approx 0.5$ and $\beta \approx 0.07$, which describes a nearly flat Gamma distribution. This does not help mitigate model mismatch, nor does it improve predictive accuracy.

EFFICIENT BUMP-VI IMPLEMENTATION

We first modify our Bayesian model notation to mimic stochastic variational inference (SVI) (Hoffman et al., 2013). Consider a Bayesian model $p(\mathbf{X}, \boldsymbol{\theta})$. Separate the latent variables $\boldsymbol{\theta}$ into a set of local $\boldsymbol{\zeta}$ and global $\boldsymbol{\beta}$ variables. Local latent variables $\boldsymbol{\zeta} = \{\boldsymbol{\zeta}_n\}_1^N$ grow with the number of observations; global latent variables $\boldsymbol{\beta}$ do not. The likelihood becomes $p(\boldsymbol{X} \mid \boldsymbol{\zeta}, \boldsymbol{\beta})$ and the prior $p(\boldsymbol{\zeta}, \boldsymbol{\beta})$.

Given the global variables $\boldsymbol{\beta}$, the local latent variable $\boldsymbol{\zeta}_n$, along with its observation $\boldsymbol{x}_n$, is conditionally independent of all other latent variables and observations

$$p(\boldsymbol{x}_n, \boldsymbol{\zeta}_n \mid \boldsymbol{x}_{-n}, \boldsymbol{\zeta}_{-n}, \boldsymbol{\beta}) = p(\boldsymbol{x}_n, \boldsymbol{\zeta}_n \mid \boldsymbol{\beta}).$$

The negative indexing notation means $\boldsymbol{x}_{-n} = \{\boldsymbol{x}_i \mid i = 1, \ldots, n-1, n+1, \ldots, N\}$. Global latent variables lack such conditional independence.

This divide is natural in many models. For example, consider latent Dirichlet allocation (LDA). The global latent variables are the topics. (The number of topics is fixed and does not vary with the number of documents.) The local latent variables are the per-document topic distributions and the per-word assignments. (There are as many of these variables as documents and words within each document.)

The local-global separation simplifies the computation of the B gradients in bumping variational inference (BUMP-VI). The variational family has two sets of variational parameters $q(\boldsymbol{\zeta}, \boldsymbol{\beta}\,; \boldsymbol{\phi}, \boldsymbol{\lambda})$ where $\boldsymbol{\phi}$ indexes the local variables and $\boldsymbol{\lambda}$ the global ones. This also splits the variational parameters in the evidence lower bound (ELBO) as $\mathcal{L}(\mathbf{X}, \boldsymbol{\phi}, \boldsymbol{\lambda})$.

Hoffman et al. (2013) show that the gradient calculation decomposes into a maximization of the local variables and a gradient with respect to the global variables. The recipe at each iteration is

$$\begin{aligned}
\boldsymbol{\phi}_\dagger &\leftarrow \arg\max_{\boldsymbol{\phi}} \mathcal{L}(\mathbf{X}, \boldsymbol{\phi}, \boldsymbol{\lambda}_{\text{prev}}) \\
\boldsymbol{g}_{\boldsymbol{\lambda}} &\leftarrow \nabla_{\boldsymbol{\lambda}} \mathcal{L}(\mathbf{X}, \boldsymbol{\phi}_\dagger, \boldsymbol{\lambda}) \\
\boldsymbol{\lambda}_{\text{next}} &\leftarrow \boldsymbol{\lambda}_{\text{prev}} + \rho \boldsymbol{g}_{\boldsymbol{\lambda}}.
\end{aligned}$$

where ρ is a scalar step-size.

In SVI, we subsample the dataset $\mathbf{X}$ and accordingly reweight the optimized local variables to construct an unbiased estimate of $\boldsymbol{g}_{\boldsymbol{\lambda}}$. Exponential family models parameterized in their natural forms enjoy a connection to their coordinate ascent updates, but the idea holds in general (Hoffman et al., 2013).

In BUMP-VI, we need B gradients of the ELBO evaluated on the bootstrapped datasets $\{\mathbf{Z}^{(b)}\}_1^B$. Luckily, subsampling is conceptually equivalent to bootstrap resampling: they both induce a weighting scheme on the local latent variables.

We propose the following efficient implementation. At each iteration:

1. Compute the optimized local variables $\boldsymbol{\phi}_\dagger$ *once* for the original dataset.
2. Compute the form of $\nabla_{\boldsymbol{\lambda}} \mathcal{L}$.
3. Generate the B gradients $\{\boldsymbol{g}_{\boldsymbol{\lambda}}^{(b)}\}_1^B$ by re-weighting the local variables according to the bootstrapped datasets $\{\mathbf{Z}^{(b)}\}_1^B$. This means weighting each local variable proportional to the number of times its paired observation appears in the bootstrapped dataset.

We implement this stragey in the accompany code: https://github.com/Blei-Lab/lda-bump-cpp.

References

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347.