Bayesian Inference for the Hyperparameters of Generalised Bayesian Inference

J.E. Lee, S. Liu and G.K. Nicholls

"One does not simply walk into Mordor"

MaxEnt 25 - Auckland, NZ 18/12/25

Bayesian inference as a belief update

Generative model: $\theta \sim \pi(\cdot)$, $\theta \in \mathbb{R}^p$ and $\mathbf{x} \sim p(\cdot|\theta)$, $\mathbf{x} \in \mathbb{R}^n$

Bayes Posterior: $\theta \sim \pi(\cdot|\mathbf{x})$

$$\pi(\cdot|\mathbf{x}) \propto \pi(\theta)\, p(\mathbf{x}|\theta).$$

Belief update[1] ψ maps prior π and data to distribution $\nu(\theta)$,

$$\psi(\theta; \pi, \mathbf{x}) = \nu(\theta).$$

The Bayesian belief update

$$\psi_{\text{Bayes}}(\theta; \pi, \mathbf{x}) = \pi(\theta|\mathbf{x})$$

is just one of many.

Can regard the choice of ψ as part of the overall inference, like the statistical modeling we use to elicit the prior and likelihood.

Loss and Gibbs posterior

Loss to data: $\ell(\theta; \mathbf{x}) \in \mathbb{R}$. For eg $\ell_{\text{Bayes}}(\theta; \mathbf{x}) = -\log(p(\mathbf{x}|\theta))$.

Choose ψ as the BU minimising[1]

$$\mathcal{L}(\nu) = \eta\, E_{\theta\sim\nu}[\,\ell(\theta;\mathbf{x})\,] \;+\; D_{\text{KL}}(\nu\,||\,\pi)$$

for $\eta \geq 0$ fixed. If $\nu^* = \arg\min_\nu \mathcal{L}(\nu)$ then

$$\nu^*(\theta) \propto \pi(\theta)\, \exp(-\eta\, \ell(\theta;\mathbf{x})), \tag{GP}$$

the Gibbs posterior. Bayes takes $\ell = \ell_{\text{Bayes}}$ and $\eta = 1$.

Example[2,5]: $S = (S_1, \dots, S_K)$ is a partition of $[n] = \{1, \dots, n\}$,

$$\ell(S;\mathbf{x}) = \sum_{k=1}^{K} \sum_{i \in S_k} (x_i - \bar{x}_k)^2 \qquad (k\text{-means loss})$$

$\pi_\eta(S|\mathbf{x}) \propto \pi(S)\, \exp(-\eta\, \ell(S;\mathbf{x}))$ is a Gibbs posterior for clustering.

In this example there is no generative model, no $p(\mathbf{x}|S)$.

Parameterising the loss

Keep $p(\mathbf{x}|\theta)$, modify BU. True generative model $X \sim p^*(\cdot)$. Risk,

$$R(\theta) = D_f(p^*(X)||p(X|\theta))$$

based on Bregman-divergence D_f,

$$D_f(p^*||p) = \int f(p^*(x))dx - \int f(p(x|\theta))dx - \int f'(p(x|\theta))(p^*(x) - p(x|\theta))dx,$$

Estimate D_f (up to constant) using data

$$\ell_f(\theta; \mathbf{x}) = n\mathbb{E}_{X|\theta}(f'(p(X|\theta))) - n\int f(p(x|\theta))dx - \sum_{i=1}^{n} f'(p(x_i|\theta))$$

If $f(x) = x\log(x) - x$ then $\ell_f = \ell_{\text{Bayes}}$ and

$$\pi_\eta(\theta|\mathbf{x}) \propto \pi(\theta)\, p(\mathbf{x}|\theta)^\eta \qquad \text{(*power posterior*)}$$

If $f(x; \beta) = (x^\beta - 1)/\beta(\beta - 1)$ then

$$\ell_\beta(\theta; \mathbf{x}) = -\frac{1}{\beta - 1}\sum_{i=1}^{n} p(x_i|\theta)^{\beta-1} + \frac{n}{\beta}\int p(x|\theta)^\beta dx \qquad (\beta\text{-loss}^{3,4})$$

$$\pi_{\eta,\beta}(\theta|\mathbf{x}) \propto \pi(\theta)\exp(-\eta\,\ell_\beta(\theta; \mathbf{x})) \qquad (\eta, \beta\text{-}\textit{posterior})$$

Recover power posterior as $\beta \to 1$.

