Compromise-free Bayesian sparse reconstruction

Higson, Handley, Hobson, Lasenby (arxiv:1809.04598)

Will Handley
wh260@cam.ac.uk

Kavli Institute for Cosmology
Cavendish Laboratory (Astrophysics Group)
University of Cambridge

March 19th 2019

Bayesian inference

- Bayes theorem for parameter estimation

$$\Pr(D|\theta, M) \times \Pr(\theta|M) = \Pr(\theta|D, M) \times \Pr(D|M)$$

$$\mathcal{L} \times \pi = \mathcal{P} \times \mathcal{Z} \qquad \text{Likelihood} \times \text{Prior} = \text{Posterior} \times \text{Evidence}$$

- Bayes theorem for model comparison

$$\Pr(M_i|D) = \frac{\Pr(D|M_i)\Pr(M_i)}{\sum_j \Pr(D|M_j)\Pr(M_j)} \equiv \frac{\mathcal{Z}_i \times \Pi_i}{\sum_j \mathcal{Z}_j \Pi_j}$$

- Model marginalisation

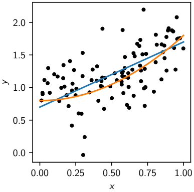
$$\Pr(\alpha|D) = \sum_j \Pr(\alpha|M_j, D)\Pr(M_j|D) \equiv \sum_j \mathcal{P}_j(\alpha) \times \Pi_j$$

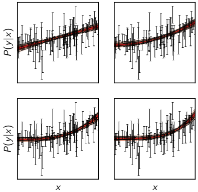
- Bayesian inference depends on parameter and model priors $\pi(\theta|M)$ and $\Pi(M)$, e.g:
  - If the prior adjusts the shape of the likelihood
  - If the prior changes its width
- My definition of Bayesianism vs Frequentism is whether you consider this prior dependency a feature or a bug.
- Other important quantities: Shannon information $\mathcal{I}$, Kullback-Leibler divergence $\mathcal{D}$ and Bayesian model dimensionality d

$$\mathcal{I}(\theta) = \log\frac{\mathcal{P}}{\pi} \qquad \mathcal{D} = \int \mathcal{P}\log\frac{\mathcal{P}}{\pi}d\theta \equiv \left\langle \log\frac{P}{\pi}\right\rangle_{\mathcal{P}} \equiv \langle\mathcal{I}\rangle_{\mathcal{P}} \qquad \frac{d}{2} = \left\langle(\mathcal{I}-\mathcal{D})^2\right\rangle_{\mathcal{P}}$$

Handley & Lemos: arXiv:1902.04029, arXiv:1903.06682

- Fit D data points $(\mathbf{x}_d, y_d)$ with some function $y = f(\mathbf{x};\theta)$ with free parameters θ





- If $\mathbf{x}$ is 2-dimensional $\Rightarrow$ image reconstruction, etc.

- Model function $f(\mathbf{x};\theta)$ as sum of N basis functions $\varphi^{(T)}$ of type T, with weights a_i and location/shape parameters p_i, so $\theta = (T, N, \mathbf{a}, \mathbf{p}_1, \ldots, \mathbf{p}_N)$:

$$f(\mathbf{x};\theta) = \sum_{i=1}^{N} a_i \varphi^{(T)}(\mathbf{x};\mathbf{p}_i)$$

- Explore (by sampling) posterior with variable (effective) dimensionality:

$$\Pr(T, N, \mathbf{a}, \{\mathbf{p}_i\}|\mathbf{y}) \propto \underbrace{\Pr(\mathbf{y}|T, N, \mathbf{a}, \{\mathbf{p}_i\})}_{\text{likelihood}} \underbrace{\Pr(\mathbf{a}|T)\Pr(\{\mathbf{p}_i\}|T, N)\Pr(N|T)\Pr(T)}_{\text{prior}}$$

- Full posterior of fit $\Pr(\mathbf{f}|\mathbf{y}) = \int \Pr(\mathbf{f}|\theta)\Pr(\theta|\mathbf{y})\,d\theta$ ...and that's it!

