Stochastic SuSiE Notes

Karl Tayeb

2022-02-07

The evidence lower bound (ELBO) for SuSiE is given by:

$$\mathcal{L}_{\text{SuSiE}}(q) = \mathbb{E}_q[\log p(y, b, \gamma)] - \mathbb{E}_q[q], \tag{1}$$

We can perform approximate inference by maximizing $\mathcal{L}_{\text{SuSiE}}$ with respect to q in the restricted family of distributions:

$$q^* = \arg\max_{q \in \mathcal{Q}} \mathcal{L}_{\text{SuSiE}}(q).$$

We will restrict our attention to the family that factorizes over the single effects $\mathcal{Q} = \{q : q = \prod q_l\}$. We use q_l as a shorthand for $q(b_l, \gamma_l)$. We will also adopt the notation q_{-l}, b_{-l}, γ_{-l}, etc. to denote the distributions, effects, and variables selected for all but the l-th single effect.

First, let's consider our objective as a function of q_l only

$$\mathcal{L}_{\text{SuSiE}}(q_l; q_{-l}) = \mathbb{E}_{q_{-l}}[\mathbb{E}_{q_l}[\log p(y, b, \gamma)]] - \mathbb{E}_{q_l}[q_l] + C \tag{2}$$

Where C is a constant with respect to q_l. In general, the expectation over $q(\gamma_{-l})$ involves summing over p^L configurations of non-zero effect variables, so we cannot directly compute the ELBO. However, we can obtain a Monte-Carlo estimate of the objective by sampling the outer expectation. Let b^*_{-l}, γ^*_{-l} denote a sample from q_{-l}, then

$$\begin{aligned}
\hat{\mathcal{L}}_{\text{SuSiE}}(q_l) &= \mathbb{E}_{q_l}\left[\log p(y, b_l, \gamma_l, b^*_{-l}, \gamma^*_{-l})\right] - \mathbb{E}_{q_l}[q_l] + C, &(3)\\
&= \mathcal{L}_{\text{SER}}(q_l; \psi^*_{-l}) + C, &(4)
\end{aligned}$$

Where $\psi_{-l} = X(\sum_{k \neq l} b_k \gamma_k)$ are the sampled predictions from the $L-1$ other single effects and $\mathcal{L}_{\text{SER}}(q_l, \psi^*_{-l})$ is the ELBO for an SER, where ψ^*_{-l} is a fixed offset in the linear prediction.

By optimizing $\mathcal{L}_{\text{SER}}(q_l, \psi^*_{-l})$, with respect to q_l, we are also optimizing an unbiased estimate of the SuSiE ELBO. We can use this fact to develop a stochastic approximation approach, where we approximate q_l by solving multiple instances of an easier SER problem.

Connection with gradient descent

In the Gaussian model, the SER prior is conjugate. $q(b|\gamma)$ is Gaussian and $q(\gamma)$ is multinomial. The coordinate updates can be related to a step of gradient descent in the natural parameter space.

Generically, Let the complete conditional $p(b|y, z)$ be an exponential family with natural parameter $\lambda(y, z)$ and $q(b)$ the same family with natural parameter η. Our goal is to optimize $q(b)$, we assume $q(b, z) = q(b)q(z)$.

$$\begin{aligned}
\mathcal{L}(\eta) &= \mathbb{E}_{q(b)}\left[\mathbb{E}_{q(z)}[\log p(b|y, z)]\right] - \mathbb{E}_{q(b)}[\log q(b)] + C && (5)\\
&= \mathbb{E}_{q(b)}\left[\mathbb{E}_{q(z)}[\langle T(b), \lambda(y, z)\rangle]\right] - \mathbb{E}_{q(b)}[\langle T(z), \lambda\rangle - A(\lambda)] + C && (6)\\
&= \mathbb{E}_{q(b)}\left[\mathbb{E}_{q(z)}[\langle T(b), \lambda(y, z) - \eta\rangle]\right] - A(\eta) + C && (7)\\
&= \langle \mathbb{E}_{q(b)}[T(b)], \mathbb{E}_{q(z)}[\lambda] - \eta\rangle - A(\eta) + C && (8)\\
&= \langle \nabla_\eta A(\eta), \mathbb{E}_{q(z)}[\lambda] - \eta\rangle - A(\eta) + C && (9)
\end{aligned}$$