Choosing loss hyperparameters I - estimate $s = (\eta, \beta)$.

Don't:

take a prior $\rho(s)$ and use Bayesian inference

$$\rho(\theta, s|\mathbf{x}) \propto \rho(s)\pi(\theta)\frac{\exp(-\eta\,\ell_\beta(\theta;\mathbf{x}))}{c(\theta, s)}.$$

$c(\theta, s)$ needed to normalise "likelihood", $\exp(-\eta\,\ell_\beta)$ over $\mathbf{x}$.

$\Rightarrow c(\theta, s)$ messes up θ dependence, doesn't give (η, β)-posterior.

Do:

Consider a *block* of test data $z \sim p^*$, $z = (z_1, \ldots, z_m)$.

Suppose goal is to predict z using posterior predictive

$$p_s(z|\mathbf{x}) = \int p(z|\theta)\,\pi_s(\theta|\mathbf{x})\,d\theta.$$

Risk for prediction is

$$\tilde{l}(s;\mathbf{x}) = E_{z\sim p^*}[-\log(p_s(z|\mathbf{x}))]$$

so best s is $s^* = \arg\min_s \tilde{l}(s^*;\mathbf{x})$.

Plan: work with held-out data and empirical risk.

Choosing loss hyperparameters II - what we do

1) From $\mathbf{x}$ hold out J blocks of m calibration samples,

$$y_{(J,m)} = (y_{(1)}, \ldots, y_{(J)}) \text{ with } y_{(j)} = (y_{(j-1)m+1}, \ldots, y_{jm}).$$

2) Define empirical rick/loss $l(s; y_{(J,m)}, \mathbf{x})$ for s-estimation,

$$l(s; y_{(J,m)}, \mathbf{x}) = -\sum_{j=1}^{J} \log(p_s(y_{(j)}|\mathbf{x})). \qquad \text{(ER/LTI)}$$

3) Update belief for $s = (\eta, \beta)$ using Gibbs posterior

$$\rho(s|y_{(J,m)}; \mathbf{x}) \propto \rho(s) \prod_{j=1}^{J} p_s(y_{(j)}|\mathbf{x}).$$

Let $\widehat{s}_{(J,m)} = \arg\min_s l(s; y_{(J,m)}, x)$ minimise empirical risk.

Remarks:

this is just Bayesian inference with a log-likelihood $-l(s; y_{(J,m)}, \mathbf{x})$;

here $\exp(-l(s; y_{(J,m)}, \mathbf{x}))$ is a normalised PDF.

the "true" parameter we want to estimate is s^*;

this BU is well specified as $\widehat{s}_{(J,m)} \to s^*$ as $J \to \infty$.

Choosing loss hyperparameters III - properties

Theorem: under regularity conditions, if $J = 1$ then

$$\rho(s|y_{(J,m)};\mathbf{x}) \xrightarrow{p} \rho(s)\pi_s(\widehat{\theta}_m|\mathbf{x})/c$$

as $m \to \infty$, where $\widehat{\theta}_m = \arg\max_\theta p(y_{(1,m)}|\theta)$ is the MLE.