Desirable properties

$$\Pr(T, N, \mathbf{a}, \{\mathbf{p}_i\}|\mathbf{y}) \propto \underbrace{\Pr(\mathbf{y}|T, N, \mathbf{a}, \{\mathbf{p}_i\})}_{\text{likelihood}} \underbrace{\Pr(\mathbf{a}|T)\Pr(\{\mathbf{p}_i\}|T, N)\Pr(N|T)\Pr(T)}_{\text{prior}}$$

- Full posterior on parameters (rather than simply optimising) $\Rightarrow$ quantify uncertainties
- Bayesian approach $\Rightarrow$ naturally penalises overcomplex models
- Sparsity can be further enforced directly by $\Pr(N)$ and marginalised over
- No regularisation parameter to be chosen (unlike L_p-norm regularisation, etc.)
- Variable number of basis functions with variable positions
- Basis functions families/shapes determined (dictionary learning) or marginalised over
- Can impose arbitrary constraints on reconstruction (not just positivity)
- Accommodates any noise type, e.g. Gaussian, Poisson, etc. (extra hyperparameters)
- Accommodates arbitrary missing and/or irregular data

Practical considerations

$$\Pr(T,N,\mathbf{a},\{\mathbf{p}_i\}|\mathbf{y}) \propto \underbrace{\Pr(\mathbf{y}|T,N,\mathbf{a},\{\mathbf{p}_i\})}_{\text{likelihood}} \underbrace{\Pr(\mathbf{a}|T)\Pr(\{\mathbf{p}_i\}|T,N)\Pr(N|T)\Pr(T)}_{\text{prior}}$$

- Transdimensional sampling (RJMCMC) costly $\Rightarrow$ use product-space approach
  – consider hypermodel H with space θ of fixed dimensionality $T_{\max}\times N_{\max}$
  – integer parameters (T,N) enumerate models H_M within H
  – for each sampled (T,N)-values, partition θ into parameters used by H_M and others
  – latter set of parameters ignored (not passed by 'wrapper' to likelihood for H_M)
- Marginalisation over $\mathbf{a}$ and $\{\mathbf{p_i}\}$ $\Rightarrow$ posterior $\Pr(T,N|\mathbf{y})$ (recovers PORs)

$$\mathcal{P}_{(T,N)}^{(T',N')} = \ln\left[\frac{\Pr(T',N'|\mathbf{y})}{\Pr(T,N|\mathbf{y})}\right]$$

- i.e. Bayesian model selection without evidences! (can also use 'vanilla' method)
- But. . .
  – posterior $\Pr(T,N,\mathbf{a},\{\mathbf{p}_i\}|\mathbf{y})$ dimensionality $N_{\rm dim}\sim 10^3-10^4$ for small images
  – posterior is highly multimodal with strong degeneracies (certainly non-convex!)
  – categorical/integer parameters $T,N \Rightarrow$ cannot use gradients
  $\Rightarrow$ use (dynamic) nested sampling to explore posterior with (dy)PolyChord
- Computationally demanding, but now possible (proof of principle) . . .

Nested sampling

- Want to compute evidence, which is high-dimensional integral over parameter space θ. Define prior volume X as fraction of prior above contour $\mathcal{L}(\theta) \ge \mathcal{L}$

$$\mathcal{Z} = \int \mathcal{L}(\theta)\pi(\theta)d\theta = \int \mathcal{L}(X)dX, \qquad X(\mathcal{L}) = \int_{\mathcal{L}(\theta)\ge\mathcal{L}} \pi(\theta)d\theta$$

- Nested sampling procedure:
  1. Draw N "live" points from the prior $\pi(\theta)$ and compute likelihoods.
  2. Remove lowest live point, replace with one drawn from prior at higher likelihood.
  3. Repeat step 2 until live points occupy a small enough prior volume.
- Procedure allows one to estimate prior volumes probabilistically, as volume contracts by factor $\approx \frac{N}{N+1}$ at each step.
- Compute evidence from M discarded points via trapezium rule:

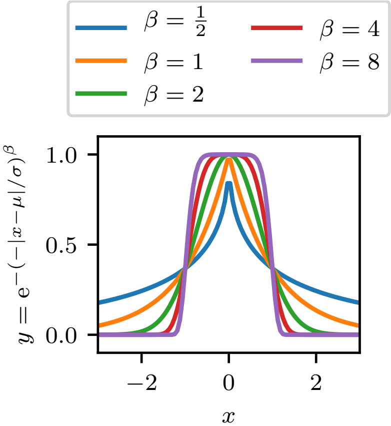
$$\mathcal{Z} \approx \sum_{i=0}^{M} \mathcal{L}_i \times \frac{1}{2}(X_{i-1} - X_{i+1}), \qquad X_0 = 1, \quad X_N = 0, \quad X_i = t_i X_{i-1}, \quad \Pr(t_i) = N t^{N-1}$$

