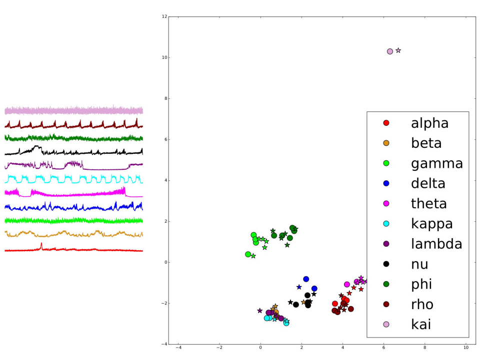
Nikos Gianniotis, Postdoc

Astroinformatics Group @
Heidelberg Institute for Theoretical Studies (HITS)

nikos.gianniotis@h-its.org

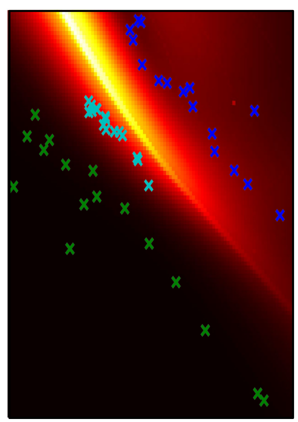
I am a computer scientist working in machine learning

Dimensionality reduction for time series

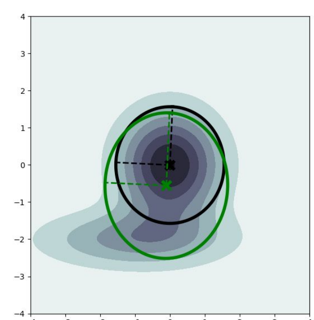


Xray series from GRS1915+105

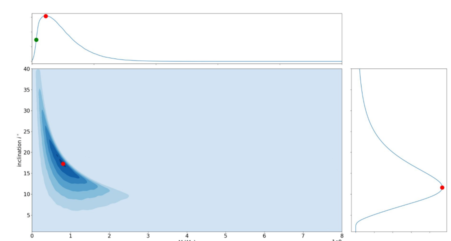
Metric properties of eclipsing binary systems



Approximate Bayesian inference



Reverberation Mapping



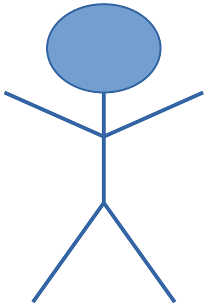
Heidelberg Institute for Theoretical Studies

HITS

Mixed Variational Inference

Rev. Thomas Bayes
1702 - 1761

$\approx$



Nikos Gianniotis

Astroinformatics Group

Acknowledgements: Christoph Schnoerr - University of Heidelberg
Christian Molkenthin - University of Potsdam

Overview

- Start by discussing why Bayesian inference is important
- Bayesian inference is difficult to carry it out in many cases
- The talk is about overcoming these computational difficulties
- Two elements:
  - How to overcome the difficulty *(spoiler: monte carlo average)*
  - How to do it efficiently *(spoiler: employ Laplace approximation)*

Bayesian Inference

- Bayesian inference is the consistent use of probabilities in reasoning
  This means we have to play by certain rules

$$p(w|\mathcal{D}) = \frac{p(\mathcal{D}|w)p(w)}{p(\mathcal{D})}$$

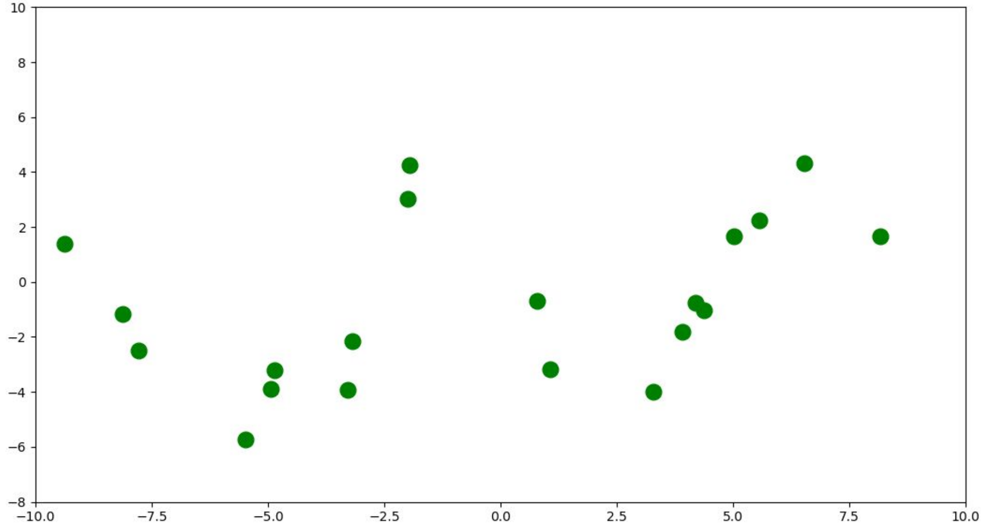
- $\mathcal{D}$ are the data, and $w \in \mathbb{R}^d$ are the model parameters

- Bayes tells us how to work out the set of all likely solutions w

Why Bayesian Inference?