If $m \geq 1$ then as $J \to \infty$ we have $\widehat{s}_{(J,m)} \to s^*$ and

$$\sqrt{J}(s - \widehat{s}_{(J,m)}) \xrightarrow{t.v.} N(0, H^{-1})$$

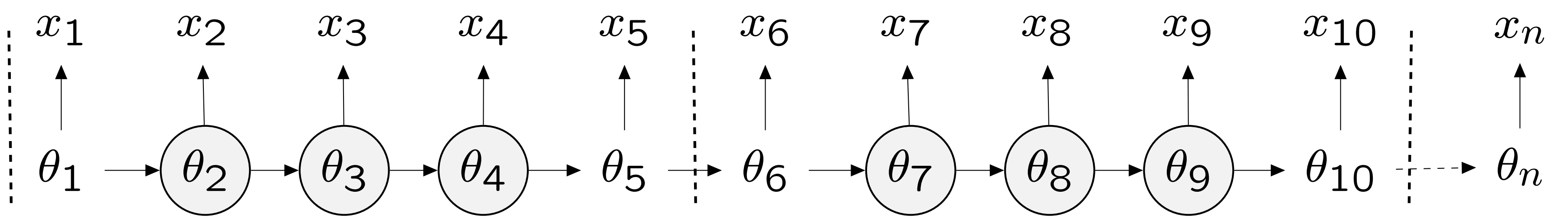
for $s \sim \rho(s|y_{(J,m)};\mathbf{x})$ with $H = \nabla_s^2 \tilde{l}(s^*;x)$ the Hessian.

Remarks:

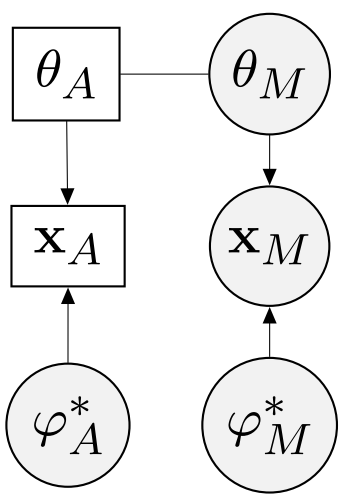
the second part is a classical BvM result from an additive log-lkd

first part has convergence to diffuse distribution if no blocking

Example: State Space Model



Latent process $\theta = (\theta_1, \ldots, \theta_n)$ and data $x = (x_1, \ldots, x_n)$.
Block size $m = 5$ with $n = J^{(x)}m$ if there are $J^{(x)}$ blocks.



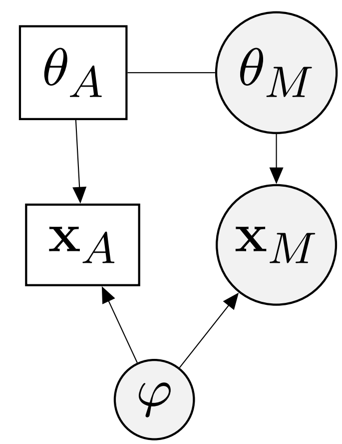
True model

$$\theta_i \sim N(\nu\theta_{i-1}, \sigma^2), \;\; i \in 2:n$$

$$x_{A,i} \sim N(\theta_{A,i}, (\varphi_A^*)^2), \;\; i \in 1:p_A$$

$$x_{M,i} \sim N(\theta_{M,i}, (\varphi_M^*)^2), \;\; i \in 1:p_M$$

$$\varphi_A^* = 1$$



Fitted model

SSM risk ratios

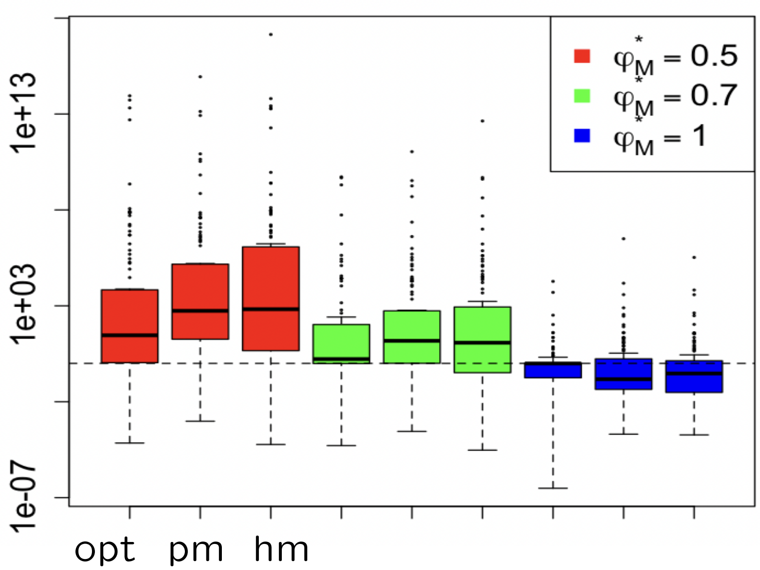
$$\pi_{\eta,\beta}(\theta,\varphi|\mathbf{x}) \propto \pi(\theta,\varphi)\ \exp(-\eta\,\ell_\beta(\theta,\varphi;\mathbf{x})) \qquad (\eta,\beta\textit{-posterior})$$