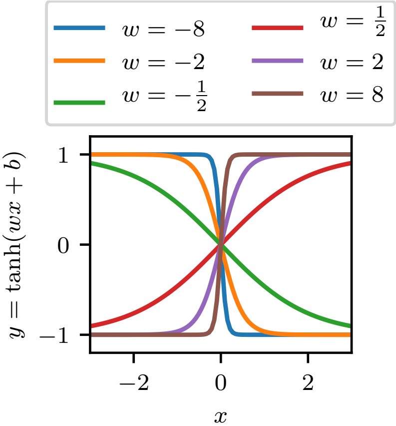
- Generates posteriors as by-product with weights $w_i = \frac{1}{\mathcal{Z}}\mathcal{L}_i \times \frac{1}{2}(X_{i-1} - X_{i+1})$
- Step 2 is by far the hardest step.

MultiNest	Ellipsoidal based rejection sampling
Galilean	Gradient-based HMC-like algorithm
Diffusive NS & Dynesty	User-based choice
PolyChord	Slice-sampling

Simple basis functions

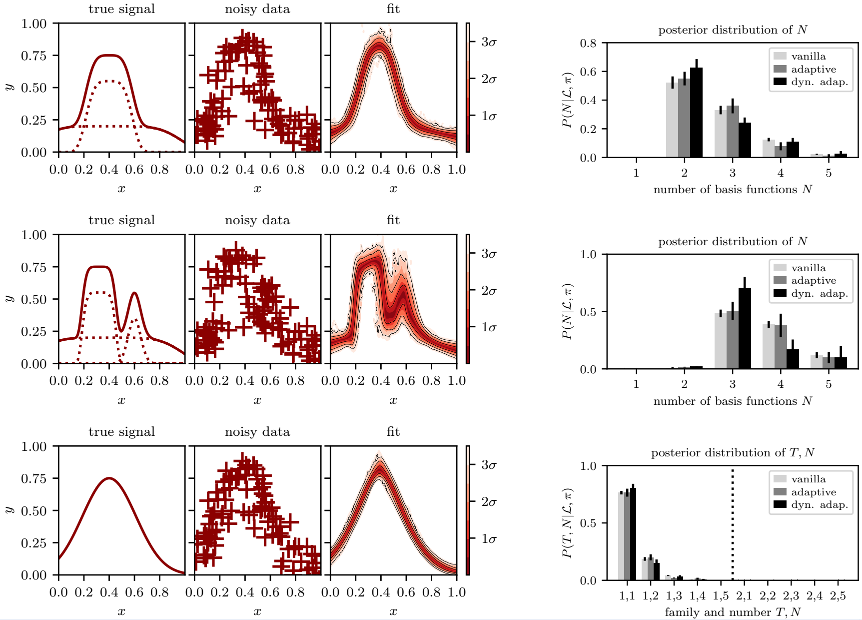
- 1-d generalised Gaussians $\varphi^{(\mathrm{g})}(x;\mathbf{p}) = \varphi^{(\mathrm{g})}(x;\mu,\sigma,\beta) = e^{-(|x-\mu|/\sigma)^\beta}$ (GGMM)
- 1-d tanh functions $\varphi^{(\mathrm{t})}(x;\mathbf{p}) = \varphi^{(\mathrm{t})}(x;w,b) = \tanh(wx+b)$ (TMM)