Quantify uncertainty

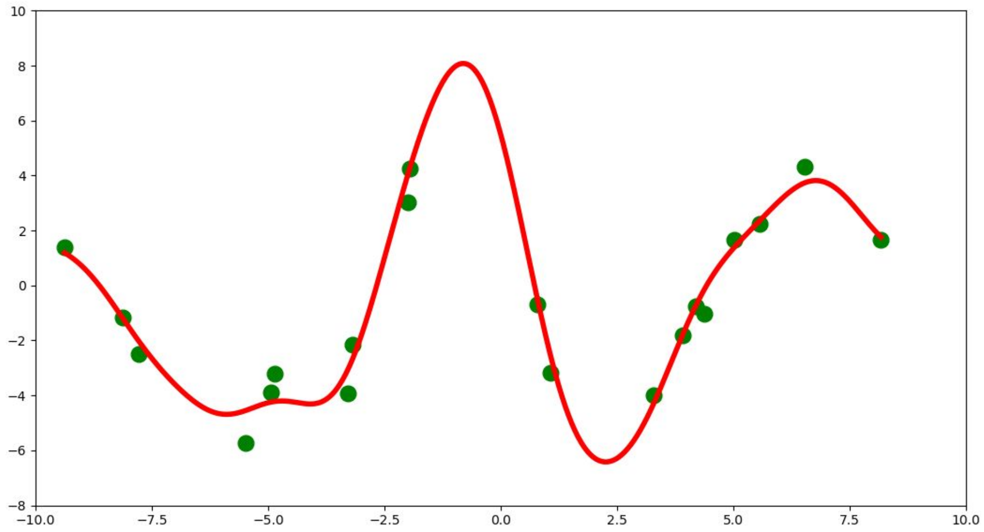
1. We learn a regression model to model dependency between x and y
2. The model possesses parameters w
3. However, there are many parameters w that are likely



Why Bayesian Inference?

Quantify uncertainty

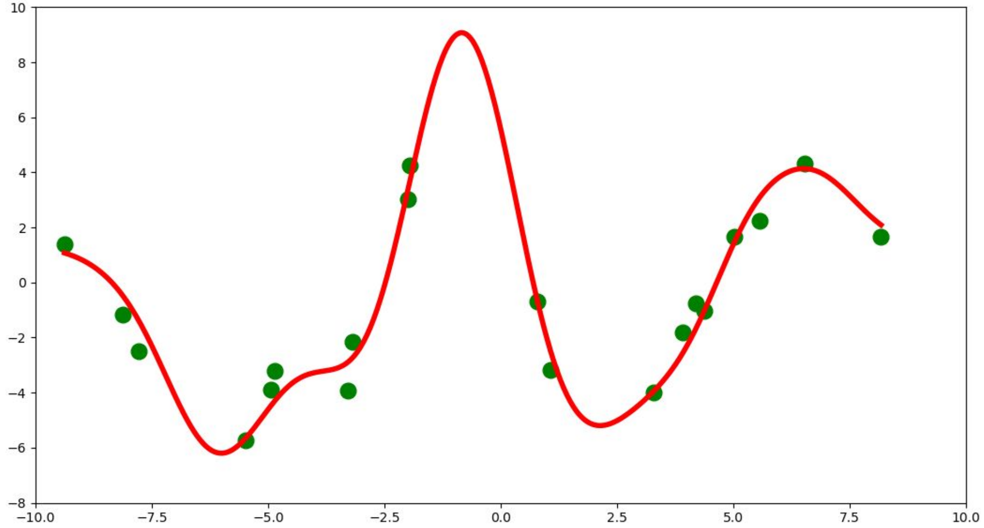
1. We learn a regression model to model dependency between x and y
2. The model possesses parameters w
3. However, there are many parameters w that are likely



Why Bayesian Inference?

Quantify uncertainty

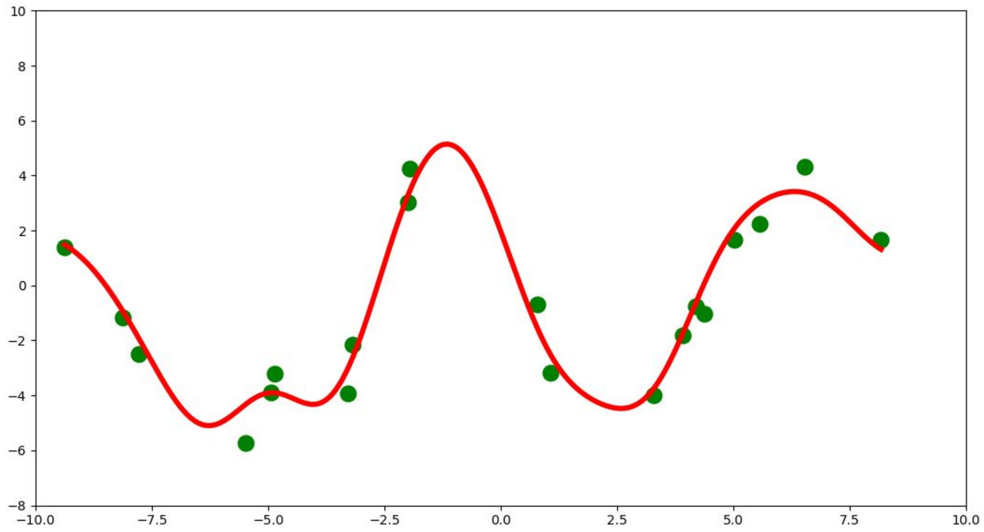
1. We learn a regression model to model dependency between x and y
2. The model possesses parameters w
3. However, there are many parameters w that are likely



Why Bayesian Inference?

Quantify uncertainty

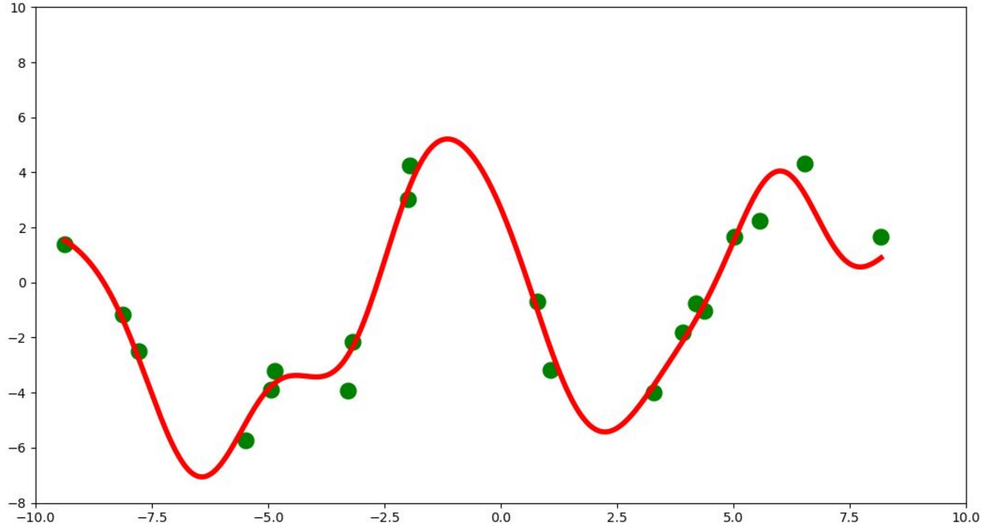
1. We learn a regression model to model dependency between x and y
2. The model possesses parameters w
3. However, there are many parameters w that are likely



Why Bayesian Inference?