Taking the gradient w.r.t η

$$\nabla_\eta \mathcal{L}(\eta) = \nabla^2_\eta A(\eta)(\mathbb{E}_{q(z)}[\lambda] - \eta)$$

Which is optimized at $\eta = \mathbb{E}_{q(z)}[\lambda]$. This says that the coordinate update for η is the expected natural parameter of the complete conditional.

We can also see that

$$\mathbb{E}_{q(z)}[\lambda] - \eta = (\nabla^2_\eta A(\eta))^{-1}\nabla_\eta \mathcal{L}(\eta)$$

The LHS can be computed by taking the coordinate update and subtracting the current parameter estimate. The RHS is a rescaled gradient.

The idea is that an unbiased estimate of $\mathbb{E}_{q(z)}[\lambda]$ (e.g. obtained from coordinate update of an unbiased estimate of the ELBO), gives a noisy version of the rescaled gradient on the RHS, which can be used in stochastic gradient descent. Call the coordinate update for $\hat{\mathcal{L}}$ $\hat{\eta}$,

$$\begin{aligned}
\eta^{(t+1)} &= \eta^{(t)} + \alpha_t(\nabla^2_\eta A(\eta))^{-1}\nabla_\eta \hat{\mathcal{L}}(\eta) && (10)\\
&= \eta^{(t)} + \alpha_t(\hat{\eta} - \eta^{(t)}) && (11)\\
&= (1 - \alpha_t)\eta^{(t)} + \alpha_t\hat{\eta} && (12)
\end{aligned}$$

We note that $q(\gamma)$ is multinomial (an exponential family, same as the prior) so this stochastic optimization approach should work for $q(\gamma)$. The troulbe we have is that for general likelihoods $p(b|\gamma, y, X)$ is not Gaussian. We could consider making a Gaussian approximation by constraining $q(b|\gamma)$ to be Gaussian. However, while ate each step $\hat{q}(b|\gamma)$ will be the best Gaussian approximation for that iteration– it is not clear that the sequence of $q^{(t)}(b|\gamma)$ would converge to the best Gaussian approximation of the SER posterior. However, if we can access samples from $p(b_{-l}|\gamma_{-l}, y, X)$ we can be sure that $q^{(t)}(\gamma)$ will converge to what we want.

We can consider strategies that help us estimate

$$\mathbb{E}_{q_{-l}}[\lambda(b_l, \gamma_l, b_{-l}, \gamma_{-l})]$$

Variance Reduction

Taking expectation over $q(b_{-l}|\gamma_{-l})$

Where possible, we can reduce variability by taking expectations over $q(b_{-l}|\gamma_{-l})$ analytically, rather than sampling.

$$
\begin{aligned}
\tilde{\mathcal{L}}_{\text{SuSiE}}(q_l) &= \mathbb{E}_{q(b_{-l}|\gamma_{-l})}\left[\mathbb{E}_{q_l}\left[\log p(y, b_l, b_{-l}, \gamma_l, \gamma^*_{-l})\right]\right] - \mathbb{E}_{q_l}[q_l] + C, && (13)\\
&= \mathbb{E}_{q(b_{-l}|\gamma_{-l})}[\mathcal{L}_{\text{SER}}(q_l; \psi_{-l})] + C, && (14)
\end{aligned}
$$

Again, we have $\mathbb{E}\left[\tilde{\mathcal{L}}_{\text{SuSiE}}\right] = \mathcal{L}_{\text{SuSiE}}$. Optimizing $\tilde{\mathcal{L}}_{\text{SuSiE}}(q_l)$ amounts to being fitting an SER with independent normal "random effect" for each observation $(\psi_{-l})_i \sim N(\sum_{k \neq l} x_{i\gamma_k}\mu_{k\gamma_k}, \sum_{k \neq l} x^2_{i\gamma_k}\nu_{k\gamma_k})$. This can be done efficiently e.g. when $\log p(y, \psi)$ is quadratic in ψ, such as when using the Jaakola-Jordan/Polya-Gamma approximation. Note that while the predictions are certainly not independent across samples, the log-likelihood seperates across samples, so only the marginal distribution of $\psi_{-l,i}$ matters here.