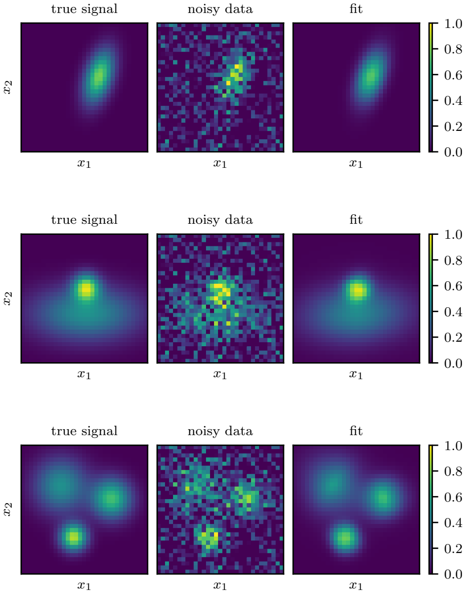


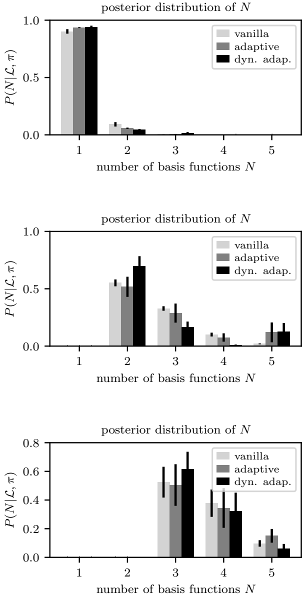
- Easily extended to higher dimensions (including anisotropic scaling and rotation)

Simple 1-D examples: generalised Gaussians data

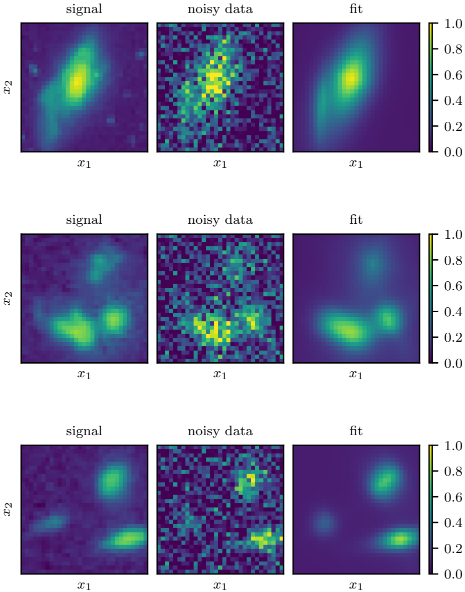


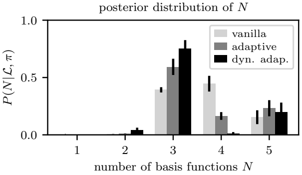
Simple 2-D examples: generalised Gaussians data

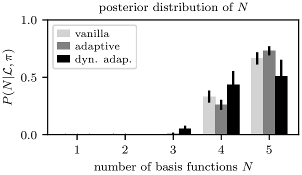


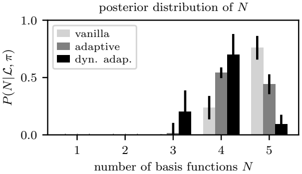


HST eXtreme Deep Field images: generalised Gaussians fit

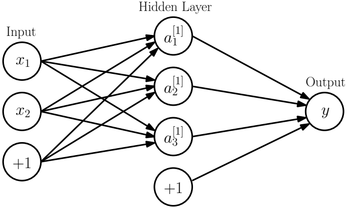








- Consider feed-forward NN, d-dimensional input $\mathbf{x}$, one hidden layer with N nodes:



$$
\begin{aligned}
a_j^{[1]} &= \phi^{[1]}\left(\sum_{i=1}^{d} x_i w_{ji}^{[1]} + b_j^{[1]}\right)\\
\hat{y}_j = a_j^{[2]} &= \phi^{[2]}\left(\sum_{i=1}^{N} a_i^{[1]} w_{ji}^{[2]} + b_j^{[2]}\right)
\end{aligned}
$$

- If activation functions $\phi^{[1]}(x) = \tanh x$ and $\phi^{[2]}(x) = x$, and $b_j^{[2]} = 0$
  $\Rightarrow$ consider (noisy) data points $(\mathbf{x}^{(t)}, y^{(t)})$ $(t = 1, 2, \ldots, T)$ as training set with objective function equal to likelihood $\Pr(\mathbf{y}|\hat{\mathbf{y}})$ (noise model)