Quantify uncertainty

1. We learn a regression model to model dependency between x and y
2. The model possesses parameters w
3. However, there are many parameters w that are likely

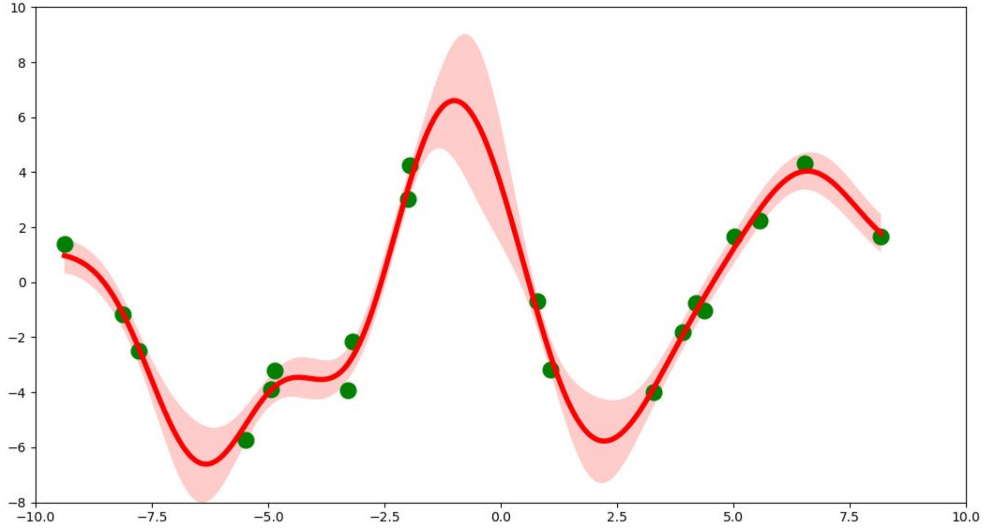


Why Bayesian Inference?

Quantify uncertainty

1. We learn a regression model to model dependency between x and y
2. The model possesses parameters w
3. However, there are many parameters w that are likely

$$p(w|\mathcal{D})$$

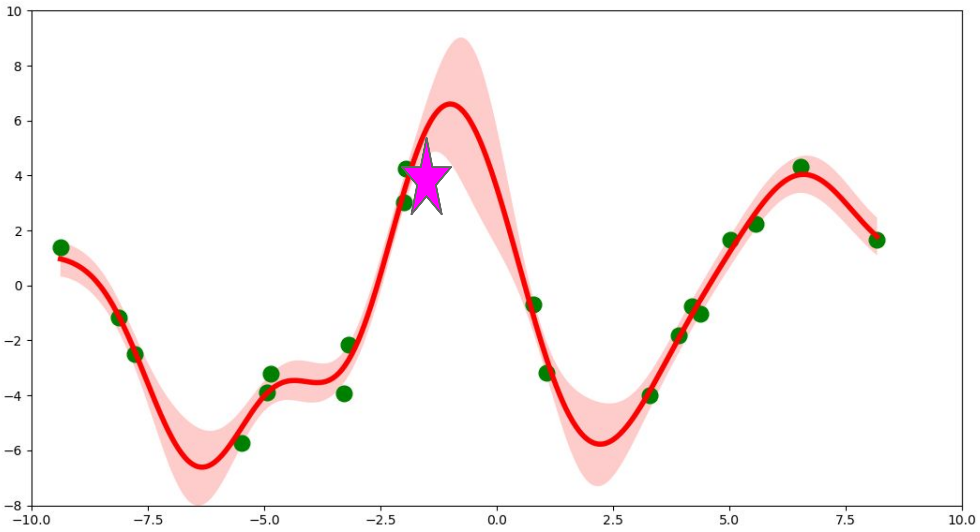


Why Bayesian Inference?

Propagate uncertainty

1. We know the density of likely solutions, i.e. the posterior distribution
2. We can pose question and reason probabilistically
3. E.g. what is the probability that p(y>6|x= -1.1)? It is 0.625

$$p(w|\mathcal{D})$$



Problem statement

- Necessary calculations often intractable! We need approximations!
- But let’s look at how things can quickly turn ugly ...
- The culprit is the denominator in

$$p(w|\mathcal{D}) = \frac{p(\mathcal{D}|w)p(w)}{p(\mathcal{D})}$$

Problem statement

- Necessary calculations often intractable! We need approximations!
- But let's look at how things can quickly turn ugly ...
- The culprit is the denominator in

$$p(w|\mathcal{D}) = \frac{p(\mathcal{D}|w)p(w)}{p(\mathcal{D})}$$

Analytical calculations possible

$$p(w|\mathcal{D}) = \frac{\prod_{n=1}^{N} \mathcal{N}(y_n|w^T x, \sigma^2)\, \mathcal{N}(w|0, \alpha^{-1} I)}{\int \prod_{n=1}^{N} \mathcal{N}(y_n|w^T x, \sigma^2)\, \mathcal{N}(w|0, \alpha^{-1} I) dw}$$