Fix $\beta = 1$ (gives power posterior). GBI for η:

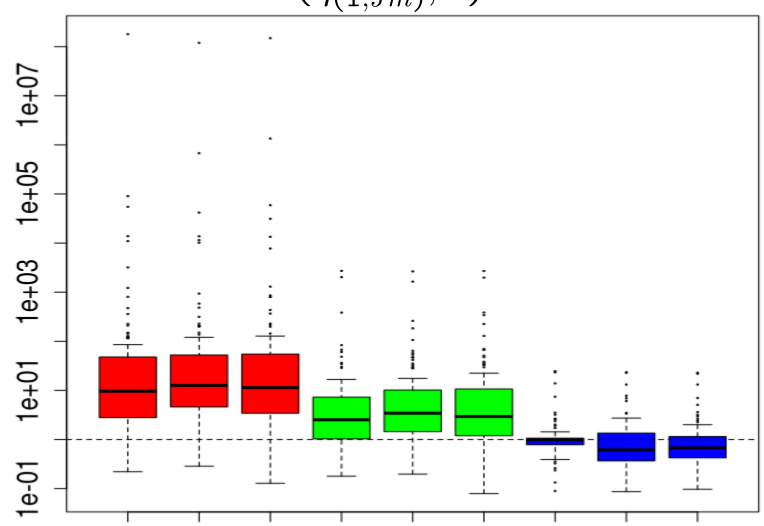
$$\rho(\eta|\mathbf{y};\mathbf{x}) \propto \rho(\eta) \textstyle\prod_{j=1}^{J} p_\eta(y_{(j)}|\mathbf{x}).$$

Compare against Bayes.