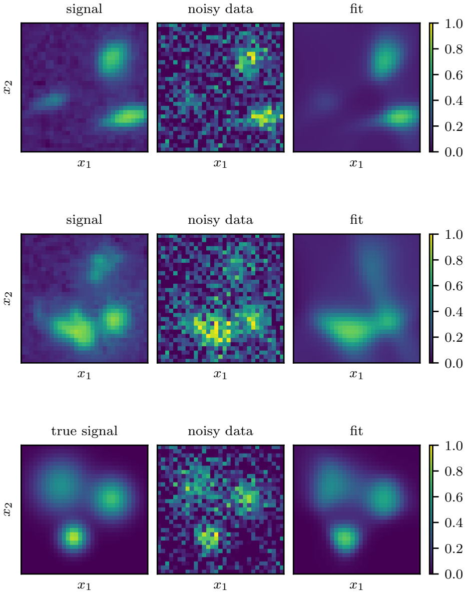
  $\Rightarrow$ regression problem with N adaptive tanh basis functions (or $\mathrm{sig}(x)$ or $\max(0, x)$)

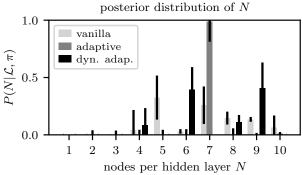
$$
\hat{y}(\mathbf{x}) = f(\mathbf{x}) = \sum_{j=1}^{N} a_j^{[1]} \tanh\left(\sum_{i=1}^{d} x_i w_{ji}^{[1]} + b_j^{[1]}\right) \qquad \text{(Activation function MM)}
$$

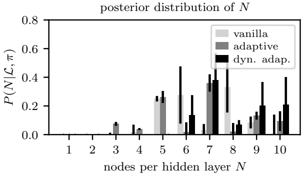
Bayesian neural networks architecture

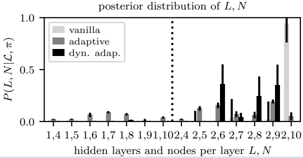
- In general: NN can have L hidden layers with nodes $N^{[1]}, \ldots, N^{[L]}$ & many outputs
- output(s) no longer direct sum(s) of inputs but method still applicable
- can determine integer parameters $(L, \{N^{[l]}\})$ and activation type T
- simultaneous training of network parameters, architecture and activation function
- full joint posterior distribution on all aspects of NN

2-D examples: generalised Gaussians & HST images









Bayesian inference from simulations

(a.k.a. Likelihood Free Inference)

- In many cases, do not have access to likelihood $\Pr(D|\theta, M)$
- Can however simulate data $D = \phi(\theta, M)$
- Must compress data in order to avoid curse of dimensionality $t = t(D)$
- Massive compression: $\dim(t) = \dim(\theta)$ (Alsing et al arXiv:1801.01497)

1. Construct proxy joint/conditional distribution $p = \Pr(t, \theta|\eta)$ with nuisance η:
   - Gaussian mixture model, $x = (t, \theta)$, $\eta = (N, A_1, \mu_1, \sigma_1, \ldots, A_N, \mu_N, \sigma_N)$:
   $$p(t, \theta|\eta) = \sum_{i=1}^{N} A_i \exp\left(-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right) \qquad \text{(Alsing et al arXiv:1801.01497)}$$
   - Neural density estimator $x = (t, \theta)$, $\eta = (\mathrm{Architecture}, \mathbf{w})$ (Alsing et al arXiv:1903.00007):
2. Compute example simulations $\{(t_i, \theta_i)\}$
3. Fit proxy to simulations via η given prior $\Pr(\eta)$, using likelihood
   $$\mathcal{L}(\eta) = \prod_i p(t_i, \theta_i|\eta)$$
4. Marginalise over proxy (ignore η column in samples), evaluated at observed data D
   $$\Pr(\theta, D) = \int p(\theta, t(D)|\eta) \Pr(\eta) d\eta$$
5. Condition on data D, either analytically or via nested sampling
   $$\Pr(\theta|D) = \Pr(\theta, D)/\Pr(D) \qquad \Pr(D) = \int \Pr(\theta, D) d\theta$$

Likelihood free inference: what's in a name?

- The term "Likelihood-free" is a misnomer – there is still very much a likelihood involved at the centre of the analysis, we just don't analytically compute it
- From the Bayesian viewpoint, in lieu of attempting an impossible calculation of a likelihood, we construct a proxy, and marginalise over our lack of knowledge.
- Before becoming involved in this hack week, I found the term LFI disconcerting.
- Alternative names
  - Simulation-based inference
  - Likelihood learning
- Proposed Hack: Come up with a different name for those outside the field