HITS

Problem statement

- Necessary calculations often intractable! We need approximations!
- But let's look at how things can quickly turn ugly ...
- The culprit is the denominator in

$$p(w|\mathcal{D}) = \frac{p(\mathcal{D}|w)p(w)}{p(\mathcal{D})}$$

Analytical calculations not possible

$$p(w|\mathcal{D}) = \frac{\prod_{n=1}^{N} \mathcal{S}t(y_n|w^T x, \sigma^2, \nu)\ \mathcal{N}(w|0, \alpha^{-1}I)}{\int \prod_{n=1}^{N} \mathcal{S}t(y_n|w^T x, \sigma^2, \nu)\ \mathcal{N}(w|0, \alpha^{-1}I)dw}$$

We need to approximate

- We cannot calculate the exact true posterior p(w|D) $\propto$ p(D|w) p(w)
- But we can find an approximation **q(w)** that is close to p(D|w) p(w)
- How good is the approximation? Use Kullback-Leibler Divergence:

$$D_{KL}(q||p) = -\int q(w) \log \frac{p(\mathcal{D}|w)p(w)}{q(w)} dw$$

- Typically we choose q(w) to be Gaussian:

$$q(w) = \mathcal{N}(w|\mu, \Sigma)$$

We need to approximate

- We cannot calculate the exact true posterior p(w|D) $\propto$ p(D|w) p(w)
- But we can find an approximation **q(w)** that is close to p(D|w) p(w)
- How good is the approximation? Use Kullback-Leibler Divergence:

$$D_{KL}(q||p) = -\int q(w) \log \frac{p(\mathcal{D}|w)p(w)}{q(w)} dw$$

- Typically we choose q(w) to be Gaussian:

$$q(w) = \mathcal{N}(w|\mu, \Sigma)$$

Free parameters to optimise

HITS

Kullback-Leibler divergence (this is how variational inference is done)

$$-\int q(w) \log \frac{p(\mathcal{D}, w)}{q(w)} dw$$

Kullback-Leibler divergence (this is how variational inference is done)

$$-\int q(w)\log\frac{p(\mathcal{D},w)}{q(w)}dw = -\int q(w)\log p(D,w)p(w) + \int q(w)\log q(w)dw$$

HITS

Kullback-Leibler divergence (this is how variational inference is done)

$$-\int q(w)\log\frac{p(\mathcal{D},w)}{q(w)}dw = -\int q(w)\log p(D,w)p(w) + \int q(w)\log q(w)dw$$

$$= -\int q(w)\log p(D,w)p(w) - \mathcal{H}[q]$$

HITS

Kullback-Leibler divergence (this is how variational inference is done)

$$-\int q(w)\log\frac{p(\mathcal{D},w)}{q(w)}dw = -\int q(w)\log p(D,w)p(w) + \int q(w)\log q(w)dw$$

$$= -\int q(w)\log p(D,w)p(w) - \mathcal{H}[q]$$

$$= -\int \mathcal{N}(w|\mu,\Sigma)\log p(\mathcal{D},w)dw - \mathcal{H}[q]$$

HITS

Kullback-Leibler divergence (this is how variational inference is done)

$$-\int q(w)\log\frac{p(\mathcal{D},w)}{q(w)}dw = -\int q(w)\log p(D,w)p(w) + \int q(w)\log q(w)dw$$

$$= -\int q(w)\log p(D,w)p(w) - \mathcal{H}[q]$$

$$= -\int \mathcal{N}(w|\mu,\Sigma)\log p(\mathcal{D},w)dw - \mathcal{H}[q]$$

$$\approx -\frac{1}{S}\sum_{s=1}^{S}\log p(\mathcal{D},w_s) - \mathcal{H}[q]$$

HITS

Kullback-Leibler divergence (this is how variational inference is done)

$$-\int q(w)\log\frac{p(\mathcal{D},w)}{q(w)}dw = -\int q(w)\log p(D,w)p(w) + \int q(w)\log q(w)dw$$

$$= -\int q(w)\log p(D,w)p(w) - \mathcal{H}[q]$$

$$= -\int \mathcal{N}(w|\mu,\Sigma)\log p(\mathcal{D},w)dw - \mathcal{H}[q]$$

$$\approx -\frac{1}{S}\sum_{s=1}^{S}\log p(\mathcal{D},w_s) - \mathcal{H}[q]$$

$$= -\frac{1}{S}\sum_{s=1}^{S}\log p(\mathcal{D}|\mu+\Sigma^{\frac{1}{2}}z_s) - \mathcal{H}[q]$$

Use property

$$z_s \sim \mathcal{N}(0,I)$$
$$w_s = \mu + \Sigma^{\frac{1}{2}}z_s$$

HITS

Kullback-Leibler divergence (this is how variational inference is done)

$$-\int q(w)\log\frac{p(\mathcal{D},w)}{q(w)}dw = -\int q(w)\log p(D,w)p(w) + \int q(w)\log q(w)dw$$

$$= -\int q(w)\log p(D,w)p(w) - \mathcal{H}[q]$$

$$= -\int \mathcal{N}(w|\mu,\Sigma)\log p(\mathcal{D},w)dw - \mathcal{H}[q]$$

$$\approx -\frac{1}{S}\sum_{s=1}^{S}\log p(\mathcal{D},w_s) - \mathcal{H}[q]$$

$$= -\frac{1}{S}\sum_{s=1}^{S}\log p(\mathcal{D}|\mu + \Sigma^{\frac{1}{2}}z_s) - \mathcal{H}[q]$$