$R(\eta_{(J,m)}, 1)$

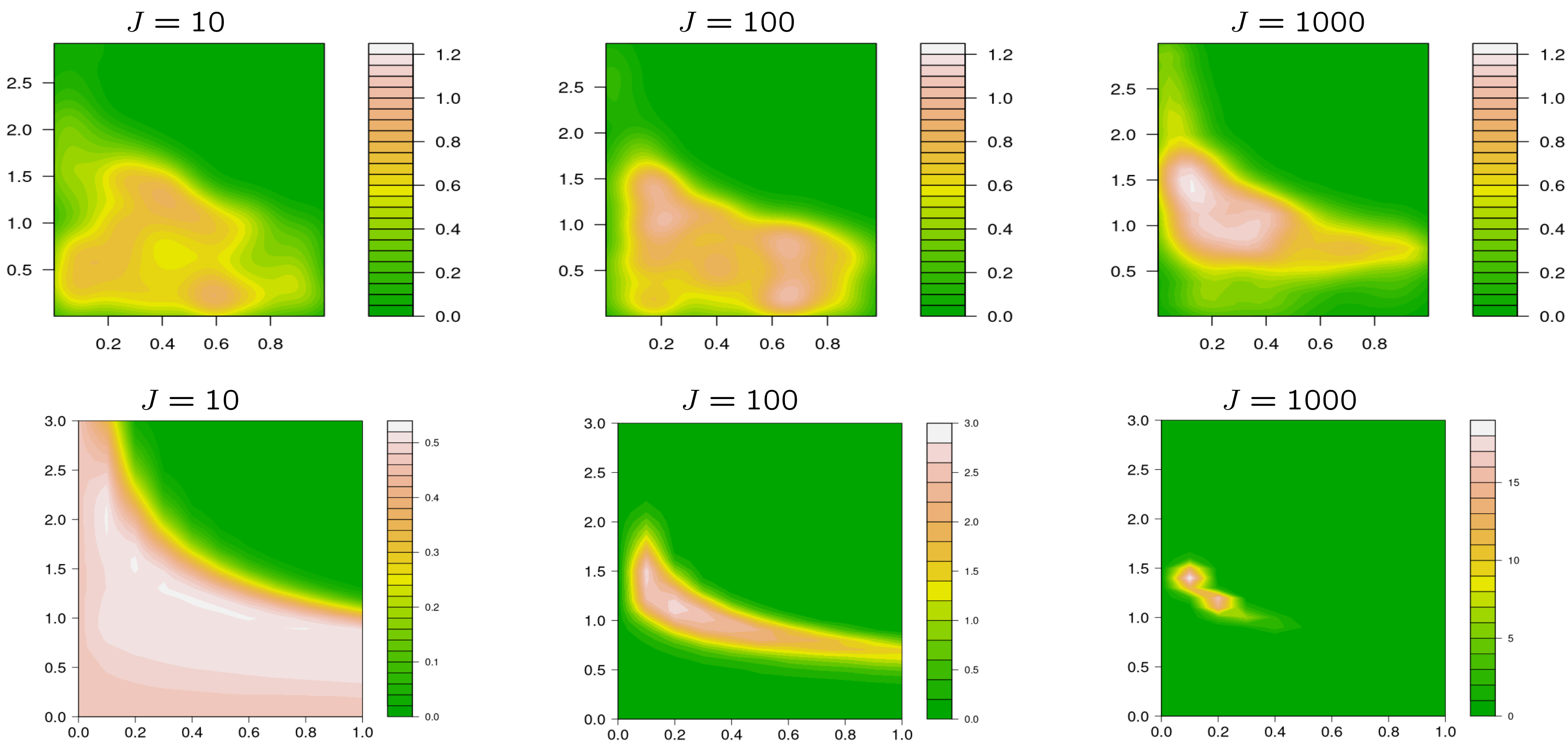


$R(\eta_{(1,Jm)}, 1)$



$$R(\widehat{s}_1,\widehat{s}_2) = \mathbb{E}_{\{z_{j,1:m}\}_{j=1}^{J^{(z)}} \sim p^*}\left[\frac{p_{\widehat{s}_1}(\{z_{j,1:m}\}_{j=1}^{J^{(z)}}|\mathbf{y}_{(J,m)},\mathbf{x})}{p_{\widehat{s}_2}(\{z_{j,1:m}\}_{j=1}^{J^{(z)}}|\mathbf{y}_{(J,m)},\mathbf{x})}\right].$$

(η, β)-posterior asymptotics at fixed $\mathbf{x}$ with $m, J^{(y)} \to \infty$



(η, β)-posterior for $\varphi^*_M = 0.7$; η on x-axis, $1/\beta$ on y-axis. Rows show posteriors for pooled/$J^{(y)} = 1$, $J = mJ^{(y)}$ calibration data (top row) and blocked/$J = J^{(y)}$ calibration data (bottom row). Fixed training data size, $J^{(x)} = 10$ and $m = 5$ throughout.

