Unscented Variational Bayesian Inference

Anaïk Olivero[†], Liva Ralaivola[*]

[†]Laboratoire d'Analyse, Topologie, Probabilités, Marseille

[*]Laboratoire d'Informatique Fondamentale, Marseille

Journée GdR ISIS "Traitement du signal de données à valeurs complexes", 12 juin 2013

- **The phase Retrieval Problem**

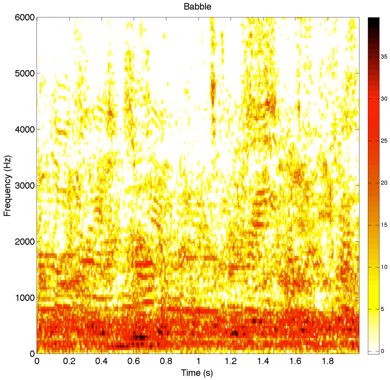
$$\text{Given } b \in \mathbb{C}^m, \text{ find } x \in \mathbb{C}^n \text{ such that } |Ax| = b,$$

where A is a matrix in $\mathbb{C}^{m\times n}$.

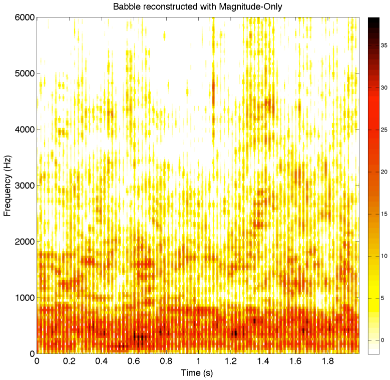
- **Audio Applications**
  - Source separation [Sturmel, 2011]
  - Sound transformation in the time-frequency domain [Olivero, 2012]
  - Synthesis via scattering coefficients [Mallat, 2012]
- **Solver for the phase retrieval problem**
  - Iterative algorithms like [Griffin and Lim, 1984]
  - Recent algorithms used convex relaxation techniques and semi-definite programming [Candes, 2011]
- **Our contribution**
  - Bayesian model
  - Unscented transform

Example in the Gabor case with a babble signal

Signal reconstructed via Ax



Signal reconstructed via $|Ax|$



$$\text{Given } b \in \mathbb{R}_+^m, \text{ find } x \in \mathbb{C}^n \text{ such that } |Ax| = b,$$

where A is a matrix in $\mathbb{C}^{m \times n}$.

$$\text{Find } Ax = y \in \mathbb{C}^m \text{ such that } \begin{cases} |y| = b \\ \underbrace{(I - AA^\dagger)}_{\Pi} y = 0 \end{cases}$$

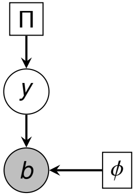
For noisy y and b, the model reads :

$$p(y) \propto exp(-\|\Pi y\|^2), \quad y \in \mathbb{C}^m$$

$$p(b_m | y_m, \phi_m) \propto \exp\left\{-\phi_m \left(|y_m| - b_m\right)^2\right\}$$

If $\phi_m = \phi \in \mathbb{R}_+$, this model amounts to

$$\min_y \|\Pi y\|^2 + \phi \||y| - b\|^2$$



Phase Retrieval Problem

$$\text{Given } b \in \mathbb{R}_+^m, \text{ find } x \in \mathbb{C}^n \text{ such that } |Ax| = b,$$

where A is a matrix in $\mathbb{C}^{m \times n}$.

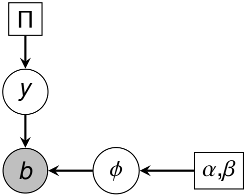
$$\text{Find } Ax = y \in \mathbb{C}^m \text{ such that } \begin{cases} |y| = b \\ \underbrace{(I - AA^\dagger)}_{\Pi} y = 0 \end{cases}$$

For noisy y and b, the model reads :

$$p(y) \propto exp(-\|\Pi y\|^2), \quad y \in \mathbb{C}^m$$

$$p(b_m|y_m, \phi_m) \propto \exp\left\{-\phi_m \left(|y_m| - b_m\right)^2\right\}$$

$$p(\phi_m) \propto \text{Gamma}(\alpha, \beta)$$



- H : hidden random variables
- V : visible random variables (i.e. observations)
- $\mathcal{M}$: model defining the probabilistic dependencies between H and V

Goal : evaluate the model posterior probablity

$$p(\mathcal{M}|V = v) \propto p(V = v|\mathcal{M})\pi(\mathcal{M})$$

- H : hidden random variables
- V : visible random variables (i.e. observations)
- $\mathcal{M}$: model defining the probabilistic dependencies between H and V

