NEW TOOLS FOR BAYESIAN INFERENCE: THE VARIATIONAL APPROXIMATION

Nikolaos P. Galatsanos
ECE Dept. Univ. of Patras, Greece.

Dimitris Tzikas
CS Dept. Univ. of Ioannina, Greece.

Introduction

Thomas Bayes (1701-1761), left, first discovered “Bayes’ theorem” in 1764. However, Bayes in his theorem used uniform priors. Pierre-Simon Laplace (1749-1827), right, unaware of Bayes’ work, discovered the same theorem in more general form in a memoir he wrote at the age of 25 and showed its wide applicability.

Applications of Variational Approximation
1. Mixture Modeling of pdfs & Clustering
2. ICA & PCA Analysis
3. Learning MRFs
4. Dynamic Systems Modeling
5. Image Recovery
6. Visual Tracking
7. Digital Communications
8. Acoustics and Speech Processing
9. Learning from Data Bases

Sample of US Patents that Use Variational Inference

- US Patent 6677894: Variational inference vector machine, issued on April 12, 2004
- US Patent 20080025627: Removing camera shake from a single photograph
- US Patent 6612404: Variational inference and learning for segmentation switching-state space models of hidden speech dynamics,
- US Patent 6079803: Data security and intrusion detection,
- US Patent 6184360: AI Player ranking with partial information
- US Patent 6312718: AI Face matching
- US Patent 6088924: AI Detecting humans via their pose

Outline

- Introduction
- Bayesian Inference Basics
- MAP Estimation
- Conjugate Distributions
- Graphical Models
- EM Algorithm
- An Alternative View of the EM
- The Variational EM framework
- Examples
  - Linear Regression
  - Blind Image Deconvolution
  - Image Restoration
  1. Constrained Variational Inference
  2. Bounded Variational Inference
  - Gaussian Mixture Modeling
- Conclusions

(a): # of papers /year in IEEE Journals “EM Algorithm”

(b): # of papers /year in IEEE Journals “Variational Methodology”
Bayesian Inference Basics

• Estimation => Parameter
• Observations: $x$
• Parameters: $\theta$
• Likelihood Function: $p(x; \theta)$ ($\theta$ – parameter)
• Maximum Likelihood Estimation

$$\hat{\theta}_{ML} = \arg \max_{\theta} p(x; \theta)$$

Bayesian Inference Basics

• Inference => $\theta$ Random Variable
• Find Posterior $p(\theta | x)$
• Posterior MORE information than point estimate
  – $E(\theta | x)$ MMSE estimate
  – $Var(\theta | x)$ accuracy of estimates

Bayesian Inference Basics

• "Hidden Variables" $z$:
  – Describe data generation mechanism (graphical model)
  – “links” between observations and parameters
  – Easy to compute $p(x | z)$
  – Introduce priors $p(z; \theta)$

Bayesian Inference Basics

• Find Likelihood, Marginalize Hidden Variables

$$p(x; \theta) = \int p(x, z; \theta) dz = \int p(x | z; \theta) p(z; \theta) dz$$

• Find Posterior

$$p(z | x; \theta) = \frac{p(x | z; \theta) p(z; \theta)}{p(x; \theta)}$$

• In most cases of interest CANNOT Marginalize

Bayesian Inference Basics

• Main effort in Bayesian Inference techniques bypass or approximate marginalization integral.
  – Random sampling methods
    • Monte Carlo
  – Deterministic approximations
    • Laplace
    • Variational

MAP Estimation

• Defined as mode of posterior:

$$\hat{\theta}_{MAP} = \arg \max_{\theta} p(\theta | x)$$

• Based on Bayes’ theorem

$$p(\theta | x) = \frac{p(x | \theta) p(\theta)}{p(x)}$$

• Can be found as:

$$\hat{\theta}_{MAP} = \arg \max_{\theta} p(x | \theta)p(\theta)$$
MAP Estimation
• No need for \( p(x) \) no marginalization
• MAP Estimator
  – Much easier to find
  – Mode of posterior
  – No information about shape of posterior
• Uses Bayes’ Theorem, however posterior not found
• MAP=Poor Man’s Bayesian Inference

Conjugate Priors
• Find prior which allows closed form marginalization of hidden variables
\[
p(x) = \int p(x, z) \, dz = \int p(x \mid z) \, p(z) \, dz
\]

Conjugate Priors
• Example #1: \( \mu \) hidden variable
\[
p(x \mid \mu, \sigma^2) = N(x; \mu, \sigma^2), \quad p(\mu, \mu_0, \sigma_0^2) = N(x; \mu_0, \sigma_0^2)
\]
\[
p(x, \mu_0, \sigma^2, \sigma_0^2) = \int p(x \mid \mu, \sigma^2) \, p(\mu, \mu_0, \sigma_0^2) \, d\mu
\]