Unsupervised sense-clustering of text snippets[6,7]

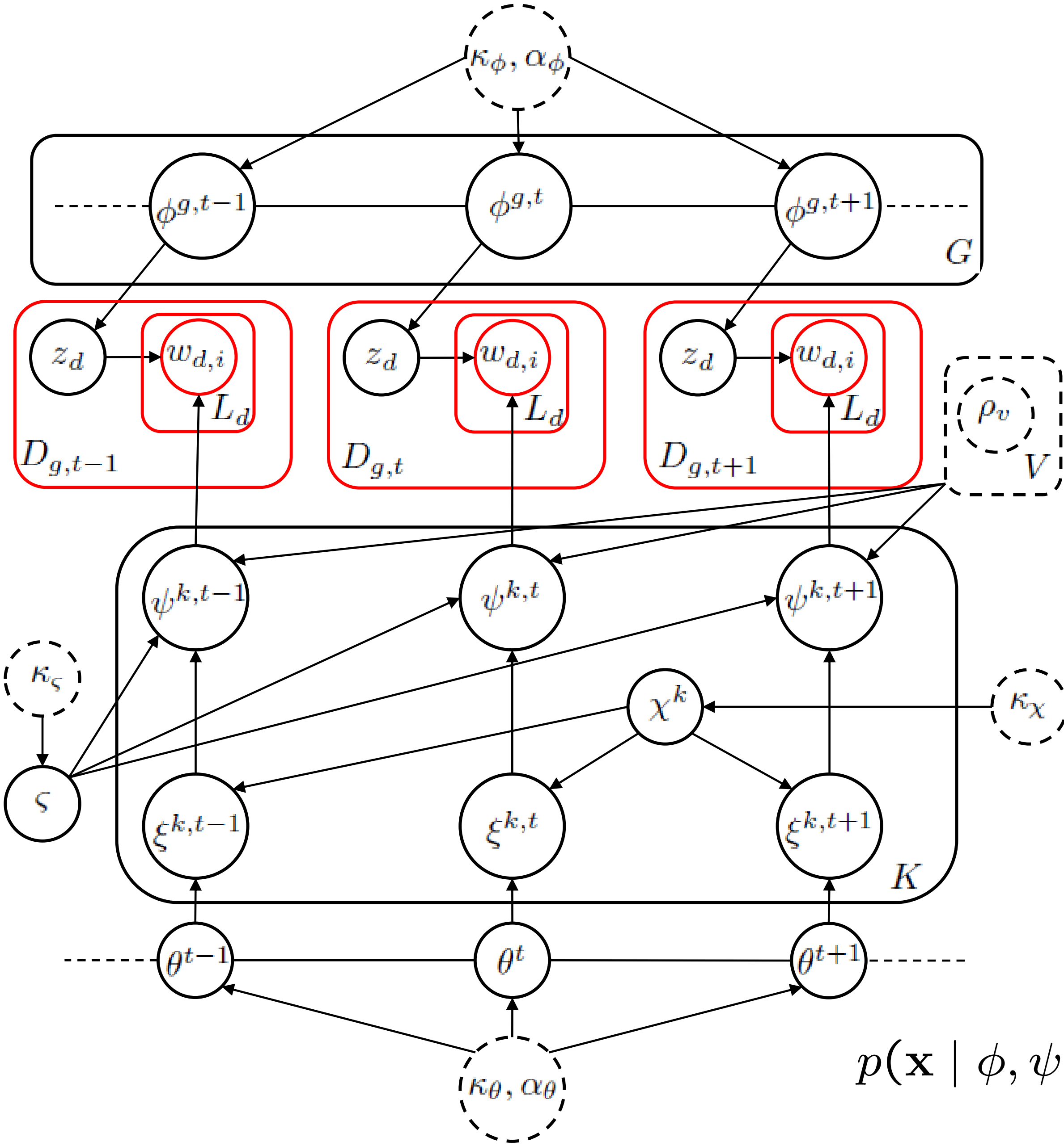
"...were sitting on the grass. A small bug landed on the picnic blanket and crawled..."

'...warranted further investigation. Federal agents planted a bug in the suspect's office to gather intelligence..."

"...released a patch to fix a major bug that was causing the application to crash..."

"...out I had finally caught the stomach bug that had been going around the office..."