Free parameters

Objective to minimise

Use property

$$z_s \sim \mathcal{N}(0,I)$$
$$w_s = \mu + \Sigma^{\frac{1}{2}}z_s$$

HITS

Kullback-Leibler divergence (this is how variational inference is done)

$$-\int q(w)\log\frac{p(\mathcal{D},w)}{q(w)}dw = -\int q(w)\log p(D,w)p(w) + \int q(w)\log q(w)dw$$

$$= -\int q(w)\log p(D,w)p(w) - \mathcal{H}[q]$$

$$= -\int \mathcal{N}(w|\mu,\Sigma)\log p(\mathcal{D},w)dw - \mathcal{H}[q]$$

$$\approx -\frac{1}{S}\sum_{s=1}^{S}\log p(\mathcal{D},w_s) - \mathcal{H}[q]$$

$$= -\frac{1}{S}\sum_{s=1}^{S}\log p(\mathcal{D}|\mu + \Sigma^{\frac{1}{2}}z_s) - \mathcal{H}[q]$$

Free parameters

Objective to minimise

M. Titsias, M. Lazaro-Gredilla, 2014
D. P. Kingma, M. Welling, 2014
N. Gianniotis, C. Schnorr etal, 2015
N. Depraetere, M. Vandebroek, 2016

Use property

$$z_s \sim \mathcal{N}(0,I)$$
$$w_s = \mu + \Sigma^{\frac{1}{2}} z_s$$

HITS

Gaussian posterior needs many parameters

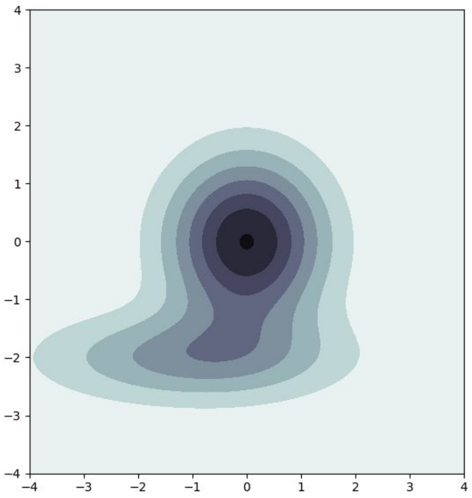
$$q(w) = \mathcal{N}(w|\mu, \Sigma)$$

- We need **d** parameters for the mean μ
- We need **d(d+1)/2** parameters for the covariance $\sum$
- E.g. a problem with $w \in \mathbb{R}^{50}$, needs 1325 Gaussian parameters!
- Strategy:reduce the number of parameters in covariance matrix
- Other work:take $\sum$ to be diagonal,**d** parameters only,lose correlations!
- This work: build covariance matrix using Laplace approximation

Interlude: Laplace approximation

Idea: put a Gaussian around the mode

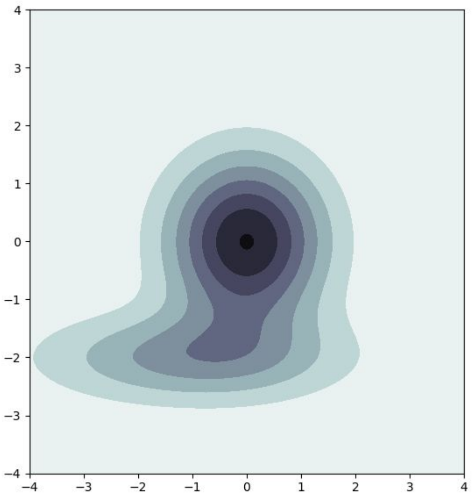
Performed in two steps



$p(w|D) \propto p(D|w)\, p(w)$

Interlude: Laplace approximation

Idea: put a Gaussian around the mode



$p(w|D) \propto p(D|w)\, p(w)$

Performed in two steps

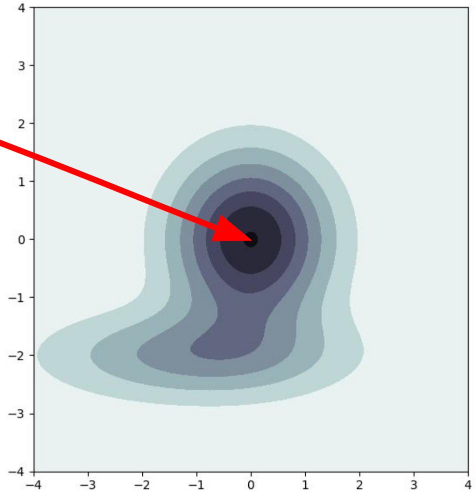
1. Locate mode m by following gradient

Interlude: Laplace approximation

Idea: put a Gaussian around the mode

Performed in two steps

1. Locate mode m by following gradient



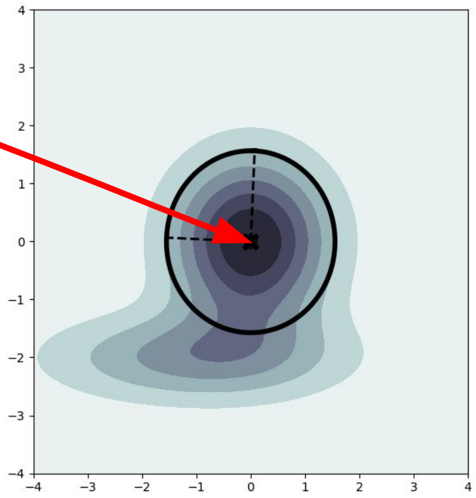
$p(w|D) \propto p(D|w)\, p(w)$

Interlude: Laplace approximation

Idea: put a Gaussian around the mode

Performed in two steps

1. Locate mode m by following gradient
2. Calculate local curvature at m as Hessian H



$p(w|D) \propto p(D|w)\, p(w)$

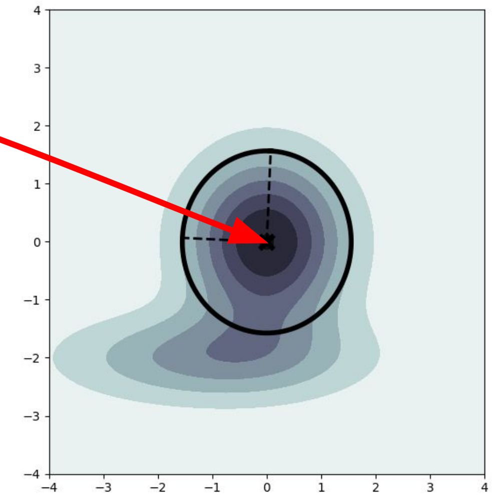
HITS

Interlude: Laplace approximation

Idea: put a Gaussian around the mode

Performed in two steps

1. Locate mode m by following gradient
2. Calculate local curvature at m as Hessian H



$p(w|D) \propto p(D|w)\, p(w)$

Gaussian posterior obtained via Laplace reads:

$$\mathcal{N}(w|\mu_{LA} = m, \Sigma_{LA} = H^{-1})$$

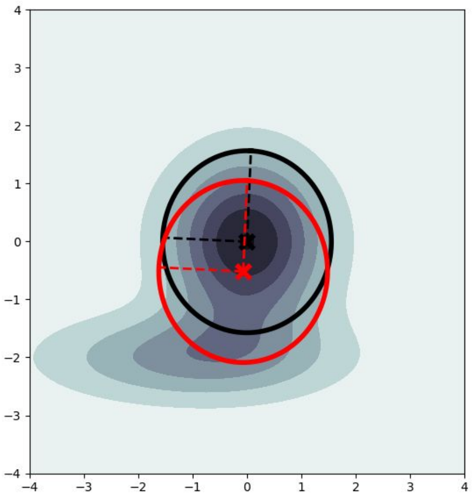
Mixed Variational Inference

Mixed Variational Inference - first proposal

- Take posterior covariance matrix Σ_{LA} from Laplace

- Define new approximate posterior with

  only **d** free parameters:



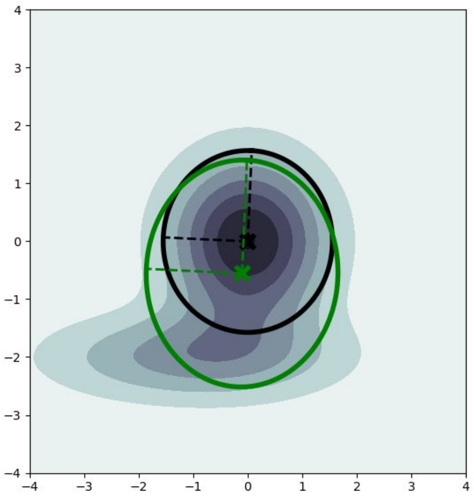
$$q(w) = \mathcal{N}(w|\mu, \Sigma_{LA})$$

$p(w|D) \propto p(D|w)\, p(w)$

Mixed Variational Inference - second proposal

- Take covariance matrix from Laplace and do eigenvalue decomposition

$$\Sigma_{LA} = U \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ & & \lambda_d \end{bmatrix} U^T$$



p(w|D)∝p(D|w) p(w)

- Define new covariance matrix as:

$$\Sigma_{eig} = U \begin{bmatrix} c_1 & & 0 \\ & \ddots & \\ & & c_d \end{bmatrix} U^T$$

- New approximate posterior had only **2d** free parameters:

$$q(w) = \mathcal{N}(w|\mu, \Sigma_{eig})$$

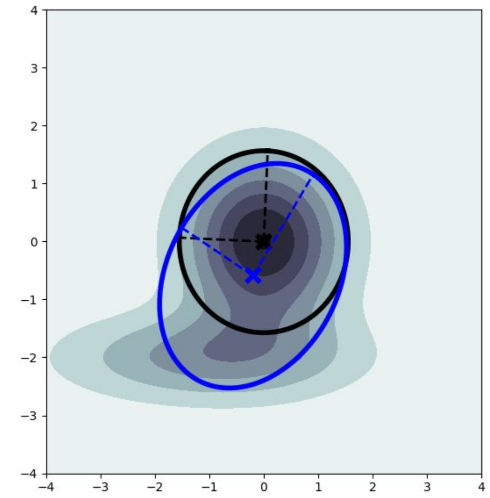
HITS

Mixed Variational Inference - third proposal

- Define two free parameter vectors $U, V \in \mathbb{R}^d$
- Take covariance from Laplace and do cholesky

$$\Sigma_{LA} = C_{LA} C_{LA}^T$$

- Define $L = C_{LA} + UV^T$



p(w|D)∝p(D|w) p(w)

- New approximate posterior had only **3d** free parameters:

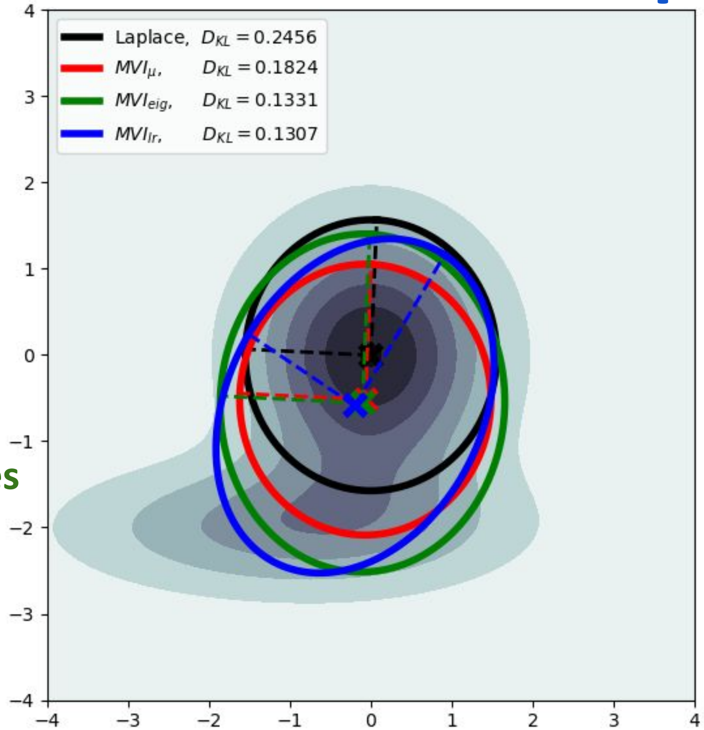
$$q(w) = \mathcal{N}(w|\mu, LL^T)$$

HITS

Mixed Variational Inference - all proposals

MVIμ
adaptation of mean
d # parameters

MVIeig
adaptation of mean, eigenvalues
2d # parameters



MVIlr
adaptation of mean and low rank
3d # parameters

In contrast:
Gaussian with full covariance
d + d(d+1)/2 # parameters

$$p(w|D) \propto p(D|w)\, p(w)$$

Numerical simulations

- We compare with Laplace and with Gaussian diagonal posterior
- Compare algorithms in terms of predictive log-likelihood, this means
    - sample S number of likely solutions from posterior

