Extended Variational Inference for Non-Gaussian Statistical Models

Zhanyu Ma
mazhanyu@bupt.edu.cn

Pattern Recognition and Intelligent System Lab.,
Beijing University of Posts and Telecommunications, Beijing, China.

VALSE Webinar
May 20, 2015
Collaborators
References


Outline

Non-Gaussian Statistical Models
- Non-Gaussian vs. Gaussian
- Advantages and Challenges

Variational Inference (VI) and Extended VI
- Formulations and Conditions
- Convergence and Bias

Related Applications
- Beta/Dirichlet Mixture Model
- BG-NMF
Non-Gaussian Statistical Models

• Definition
  – Statistical model for non-Gaussian data
  – Belong to exponential family

  - von Mises-Fisher
    • Directional data
    • $L_2$ norm = 1

  - Dirichlet / Beta
    • Bounded support
    • $L_1$ norm = 1

  - Gamma
    • Semi-bounded support
Why non-Gaussian? OR Why *not* Gaussian?

- Real-life data are not Gaussian
  - Speech Spectra
  - Image pixel value
  - Edge strength in complex network
  - DNA methylation level
  - ............
Non-Gaussian Statistical Models

**Gaussian distribution**

✓ **Advantages**
  - the widely used probability distribution
  - analytically tractable solution
  - Gaussian mixture model can model arbitrary distribution
  - vast applications

✓ **Disadvantages**
  - not all the data are Gaussian distributed
  - unbounded support and symmetric shape for bounded/semi-bounded/well-structured data
  - flexible model with the cost of high model complexity
Non-Gaussian Statistical Models

- Non-Gaussian distribution

  ✔ Advantages
  - well defined for bounded/semi-bounded/well-structured data
  - belong to exponential family $\rightarrow$ mathematical convenience and conjugate match
  - non-Gaussian mixture model can model data more efficiently

  ✔ Disadvantages
  - numerically challenging in parameter estimation, both ML and Bayesian estimations!
  - lack of closed-form solution for real applications
Example 1: beta distribution

\[
beta(x; u, v) = \frac{\Gamma(u + v)}{\Gamma(u)\Gamma(v)} x^{u-1}(1-x)^{v-1}, \quad \Gamma(z) = \int_0^\infty t^{z-1}e^{-t} dt
\]

- Bounded support and flexible shape
- Image processing, speech coding, DNA methylation analysis
Example 2: Dirichlet distribution (neutral vector)

\[
\text{Dir}(\mathbf{x}; \mathbf{a}) = \frac{\Gamma\left(\sum_{k=1}^{K} a_k\right)}{\prod_{k=1}^{K} \Gamma(a_k)} \prod_{k=1}^{K} x_k^{a_k - 1}, \quad \sum_{k=1}^{K} x_k = 1, \quad x_k > 0, \quad a_k > 0.
\]

- Conventionally used as conjugate prior of multi categorical distribution or multinomial distribution, describing mixture weights in mixture modeling
- Recently, it was applied to model proportional data (i.e., data with L1 norm)
- Speech coding, skin color detection, multiview 3D enhancement, etc.
Example 3: von Mises-Fisher distribution

\[ f(x; \lambda, \mu) = \frac{\lambda^{\frac{K-1}{2}}}{(2\pi)^{\frac{K}{2}} I_{\frac{K-1}{2}}(\lambda)} e^{\lambda \mu^T x}, \quad x^T x = 1 \]

\( I_p(\nu) \) denotes the modified Bessel function of the first kind

- Distributed on K-dimensional sphere
- Two-dimensional vMF = circle
- Directional statistics, gene expressions, speech coding
Non-Gaussian Statistical Models

• Summary
  – Non-Gaussian distribution represents a family of distributions which are not Gaussian distributed
  – Not conflicting with central limit theorem
  – Well-defined for bounded/semi-bounded/structured data
  – More efficient than Gaussian distribution
  – Hard to estimate, computationally costly, and difficult to use in practice
Outline

Non-Gaussian Statistical Models
- Non-Gaussian vs. Gaussian
- Advantages and Challenges

Variational Inference (VI) and Extended VI
- Formulations and Conditions
- Convergence and Bias

Related Applications
- Beta/Dirichlet Mixture Model
- BG-NMF
Formulation and Conditions