Target word	Snippets (N)	Vocab (V)	Length (L)	True senses (K^*)	Model senses (K)	Genres (G)	Train, Cal, block ($n, \|\mathbf{y}\|, J$)	Time periods (T)	detail
bank split 1	704	736	14	2	2	1	500, 204, 34	10	1810–2010
bank split 2	708	717	14	2	2	1	500, 208, 34	10	1810–2010
bank split 3	703	728	14	2	2	1	500, 203, 34	10	1810–2010
bank split 4	704	742	14	2	2	1	500, 204, 34	10	1810–2010
bank split 5	706	735	14	2	2	1	500, 206, 34	10	1810–2010
chair	745	3,180	20	2	2	4	500, 245, 41	10	1820–2020
apple	1,154	3,737	20	2	2	4	800, 354, 59	5	1960–2020
gay	650	3,071	20	2	4	3	450, 200, 33	5	1920–2020
mouse	584	2,439	20	2	3	3	400, 184, 31	4	1940–2020
bug	522	2,475	20	4	4	3	400, 122, 20	8	1980–2020



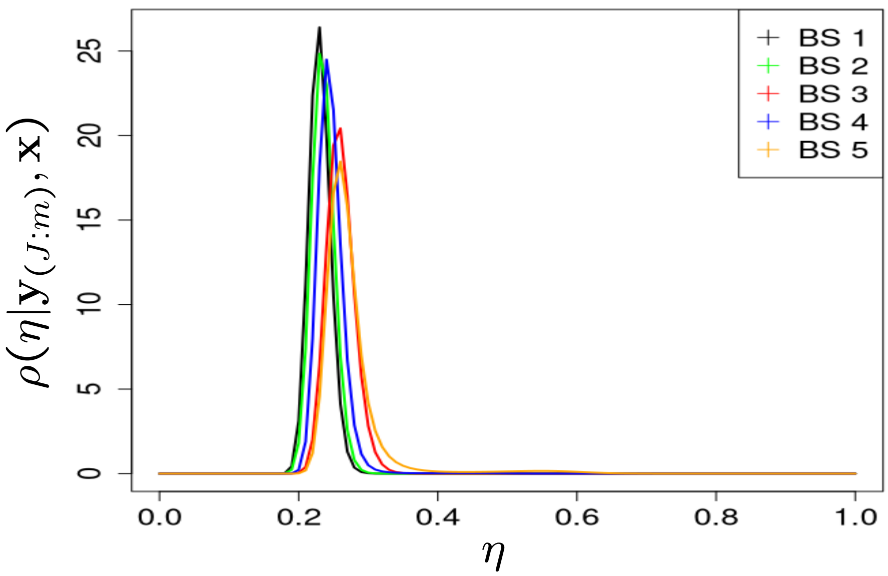
EDiSC $t \in \{1, \ldots, T\}$.

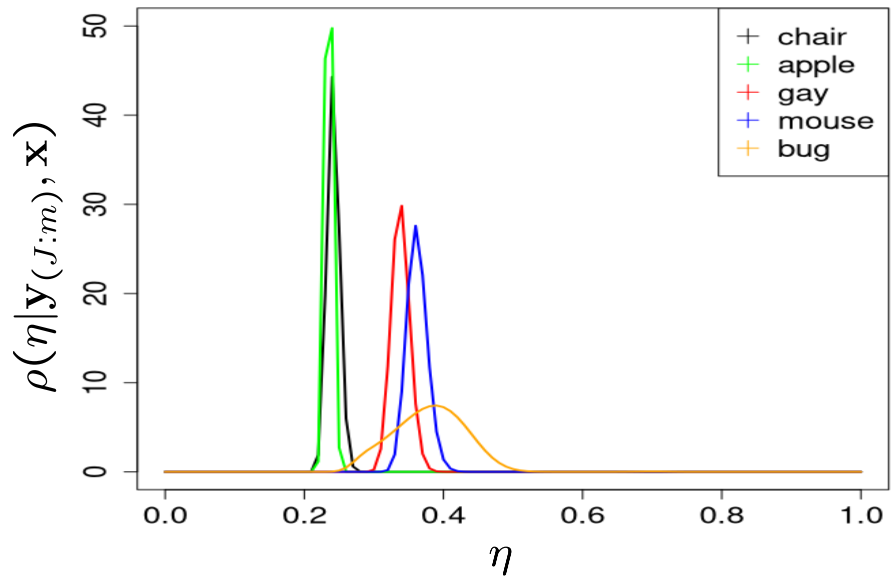
Snippets $\mathbf{x}$ are data, sense assignments $z = (z_1, \ldots, z_n)$, $z_i \in [K]$.