Multiple samples

Another easy and more general way to reduce variance is simply to draw more samples. We will consider how to optimize the MC estimate of the $\mathcal{L}_{\text{SuSiE}}$ obtained by averaging multiple draws of $\hat{\mathcal{L}}_{\text{SuSiE}}$ or $\tilde{\mathcal{L}}_{\text{SuSiE}}$.

$$
\hat{\mathcal{L}}_{\text{SuSiE},M} = \frac{1}{M}\sum_m \hat{\mathcal{L}}_{\text{SuSiE}}(q_l)^{(m)} \tilde{\mathcal{L}}_{\text{SuSiE},M} = \frac{1}{M}\sum_m \tilde{\mathcal{L}}_{\text{SuSiE}}(q_l)^{(m)}
$$

In practice this may be an attractive option if we can get a good approximation of the posterior for moderate M. In this case we can closely approximate the (non-stochastic) coordinate wise optimization of the ELBO by fitting the stochastic SER to convergence.

This seems particularly plausible for $\tilde{\mathcal{L}}_{\text{SuSiE},M}$, where as we approach the optimum solution there might not be too much variability across samples. We expect this because the $L - 1$ other single effects should either be concentrated on a few highly correlated variables (giving similar predictions), or not contribute much prediction at all. As a plus, if we can estimate the prior variance for these diffuse components, we will estimate a small prior variance ("automatic relevance determination", "ARD"), which will induce strong shrinkage and further reduce variability in the predictions.

For each sample we can fit an SER to approximate the posterior of q_l for SuSiE. A key question is if we can effectively combine the posterior estimates for multiple SERs.

Typically, stochastic approximation is performed sequentially. At a high level, a stochastic estimate of the loss function or gradient is taken, and used to update q. This step usually depends on the current state of q, and so must be performed sequentially. Over a large number of iterations, we can expect our parameter estimates to drift towards there true value and bounce around there. Stochastic optimization runs the procedure for a long time, slowly ignoring the new gradient information. If this "ignoring" happens at a slow enough rate, famous results due to Robbins and Monroe guarantee the convergence (in probability?) of our parameter estimates to their true optimum.

In our case, the stochasticity we are using does not depend at all on q_l. Intuitively, we should hope the way we combine the SER posteriors to treat each posterior symmetrically, or at least in a way that does not depend on the ordering of the posterior SERs. Below we provide a sketch of an argument for why averaging the natural parameters is reasonable. Importantly, it seems to work in practice.

Combining SER posteriors

Suppose $\eta^{(m)}$ optimizes $\mathcal{L}^{(m)}_{\text{SER}}$ (i.e. $\nabla_\eta \mathcal{L}^{(m)}_{\text{SER}}(\eta^{(m)}) = 0$). Then, we can approximated the gradient of $\tilde{\mathcal{L}}_{\text{SuSiE},M}$ by taking a 1st order Taylor series expansion around $\eta^{(m)}$ for each $\nabla_\eta \mathcal{L}^{(m)}_{\text{SER}}$

$$\nabla_\eta \tilde{\mathcal{L}}_{\text{SuSiE},M}(\eta) = \frac{1}{M}\sum_m \nabla_\eta \tilde{\mathcal{L}}^{(m)}_{\text{SuSiE}}(\eta) \tag{15}$$

$$= \frac{1}{M}\sum_m \nabla_\eta \mathcal{L}^{(m)}_{\text{SER}}(\eta) \tag{16}$$

$$\approx \frac{1}{M}\sum_m \nabla_\eta \mathcal{L}^{(m)}_{\text{SER}}(\eta^{(m)}) + \nabla^2_\eta \mathcal{L}^{(m)}_{\text{SER}}(\eta^{(m)})(\eta - \eta^{(m)}) \tag{17}$$

$$= \frac{1}{M}\sum_m \nabla^2_\eta \mathcal{L}^{(m)}_{\text{SER}}(\eta^{(m)})(\eta - \eta^{(m)}) \tag{18}$$