• Maximum likelihood (ML) estimation
  – Widely used for point estimation of the parameters
  – Expectation-maximization (EM) algorithm
  – Converge to local maxima and may yield overfitting
  – No analytically tractable solution for most non-Gaussian distributions
Formulation and Conditions

- Bayesian estimation
  - Estimating the distributions of the parameters, rather than point estimate
  - Conjugate match in exponential family
  - No overfitting, feasible for online learning
  - Without approximation, there is no analytically tractable solution for non-Gaussian distributions
Example: ML estimation for beta mixture model[1]

- M step

\[
\begin{align*}
\psi(u + v) - \psi(u) + \frac{1}{N} \sum_{n=1}^{N} \ln x_n \\
\psi(u + v) - \psi(v) + \frac{1}{N} \sum_{n=1}^{N} \ln(1 - x_n)
\end{align*}
\]

\[= 0\]

\[
\psi(z) = \frac{d \ln \Gamma(z)}{dz} = \int_{0}^{\infty} \left( \frac{e^{-t}}{t} - \frac{e^{-zt}}{1 - e^{-t}} \right) dt
\]

- Numerical solution, Gibbs sampling, Newton-Raphson method, MCMC, etc.

Formulation and Conditions

- Example: Bayesian estimation of beta distribution\[1\]
  - Prior
  \[p(u, v; \alpha_0, \beta_0, \nu_0) \propto \frac{\Gamma(u + v)}{\Gamma(u)\Gamma(v)} e^{-\alpha_0(u-1)} e^{-\beta_0(v-1)}\]
  - Likelihood
  \[\text{beta}(x; u, v) = \frac{\Gamma(u + v)}{\Gamma(u)\Gamma(v)} x^{u-1} (1 - x)^{v-1}\]
  - Posterior
  \[p(u, v | X; \alpha_0, \beta_0, \nu_0) \propto \frac{\Gamma(u + v)}{\Gamma(u)\Gamma(v)} e^{\left(\alpha_0 - \sum_{n=1}^{N} \ln x_n\right)(u-1)} e^{\left(\beta_0 - \sum_{n=1}^{N} \ln(1-x_n)\right)(v-1)}\]

- No closed-form expression for mean, variance, etc.
- No analytically tractable solution for mixture model
- Not applicable in practice

• Variational inference\[1\]
  – Mean field theory in physics, dates back to 18th century, by Euler, Lagrange, etc.
  – Function over function
  – Closed form solution with certain constraints

\[
f(x) = \int f(x | \Theta) f(\Theta) \, d\Theta
\]

\[
\ln f(x) = \int g(\Theta) \ln \frac{f(x, \Theta)}{g(\Theta)} \, d\Theta + \int g(\Theta) \ln \frac{g(\Theta)}{f(\Theta | x)} \, d\Theta
\]

\[
= \mathcal{L}(g) + \text{KL}(g \parallel f)
\]

– **Goal**: approximate \( f(\Theta | x) \) by \( g(\Theta) \) via either maximizing \( \mathcal{L}(g) \) or minimizing \( \text{KL}(g \parallel f) \)

Formulation and Conditions

• Factorized approximation\textsuperscript{[1]}

\[ g(\theta) \approx \prod_i g_i(\theta_i) \]

\[ \ln g_i^*(\theta_i) = E_{j \neq i} [\ln f(x, \theta)] + C \]

– No constraints on the form of \( g_i(\theta_i) \)
– Directly maximizing \( L(g) \)
– Always converges but may fall in local maxima
– Analytically tractable form solution for Gaussian

\textsuperscript{[1]} C. M. Bishop, 'Pattern Recognition and Machine Learning', \textit{Springer}, 2006
Formulation and Conditions

• Extended factorized approximation (EFA)\cite{1,2}

\[
\mathcal{L}(g) = E[f(x, \theta)] - E[g(\theta)] \\
\geq E[\tilde{f}(x, \theta)] - E[g(\theta)]
\]

– Optimal solution: \( \ln g_i^*(\theta_i) = E_{j \neq i} [\ln \tilde{f}(x, \theta)] + C \)

• Strong requirement with larger gap: \( f(x, \theta) \geq \tilde{f}(x, \theta) \)
• Weak requirement with smaller gap: \( E[f(x, \theta)] \geq E[\tilde{f}(x, \theta)] \)