Dashed nodes are constant, solid black are latent variables, solid red are observed.

$$p(\mathbf{x} \mid \phi, \psi) = \prod_{d=1}^{n} \sum_{k=1}^{K} \tilde{\phi}_k^{\gamma_d, \tau_d} \prod_{w \in x_d} \tilde{\psi}_w^{k, \tau_d}$$

$$\pi_\eta(\phi, \psi \mid \mathbf{x}) \propto \pi(\phi, \psi) p(\mathbf{x} \mid \phi, \psi)^\eta.$$





Sense	**Top 9 context words** $\eta = 1$							
1	say	p	year	computer	get	new	make	one	company
2	system	fix	computer	update	new	use	device	company	security
3	insect	spray	bug	find	mosquito	eat	assassin	little	beetle
4	p	cause	bacterium	new	plant	also	make	people	find

Sense	**Top 9 context words** $\eta = \bar{\eta} = 0.4$							
1	p	computer	new	say	year	company	software	make	get
2	say	new	federal	security	agent	phone	office	system	p
3	insect	bug	spray	mosquito	find	beetle	say	like	little
4	p	cause	make	bacterium	say	virus	get	people	one

Sense	**Top 9 context words** $\eta = 0.2$							
1	p	say	new	computer	get	make	find	year	one
2	p	say	new	computer	make	get	year	find	one
3	p	say	make	new	get	insect	find	one	use
4	say	p	bug	insect	get	make	spray	find	like

Conclusions

Generalising Bayesian inference gives another degree of freedom for "modeling".

Comes with additional burden of (abstract) statistical modeling - model the inference - choose loss and loss hyperparameters.

*References

[1] P. G. Bissiri, C. C. Holmes, and S. G. Walker. "A general framework for updating belief distributions". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 78.5 (Nov. 2016), pp. 1103–1130. ISSN: 13697412. DOI: `10.1111/rssb.12158`.

[2] Emilie Eliseussen, Arnoldo Frigessi, and Valeria Vitelli. *Rank-based Bayesian clustering via covariate-informed Mallows mixtures*. 2024. arXiv: `2312.12966 [stat.ME]`. URL: `https://arxiv.org/abs/2312.12966`.

[3] Jack Jewson, Jim Q. Smith, and Chris Holmes. "On the Stability of General Bayesian Inference". In: *Bayesian Analysis* (2024), pp. 1–31. DOI: `10.1214/24-BA1502`. URL: `https://doi.org/10.1214/24-BA1502`.

[4] Jeremias Knoblauch, Jack E Jewson, and Theodoros Damoulas. "Doubly Robust Bayesian Inference for Non-Stationary Streaming Data with β-Divergences". In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio et al. Vol. 31. Curran Associates, Inc., 2018.

[5] Tommaso Rigon, Amy H Herring, and David B Dunson. "A generalized Bayes framework for probabilistic clustering". In: *Biometrika* 110.3 (Jan. 2023), pp. 559–578. ISSN: 1464-3510. DOI: `10.1093/biomet/asad004`. eprint: `https://academic.oup.com/biomet/article-pdf/110/3/559/51111533/asad004.pdf`. URL: `https://doi.org/10.1093/biomet/asad004`.

[6] Schyan Zafar and Geoff K Nicholls. "An embedded diachronic sense change model with a case study from ancient Greek". In: *Computational Statistics & Data Analysis* 199 (2024), p. 108011.

[7] Schyan Zafar and Geoff K Nicholls. "Exploring Learning Rate Selection in Generalised Bayesian Inference using Posterior Predictive Checks". In: *arXiv preprint arXiv:2410.01475* (2024).