$$w_s \sim q(w)$$

    - plug solution in likelihood to see how well we explain the data

$$\frac{1}{S}\sum_{s=1}^{S} p(\mathcal{D}_{test}|w_s)$$

Numerical simulations - Logistic regression

MEDIAN LPD ON TEST DATA FOR LOGISTIC REGRESSION OVER 100 RUNS ON THE DATASETS (HIGHER IS BETTER).

Dataset	Q	N	N_{test}	**Laplace**	$\mathbf{MVI}_{\mu}$	$\mathbf{MVI}_{\mathbf{eig}}$	$\mathbf{MVI}_{\mathbf{lr}}$	$\mathbf{VI}_{\mathbf{diag}}$
Banana	2	400	4900	-1238.76	-1219.19	-1221.41	$\mathbf{-1212.19}^{\bullet}$	-1253.36
Breast cancer	9	200	77	-42.82	-42.65	-42.53	$\mathbf{-42.38}$	-45.42
Diabetis	8	468	300	-145.98	-145.468	-145.31	$\mathbf{-144.89}^{\bullet}$	-193.27
Solar	9	666	400	-232.64	-232.37	-232.42	$\mathbf{-232.07}^{\bullet}$	-234.52
German	20	700	300	-151.71	-151.42	-151.31	$\mathbf{-150.70}^{\bullet}$	-179.29
Heart	13	170	100	-39.25	-38.973	-38.96	$\mathbf{-38.62}$	-48.37
Image	18	1300	1010	-304.33	-291.91	-284.19	-284.60	$\mathbf{-283.80}$
Ringnorm	20	400	7000	-309.87	$\mathbf{-308.80}$	-319.67	-309.80	-342.627
Splice	60	1000	2175	-1156.80	-900.00	$\mathbf{-897.33}$	-900.30	-899.501
Thyroid	5	140	75	-11.01	-10.189	-10.189	$\mathbf{-9.844}^{\bullet}$	-10.280
Titanic	3	150	2051	$\mathbf{-1018.92}^{\bullet}$	-1019.59	-1021.7	-1020.62	-1023.91
Twonorm	20	400	7000	-452.57	-450.716	-461.10	$\mathbf{-447.28}$	-543.115
Wavenorm	21	400	4600	-947.66	$\mathbf{-946.49}$	-948.62	-950.31	-969.61

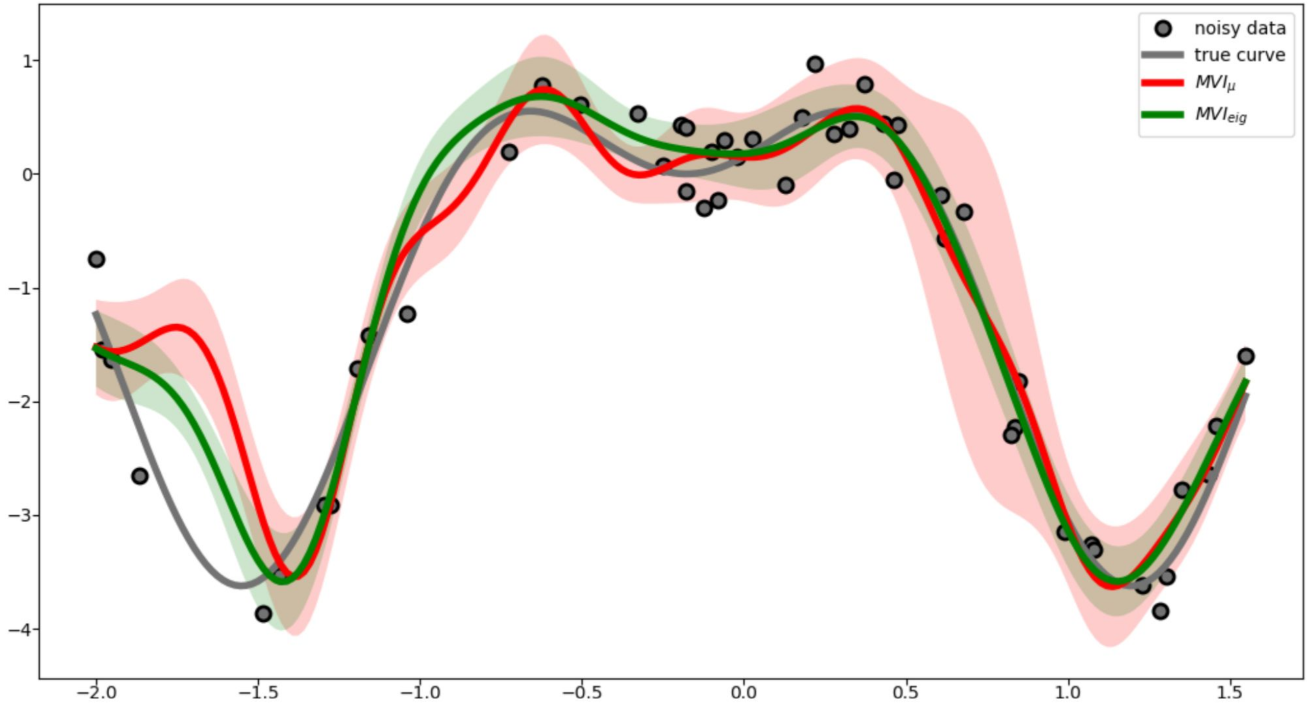
Numerical simulations - Multiclass regression

MEDIAN LPD ON TEST DATA FOR MULTICLASS LOGISTIC REGRESSION OVER 100 RUNS ON THE DATASETS (HIGHER IS BETTER).