Where we take note that $\nabla_\eta \mathcal{L}^{(m)}_{\text{SER}}(\eta^{(m)}) = 0$. In the case where we have a Gaussian likelihood, the SER prior is conjugate, and $\nabla_\eta \mathcal{L}_{\text{SER}}(\eta^{(m)}) = -\nabla^2_\eta A(\eta^{(m)})$.

$$\frac{1}{M}\sum_m \nabla^2_\eta \mathcal{L}^{(m)}_{\text{SER}}(\eta^{(m)})(\eta - \eta^{(m)}) = 0 \implies \eta = \left(\sum \nabla^2_\eta \mathcal{L}^{(m)}_{\text{SER}}(\eta^{(m)})\right)^{-1} \left(\nabla^2_\eta \mathcal{L}^{(m)}_{\text{SER}}(\eta^{(m)})\eta^{(m)}\right) \tag{19}$$

However, $q(b|\gamma)$ is univariate Normal, the Fisher information matrix is diagonal, and only depends on the posterior variance. Assuming a normalized X and a fixed residual variance, this corresponds with a straight average of the natural parameters across M SERs.

For $q(\gamma)$, we note that the natural parameters $\{\eta_i : i \in [p-1]\}$ are the log Bayes factors comparing selecting variable i over selecting variable p. This can be written in terms of the log Bayes factors against the null

$$\eta_i = \text{lbf}_i - \text{lbf}_p$$

Then

$$\bar{\eta}_i = \overline{\text{lbf}}_i - \overline{\text{lbf}}_p$$

where $\overline{\text{lbf}}_i = \frac{1}{M}\sum \text{lbf}^{(m)}_i$. Recognizing that $\text{lbf}^{(m)}_i$ is an unbiased estimate of lbf_i in the full SuSiE model, it also seems reasonable to to take the average here. This is different from the 19. In preliminary results, averaging seems to work better, the information matrix can be poorly conditioned, and seems to perform worse empirically.

You might also think it's reasonable to say $\nabla^2 \mathcal{L}^{(m)}_{\text{SER}} \approx K$ for all m. The curvature of the objective function at it's optimum is dictated mostly by the information in the data. The information content in each MC sample differs slightly due to sampling q_{-l}, but should be simlar/the same in some averaged sense. For a constant K, it follows that $\bar{\eta} = \frac{1}{M}\sum_m \eta^{(m)}$ is close the the optimum of $\tilde{\mathcal{L}}_{\text{SuSiE},M}$.

Justification for averaging the natural paremeters

We can approximate the posterior inference, $q(b_l, \gamma_l) \propto \exp\{\log p(y|b_l, \gamma_l)\}\, p(b_l, \gamma_l)$ by plugging in our MC estimate for $\log p(y|b_l, \gamma_l)$

$$\hat{q}(b_l, \gamma_l) \propto \exp\left\{\frac{1}{M}\sum_m \log p(y|b_l, \gamma_l, b^{(m)}_{-l}, \gamma^{(m)}_{-l})\right\} p_l(b_l, \gamma_l) \tag{20}$$

$$= \prod_m \left\{p(y|b_l, \gamma_l, b^{(m)}_{-l}, \gamma^{(m)}_{-l}) p_l(b_l, \gamma_l)\right\}^{\frac{1}{M}} \tag{21}$$

$$\tag{22}$$

For exponential families, conjugacy reduces posterior computation to a sum,

$$p(y|z)p(z) \propto \exp\left[\langle T(z), \eta(y)\rangle\right] \exp\left[\langle T(z), \eta_0\rangle\right] = \exp\left[\langle T(z), \eta(y) + \eta_0\rangle\right]$$