– An efficient way to derive analytically tractable solution for non-Gaussian distribution

– SLB vs MLB \cite{2}

Convergence and Bias

- Multiple lower-bound (MLB) approximation\textsuperscript{[1][2]}
  - Different auxiliary functions for different variable (group)
    \[
    p(X, Z) \geq \tilde{p}_1(X, Z_1) \\
    \vdots \\
    p(X, Z) \geq \tilde{p}_i(X, Z_i) \\
    \vdots \\
    p(X, Z) \geq \tilde{p}_M(X, Z_M)
    \]
  - Optimal solution for each variable (group)
    \[
    \ln \tilde{q}_i^*(Z_i) = E_{Z \backslash Z_i} \left[ \ln \tilde{p}_i(X, Z_i) \right] + \text{const.}
    \]

Convergence and Bias

Update $Z_1$ and $Z_2$ iteratively:

$$\ln \tilde{q}_1^*(Z_1) = E_{Z \backslash Z_1} [\ln \tilde{p}_1(X, Z)] + \text{const}$$
$$\ln \tilde{q}_2^*(Z_2) = E_{Z \backslash Z_2} [\ln \tilde{p}_2(X, Z)] + \text{const}$$

Update $Z_1$ and $Z_2$ iteratively:

$$2 \times \mathcal{L} \geq \hat{\mathcal{L}}_1 + \hat{\mathcal{L}}_2$$
$$= E_Z [\ln \tilde{p}_1(X, Z_1)] - E_Z [\ln q(Z)]$$
$$+ E_Z [\ln \tilde{p}_2(X, Z_2)] - E_Z [\ln q(Z)]$$

Convergence not guaranteed!

Convergence and Bias

- Single lower-bound (SLB) approximation\(^1\)\(^2\)
  - One auxiliary functions for all the different variable (group)

\[ \ln \hat{q}_i^*(Z_i) = \mathbb{E}_{Z \setminus Z_i} [\ln \hat{p}(X, Z)] + \text{const} \]

---


Convergence and Bias

Bias always exists, due to factorized approximations and lower-bound approximation.

Bias will vanish when increasing the amount of training data.

True posterior distribution vs. approximating distribution\[1\]. Dirichlet distribution with $u=[3 \ 5 \ 8]$.

Convergence and Bias

• Summary
  – EVI provides a flexible way to carry out Bayesian estimation of NG statistical model
  – Certain requirements should be fulfilled when implementing EVI
    – MLB vs. SLB
  – Systematic gap
Outline

Non-Gaussian Statistical Models
- Non-Gaussian vs. Gaussian
- Advantages and Challenges

Variational Inference (VI) and Extended VI
- Formulations and Conditions
- Convergence and Bias

Related Applications
- Beta/Dirichlet Mixture Model
- BG-NMF
Dirichlet Mixture Model

Graphical Model of DMM\[^1\]

Dirichlet Mixture Model

- Auxiliary function
  - Step I

\[
\mathbb{E}_Z[\ln f(X, Z)] \geq \sum_{n=1}^{N} \sum_{i=1}^{I} \mathbb{E}[Z_{ni}] \left( \ln \frac{\Gamma(\sum_{k=1}^{K+1} \bar{u}_{ki})}{\prod_{k=1}^{K+1} \Gamma(\bar{u}_{ki})} \right) \\
+ \sum_{k=1}^{K+1} \mathbb{E}_{u_{p \geq k_i}} \left\{ \psi \left( \sum_{m=1}^{k} \bar{u}_{mi} + \sum_{l=k+1}^{K+1} u_{li} \right) - \psi(\bar{u}_{ki}) \right\} \\
\times \bar{u}_{ki} (\ln u_{ki} - \ln \bar{u}_{ki}) \right\} + \mathcal{R}^\dagger
\]