Dataset	K	Q	N	N_{test}	**Laplace**	**MVI$_{\mu}$**	**MVI$_{\mathrm{eig}}$**	**MVI$_{\mathrm{lr}}$**	**VI$_{\mathrm{diag}}$**
Ecoli	8	7	236	100	-50.32	-48.80	-49.31	$\mathbf{-48.52}^{\bullet}$	-51.39
Crabs	4	5	140	60	-64.59	-64.11	-64.28	$\mathbf{-64.10}$	-68.92
Iris	3	4	105	45	-9.06	-7.53	$\mathbf{-6.42}$	-7.46	-8.17
Soybean	4	35	33	14	-4.10	-2.35	$\mathbf{-0.66}^{\bullet}$	-2.36	-1.67
Wine	3	13	125	53	-5.72	-4.01	$\mathbf{-3.33}^{\bullet}$	-3.94	-4.66
Glass	6	9	150	64	-61.35	-60.39	$\mathbf{-59.79}$	-60.44	-76.26
Vehicle	4	18	593	293	-159.783	$\mathbf{-158.39}^{\bullet}$	-158.60	-159.15	-174.725
Balance	3	4	438	187	-23.5734	$\mathbf{-22.7321}$	-23.2197	-23.078	-31.587

Numerical simulations - Regression with Cauchy errors

	Laplace	MVI_μ	MVI_{eig}	MVI_{lr}	VI_{diag}
LPD	-0.818	-0.771	$\mathbf{-0.722}$	-0.726	-0.736



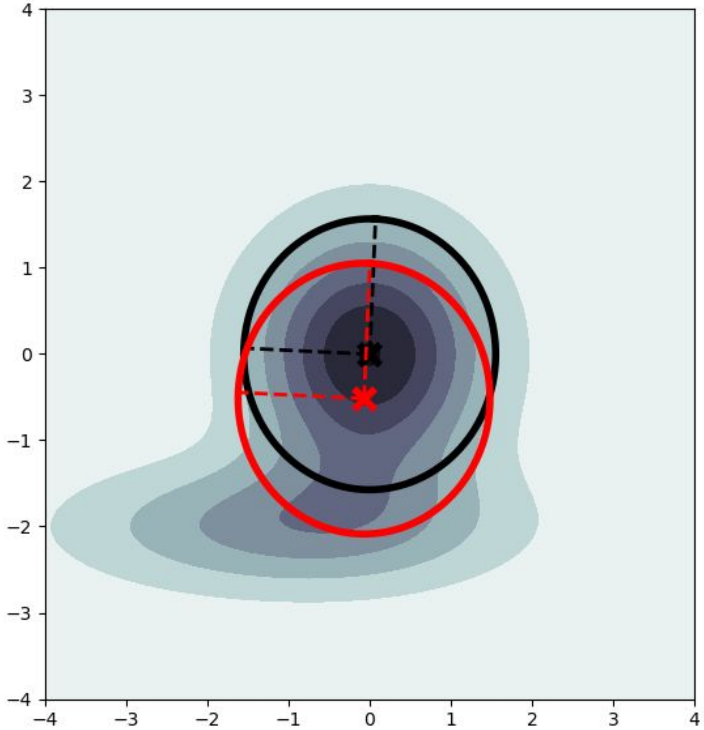
HITS

Conclusions

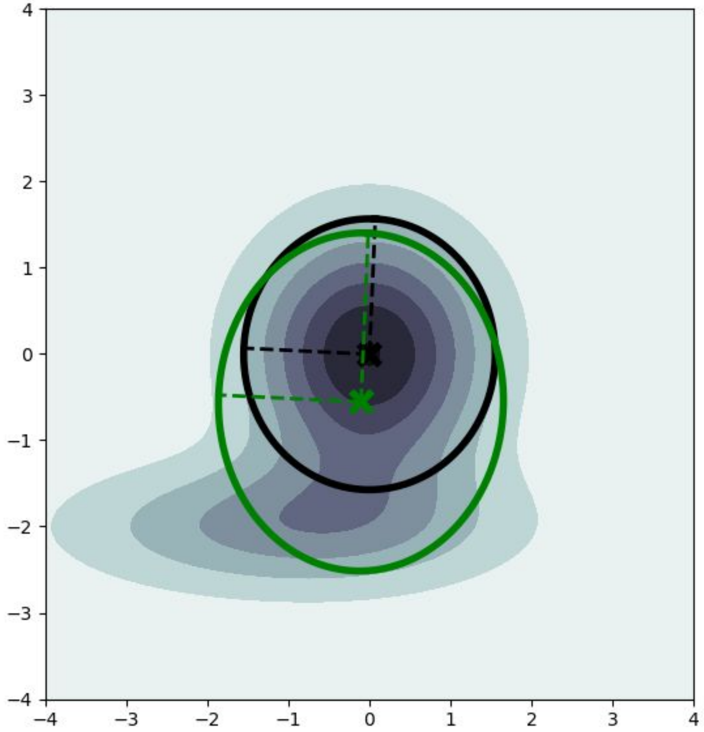
- Proposed a way to make Bayesian inference
- Contributions:
  - applicable when calculations are intractable (e.g. non-conjugate)
  - we manage to limit the numbers of free parameters
  - proposed q(w) retains correlations in contrast to diagonal posterior
- Demonstrated practical advantages in benchmark problems

HITS

Mixed Variational Inference - first proposal



Mixed Variational Inference - second proposal



Mixed Variational Inference - third proposal