In the conjugate case, where $\eta^{(m)}$ is the natural parameter for $q_l^{(m)} \propto p(y, b_l, \gamma_l, b_{-l}^{(m)}, \gamma_{-l}^{(m)})$, then $\bar{\eta} = \frac{1}{M}\sum \eta^{(m)}$ is the natural parameter for $\hat{q}_l$. To see this note that we can write $\eta^{(m)} = \tilde{\eta}^{(m)} + \eta_0$. Then we can see that the natural parameter for $\hat{q}_l$ can be written,

$$\eta_0 + \frac{1}{M}\sum \tilde{\eta}^{(m)} = \frac{1}{M}\sum \eta^{(m)} = \bar{\eta}$$

So actually, averaging the natural parameters can be justified for large M. We can think of this as normal Bayesian inference incorporating a collection of "partial observations", one from each MC sample with "weight" $1/M$.

In the non-conjugate case, we form an approximation for each MC sample, $q_l^{(m)} \approx \hat{q}_l^{(m)}$ with natural parameter $\eta^{(m)}$. Then, we think, that $\bar{\eta}$ is close to the natural parameter for $\hat{q}_l$. But now we also need to find a way to argue that the average of the natural parameters of these "projected posteriors" (projected to the prior family) is close to the "projected posterior" formed by projecting the "exact" posterior across all M samples.

Obviously, while $\mathbb{E}\left[\frac{1}{M}\sum_m \log p(y|b_l, \gamma_l, b_{-l}^{(m)}, \gamma_{-l}^{(m)})\right] = \log p(y|b_l, \gamma_l)$,

$$\mathbb{E}\left[\exp\left\{\frac{1}{M}\sum_m \log p(y|b_l, \gamma_l, b_{-l}^{(m)}, \gamma_{-l}^{(m)})\right\}\right] \neq p(y|b_l, \gamma_l),$$

But the bias term shrinks as $M \to \infty$, and variability in the average disappears. So, I want to say $\hat{q}_l \to q_l$ in some sense.

Conjugacy: There is only one family of distributions on Categorical data (the Categorical distribution). $q(\gamma_l)$ is Categorical, so is $p(\gamma_l)$.

We may constrain $q(b|\gamma)$ to be Normal, in which case we need to account for the approximation error (more analysis needed). Otherwise we can find some other means to sample from $q(b|\gamma)$. Can we approximate with the mixture $q(b|\gamma) \sim \sum \frac{1}{M} q^{(m)}(b|\gamma)$? We can make a normal approximation for each $q^{(m)}$

Stochastic SuSiE is highly paralellizable

Compared to regularly SuSiE, stochastic SuSiE will require many more calls to the SER routine. However, computing the posterior in an SER is embarassingly parallelizable– as it basically involve fitting p independent regression problems to estimate $q(b|\gamma)$, and normalizing the Bayes factors to compute $q(\gamma)$. Furthermore, we can fit SERs for all M samples in parallel.

Algorithm

Algorithm 1 Stochastic SuSiE

```
Require: Optimization schedule $(\alpha_t)_{t=1}^{\infty}$ $s.t. \sum \alpha_t = \infty, \sum \alpha_t^2 < \infty$,
Require: Initialzation $q_1^{(0)}, \dots q_L^{(0)}$
  repeat
      $t \leftarrow 0$
      for $l = 1, \dots, L$ do
          for $m = 1, \dots, M$ do
              $b_{-l}^{(m)}, \gamma_{-l}^{(m)} \sim q_{-l}^{(t)}$        ▷ Sample other effects
              $\psi_{-l}^{(m)} \leftarrow X \sum_{k \neq l} \gamma_k^{(m)} b_k^{(m)}$        ▷ Linear prediction
              $\hat{q}_l^{(m)} \leftarrow \text{SER}(y, X, \psi_{-l}, \pi, \sigma_0^2)$        ▷ Fit SER with fixed offset
          end for
          $\hat{q}_l \leftarrow \frac{1}{M} \sum \hat{q}_l^{(m)}$        ▷ Using summation to denote "combining" approximate posterior distributions
          $q_l^{(t+1)} \leftarrow (1 - \alpha_t) q_l^{(t)} + \alpha_t \hat{q}_l$
      end for
      $t \leftarrow t + 1$
  until forever
```