- Step II

\[
\mathbb{E}_{u_{p \geq k_i}} \left\{ \psi \left( \sum_{m=1}^{k} \bar{u}_{mi} + \sum_{l=k+1}^{K+1} u_{li} \right) - \psi(\bar{u}_{ki}) \right\} \bar{u}_{ki} (\ln u_{ki} - \ln \bar{u}_{ki})
\]

\[
= \left\{ \mathbb{E}_{u_{p > k_i}} \left[ \psi \left( \sum_{m=1}^{k} \bar{u}_{mi} + \sum_{l=k+1}^{K+1} u_{li} \right) \right] - \psi(\bar{u}_{ki}) \right\} \bar{u}_{ki} \left( \frac{\mathbb{E}_{u_{k_i}} [\ln u_{k_i}] - \ln \bar{u}_{k_i}}{\leq 0} \right)
\]

\[
\geq \psi \left( \sum_{m=1}^{k} \bar{u}_{mi} + \mathbb{E}_{u_{p > k_i}} \left[ \sum_{l=k+1}^{K+1} u_{li} \right] \right) - \psi(\bar{u}_{ki}) \bar{u}_{k_i} (\mathbb{E}_{u_{k_i}} [\ln u_{k_i}] - \ln \bar{u}_{k_i})
\]

\[
= \psi \left( \sum_{k=1}^{K+1} \bar{u}_{k_i} \right) - \psi(\bar{u}_{k_i}) \bar{u}_{k_i} (\mathbb{E}_{u_{k_i}} [\ln u_{k_i}] - \ln \bar{u}_{k_i})
\]

• Speech coding\(^{[1]}\)
  – Quantization of line spectral frequency (LSF)
  – Well-structured vector
    • all the elements are in \((0, \pi)\)
    • strictly ordered

• Speech coding
  – Solution: Dirichlet mixture model\cite{1,2}
    • transfer LSF vector to ΔLSF vector
    • well-structured: nonnegative elements, L1 norm equals one
    • a neutral vector that can be nonlinearly decorrelated
      (comparable to KLT/PCA for Gaussian source!)

• Speech coding
  – Solution: Dirichlet mixture model\textsuperscript{[1,2]}
    • transfer LSF vector to $\Delta$LSF vector
    • well-structured: nonnegative elements, L1 norm equals one
    • a neutral vector that can be nonlinearly decorrelated
      (comparable to KLT/PCA for Gaussian source!)

\begin{table}
\centering
\begin{tabular}{|c|c|c|c|c|c|c|c|c|c|}
\hline
& \multicolumn{3}{|c|}{Coding performance of DVQ} & \multicolumn{3}{|c|}{Coding performance of BVQ} & \multicolumn{3}{|c|}{Coding performance of UGVQ} \\
\hline
Model & bits/vec & LSD (dB) & LSF outliers (in %) & LSD outliers (in %) & LSD (dB) & LSF outliers (in %) & LSD (dB) & LSF outliers (in %) & LSD (dB) \\
\hline
32 mixture components & 47 & 1.017 & 1.035 & 0.000 & 48 & 1.037 & 0.780 & 0.000 & 48 & 1.009 & 0.935 & 0.000 \\
64 mixture components & 48 & 0.982 & 0.825 & 0.000 & 49 & 0.995 & 0.460 & 0.000 & 49 & 0.793 & 0.700 & 0.000 \\
128 mixture components & 46 & 1.032 & 1.160 & 0.000 & 47 & 0.994 & 0.387 & 0.001 & 47 & 1.008 & 1.090 & 0.002 \\
256 mixture components & 45 & 0.987 & 0.840 & 0.000 & 45 & 0.985 & 1.030 & 0.005 & 48 & 0.956 & 0.700 & 0.003 \\
\hline
\end{tabular}
\end{table}

• PRObabilistic Multiview Depth Enhancement (PROMED)\textsuperscript{[1]}

Free-viewpoint TV

Multiview video imagery

PROMDE Flow Chart

Dirichlet Mixture Model

Two concatenated Newspaper views with approximately superpixels as obtain by using SLIC\textsuperscript{[1]}.

Dirichlet Mixture Model

Dirichlet Mixture Model

Selected regions of synthesized virtual views of test sequences as generated by VSRS 3.5 using MPEG depth maps and enhanced depth maps from our depth enhancement algorithm.

The objective quality of three intermediate virtual views as generated by VSRS 3.5 using the large baseline setting.