Goal : evaluate the model posterior probablity

$$p(\mathcal{M}|V = v) \propto p(V = v|\mathcal{M})\cancel{\pi(\mathcal{M})}$$

which comes down to evaluate the log-evidence

$$\mathcal{E}_{\mathcal{M}}(v) = \ln p(V = v|\mathcal{M})$$

- H : hidden random variables
- V : visible random variables (i.e. observations)
- $\mathcal{M}$: model defining the probabilistic dependencies between H and V

Goal : evaluate the model posterior probablity

$$p(\mathcal{M}|V = v) \propto p(V = v|\mathcal{M})\cancel{\pi(\mathcal{M})}$$

which comes down to evaluate the log-evidence

$$\mathcal{E}_{\cancel{\mathcal{M}}}(v) = \ln p(V = v|\cancel{\mathcal{M}})$$

Now, we will drop the subscript $\mathcal{M}$

$$
\begin{aligned}
\mathcal{E}(v) &= \ln p(V = v) \\
&= \ln \sum_h p(V = v, H = h) \\
&= \ln \sum_h q(h) \frac{p(V = v, H = h)}{q(h)}, \quad \text{for any } q \in \mathcal{Q} \\
&\geq \underbrace{\underbrace{\mathbb{E}_{H \sim q} \{\ln P(V = v, H)\}}_{\mathcal{F}_V(q)} + \underbrace{\mathbb{E}_{H \sim q} \{-\ln q(H)\}}_{\text{Entropy : } \mathcal{G}(q)}}_{\mathcal{L}_V(q)}
\end{aligned}
$$

$\mathcal{L}_v(q)$ is a lower bound of the evidence (ELBO).

$$
\mathcal{E}(v) = \max_{q \in \mathcal{Q}} \mathcal{L}_v(q),
$$

The optimal distribution is equal to the posterior distribution $p(H|V)$, but ....

$$\mathcal{E}(v) = \max_{q \in \mathcal{Q}} \mathcal{L}_v(q) = \underbrace{\mathbb{E}_{H \sim q}\{\ln P(V = v, H)\}}_{\mathcal{F}_V(q)} + \underbrace{\mathbb{E}_{H \sim q}\{-\ln q(H)\}}_{\text{Entropy}:\, \mathcal{G}(q)}$$

1. $\mathcal{Q}$ is the family of all probabilities of distributions on H.
2. How can we compute $\mathcal{F}_v(q)$ and $\mathcal{G}(q)$?

- The popular Mean-Field approximation restricts $\mathcal{Q}$ to factorized distributions [Jordan and al, 1999].
- [Gershman, 2012] proposed a new method is also called "Nonparametric Variational Inference".

$$\mathcal{E}(v) = \max_{q \in Q} \mathcal{L}_v(q) = \underbrace{\mathbb{E}_{H\sim q}\{\ln P(V=v,H)\}}_{\mathcal{F}_V(q)} + \underbrace{\mathbb{E}_{H\sim q}\{-\ln q(H)\}}_{\text{Entropy : }\mathcal{G}(q)}$$

- [Gershman, 2012]
  1. $q(h) = \frac{1}{N}\sum_{n=1}^{N} \frac{1}{(2\pi\sigma_n^2)^{D/2}} \exp\left\{-\frac{1}{2\sigma_n^2}\|h-\mu_n\|^2\right\}$.
  2. $\ln P(V=v,\cdot)$ is obtained with a second order approximation.
  3. The entropy is lowly bounded.

- We build upon and extend this approach :
  1. we will consider a **non isotropic mixture of gaussians**
  2. we use **unscented transform** to evaluate the expectations.

[Julier and Uhlmann, 1995]

Let X be a random variable with $\mathbb{E}[X] = \overline{X}$ and $cov(X) = RR^T$ and f a twice differentiable function :

$$
\begin{aligned}
Z = f(X) &\approx f(\overline{X}) + \nabla_{\overline{X}} f^T (X - \overline{X}) + \frac{1}{2}(X - \overline{X})^T Hf_{\overline{X}}(X - \overline{X}) \\
\mathbb{E}[f(X)] &\approx f(\overline{X}) \quad + \quad 0 \quad + \quad \frac{1}{2} Tr\{\mathbb{E}[(X - \overline{X})^T (X - \overline{X})] Hf_{\overline{X}}\} \qquad (1) \\
&= f(\overline{X}) + \frac{1}{2} Tr\{cov(X) Hf_{\overline{X}}\}
\end{aligned}
$$

For all random variable with mean $\overline{X}$ and covariance $cov(X)$, we obtain a second-order approximation of $\mathbb{E}[f(X)]$.