Beta Gamma-NMF (BG-NMF)

Graphical Model of BG-NMF\[1\]

Beta Gamma-NMF (BG-NMF)

- Bayesian matrix factorization for bounded support data → BGNMF
- Handle highly sparse matrix → low rank matrix approximation

\[
X_{p,t} \sim \text{Beta}(X_{p,t}; a_{p,t}, b_{p,t}) = \frac{\Gamma(a_{p,t} + b_{p,t})}{\Gamma(a_{p,t})\Gamma(b_{p,t})} X_{p,t}^{a_{p,t}-1} (1 - X_{p,t})^{b_{p,t}-1}
\]

\[
A_{p,k} \sim \text{Gam}(A_{p,k}; \alpha_{p,k}, \mu_{p,k}) \propto A_{p,k}^{\mu_{p,k}-1} e^{-\alpha_{p,k}A_{p,k}}
\]

\[
B_{p,k} \sim \text{Gam}(B_{p,k}; \beta_{p,k}, \nu_{p,k}) \propto B_{p,k}^{\nu_{p,k}-1} e^{-\beta_{p,k}B_{p,k}}
\]

\[
z_{k,t} \sim \text{Gam}(z_{k,t}; \zeta_{k,t}, \rho_{k,t}) \propto z_{k,t}^{\rho_{k,t}-1} e^{-\zeta_{k,t}z_{k,t}}
\]

\[
\hat{X} = \frac{Az}{Az + Bz}
\]

\[ \ln q^*(A_{pk}) = \mathbb{E}_{q(A_{pk})}[\ln p(X, Z)] + \text{const} \]

\[ = \sum_t \mathbb{E}_{q(A_{pk})} \left[ -\ln B \left( \sum_k A_{pk} H_{kt}, \sum_k B_{pk} H_{kt} \right) \right]^{F(A_{pk}, B_{pk}, H_{kt})^2} \]

\[ + \left( \sum_i H_{kt} \ln X_{pt} \right) A_{pk} \]

\[ + (\mu_0 - 1) \ln A_{pk} - \alpha_0 A_{pk} + \text{const}, \]

Objective function. Need to find auxiliary function for the LIB function \( F(A_{pk}, B_{pk}, H_{kt}) \)

Beta Gamma-NMF (BG-NMF)

\[
E_{q(Z)}[F_{pt}] \geq - \ln \mathcal{B}\left(\sum_k \overline{A}_{pk} \overline{H}_{kt}, \sum_k \overline{B}_{pk} \overline{H}_{kt}\right)
\]

\[
+ \left[\psi\left(\sum_k (\overline{A}_{pk} \overline{H}_{kt} + \overline{B}_{pk} \overline{H}_{kt})\right) - \psi\left(\sum_k \overline{A}_{pk} \overline{H}_{kt}\right)\right]
\times \sum_k \overline{A}_{pk} \overline{H}_{kt} \left\{E_{q(A_{p,k})q(H_{k,t})}\left[\ln(\overline{A}_{pk} \overline{H}_{kt})\right] - \ln(\overline{A}_{pk} \overline{H}_{kt})\right\}
\]

\[
+ \left[\psi\left(\sum_k (\overline{A}_{pk} \overline{H}_{kt} + \overline{B}_{pk} \overline{H}_{kt})\right) - \psi\left(\sum_k \overline{B}_{pk} \overline{H}_{kt}\right)\right]
\times \sum_k \overline{B}_{pk} \overline{H}_{kt} \left\{E_{q(B_{p,k})q(H_{k,t})}\left[\ln(\overline{B}_{pk} \overline{H}_{kt})\right] - \ln(\overline{B}_{pk} \overline{H}_{kt})\right\}
\]

Auxiliary function with relative convexity,
Jensen inequality.

• DNA methylation analysis [1]
  – Motivation: using statistical model as a robust analysis tool in bioinformatics area
  – Improve analyzing performance comparing with benchmark methods
  – DNA methylation matrix of 27k × 136
  – Methylation level in [0,1]
  – Preprocessing: feature selection via variance.
    – 27k→5000

Beta Gamma-NMF (BG-NMF)

- PCA + VB-GMM (5000→14)

(a) Data visualization.
(b) Clustering result via PCA+VBGMM.

9 cancers to normal
0 normal to cancer
9 out of 136
Beta Gamma-NMF (BG-NMF)

- BGNMF+ VB-BMM (5000 → 14)

4 cancers to normal
1 normal to cancer
5 out of 136
124 sec. < 139 sec. (RPBMM)
Related Applications

• Summary
  – EVI-based NG statistical model shows advantages in several applications.
  – Fitting data better $\rightarrow$ improved performance
  – Needs a lot of effort to design and derive.
References


Thanks!