- We will construct a discrete random variable χ with values χ_i and probabilities ω_i.

(1) $a^T b = Tr(a^T b) = Tr(ba^T)$

Let X be a random variable with $\mathbb{E}[X] = \overline{X}$ and $cov(X) = RR^T$.

A set of sigma points $\{\chi_i\}$ and weights $\{\omega_i\}$ has the same mean and covariance.

$$\chi_0 = \overline{X} \qquad \omega_0 = \frac{\kappa}{n+\kappa}$$

$$\chi_i = \overline{X} + \sqrt{n+\kappa} R_i \qquad \omega_i = \frac{1}{2(n+\kappa)}$$

$$\chi_{i+n} = \overline{X} - \sqrt{n+\kappa} R_i \qquad \omega_{i+n} = \frac{1}{2(n+\kappa)}, \quad i = 1, .., n$$

R_i is the i-th column of R

Proposition

If f is twice differentiable, then $\mathbb{E}[f(\chi)] = \sum_{i=0}^{2n} \omega_i f(\chi_i)$ is a second order approximation of $\mathbb{E}[f(X)]$.

Let X be a random variable with $\mathbb{E}[X] = \overline{X}$ and $cov(X) = UU^T + VV^T$.

We define a set of *sigma points* : $\{\mathcal{X}_i^u, \mathcal{X}_j^v\}$ with corresponding weights $\{\omega_i^u, \omega_j^v\}$.

For $i = 1, .., n_u$, $j = 1, .., n_v$:

$$\mathcal{X}_i^u = \begin{cases} \overline{X} \\ \overline{X} + \alpha_i^u u_i \\ \overline{X} - \alpha_{i+n}^u u_i \end{cases} \qquad \text{and} \quad \mathcal{X}_j^v = \begin{cases} \overline{X} \\ \overline{X} + \alpha_j^v v_j \\ \overline{X} - \alpha_{j+n}^v v_j \end{cases}$$

with $\sum\limits_i \omega_i^u + \sum\limits_j \omega_j^v = 1$

Proposition

If f is twice differentiable, then $\mathbb{E}[f(\chi)] = \sum\limits_k \omega_k f(\chi_k)$ is a second order approximation of $\mathbb{E}[f(X)]$.

$$\mathcal{E}(v) = \max_{q \in Q} \mathcal{L}_v(q) = \underbrace{\mathbb{E}_{H \sim q}\{\ln P(V = v, H)\}}_{\mathcal{F}_V(q)} + \underbrace{\mathbb{E}_{H \sim q}\{-\ln q(H)\}}_{\text{Entropy}\,:\,\mathcal{G}(q)}$$

In our approach,

1. The family of distributions is described with

$$q(h) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{\sqrt{(2\pi)^D |\Sigma_n|}} \exp\left\{-\frac{1}{2}(x - \mu)^\top \Sigma_n^{-1}(x - \mu)\right\}$$

$$\text{where} \begin{cases} \Sigma_n = \sigma_n^2 I + W_n W_n^\top \\ W_n \in \mathbb{R}^{p \times D} \text{ with } p << D \end{cases}$$

2. The **unscented transform** gives a second order approximation of the expectations.

- Obervations : $V = \{x_t, c_t\}_{t=1}^{T}$, where $x_t \in \mathbb{R}^D$, $c_t \in \{\pm 1\}$.
- Hidden variables : $H = \{w, \alpha\}$, where $w \in \mathbb{R}^D$, $\alpha \in \mathbb{R}_+$

The probabilities describing the possible outcomes are modeled using a logistic function :

$$p(c_t|w, x_t) = \frac{1}{1 + e^{-c_t w^T x_t}}$$

The regression parameters are given with probabilities :

$$p(w|\alpha) = g(w; 0, \alpha^{-1})$$
$$p(\alpha) = \gamma(\alpha; a, b)$$

We compare our approach with the one proposed in [Gershmann, 2012] for simulated data and UCI datasets

- Comparable results

We show a promising new framework for Variational Bayesian Inference, with few hypothesis on the joint probabilitiy involved (twice differentiability).

Advantages :

- Flexibility
- Confidence intervals

Drawbacks :

- Computational cost

Future directions :

- Estimation of non-uniform mixture coefficients
- Situations where the function is L-Lipschitz differentiable
- Model selection
- Intensive experiments for the phase retrieval problem

- ANR Project ASAP
- Julien Audiffren

Thanks for your attention