Conjugate Priors
• Marginalizing \( \mu \) possible (Gaussian Integral):
\[
p(x, \mu, \sigma^2, \sigma_0^2) = \int p(x \mid \mu, \sigma^2) \, p(\mu, \mu_0, \sigma_0^2) \, d\mu
\]
\[
= N\left(x; \mu_0, \frac{1}{\sigma_0^2 + \frac{1}{\sigma^2}}\right)
\]
• Posterior:
\[
p(\mu \mid x, \sigma^2, \mu_0, \sigma_0^2) = N\left(\mu; \frac{\sigma^2 x + \sigma_0^2 \mu_0}{\sigma^2 + \sigma_0^2}, \frac{1}{\sigma^2 + \frac{1}{\sigma_0^2}}\right)
\]

Conjugate Priors
• Example #2: a “hidden variable”
\[
p(x \mid a) = N\left(x; 0, a^{-1}\right),
\]
\[
p(a; b, c) = \text{Gamma}(a; b, c) = \frac{c^b a^{b-1} \exp(-ca)}{\Gamma(b)}
\]
Conjugate Priors
• w.r.t. to a both \( p(x|a) \) and \( p(a;b,c) \) have the same form (Gamma).
\[
p(x|a) = f(a) \propto a^{1/2} \exp\left(-\frac{ax^2}{2}\right)
\]
\[
p(a;b,c) = g(a) \propto a^{-3/2} \exp(-ac)
\]
• \( p(x|a) \) and \( p(a;b,c) \) conjugate

Conjugate Priors
• Marginalize a possible
\[
p(x;b,c) = \frac{\Gamma(b + 1/2)}{\Gamma(b)} \left( \frac{1}{2\pi} \right)^{1/2} \left( c + \frac{x^2}{2} \right)^{b-1/2}
\]
• Can write as Student’s-t with \( \nu = 2b, \lambda = b/c \)
\[
p(x;\lambda,\nu) = \frac{\Gamma((\nu/2)+1/2)}{\Gamma(\nu/2)} \left( \frac{\lambda}{\nu} \right)^{\nu/2} \left( 1 + \frac{x^2}{\nu} \right)^{-(\nu/2)+1/2}
\]
• Posterior
\[
p(a|x;b,c) = \text{Gamma} \left( a + \frac{1}{2}, c + \frac{x^2}{2} \right)
\]

Conjugate Priors

<table>
<thead>
<tr>
<th>Likelihood</th>
<th>Conjugate Prior Distribution</th>
<th>Posterior Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N(\mu, \Sigma) )</td>
<td>( N(\mu, \Sigma) )</td>
<td>( N(\mu, \Sigma + \sum_{i=1}^{s} (x_i - \mu)^2) )</td>
</tr>
<tr>
<td>( N(\mu, \sigma^2) )</td>
<td>( \text{Gamma}(\sigma^2, a, b) )</td>
<td>( \text{Gamma}(\sigma^2/a + n/2, b + \sum_{i=1}^{s} (x_i - \mu)^2/2) )</td>
</tr>
<tr>
<td>( \text{Wishart}(\Sigma, \nu) )</td>
<td>( \text{Wishart}(\Sigma, \nu + n, \nu + \sum_{i=1}^{s} (X_i - \mu)^T (X_i - \mu)) )</td>
<td></td>
</tr>
<tr>
<td>( \text{Dir}(\pi</td>
<td>\alpha) )</td>
<td>( \text{Dir}(\pi</td>
</tr>
</tbody>
</table>

Graphical Models
• Represent dependencies between rv.s in a statistical model
• Graph nodes represent rv.s and edges dependencies
• Directed and undirected graphs
• Undirected Markov Random Fields
• Rest of presentation: directed, no cycles graphs

Graphical Models

- \( x_s \) rv associated node \( s \), \( \pi(s) \) parents of \( s \)
- \( p(x_s|x_{s\pi(s)}) \) conditional pdf
- Joint pdf over all variables
\[
p(x) = \prod_s p(x_s|x_{s\pi(s)})
\]

Graphical Models

- Example:
\[
p(a,b,c,d;\theta) = p(a;\theta_1) p(b|a;\theta_2) p(c|a;\theta_3) p(d|b,c;\theta_4)
\]
**EM Algorithm**

- \( x \)- observations, \( z \) - “hidden variables”, \( \theta \)-parameters.
- Define:
  \[
  Q(\theta, \theta^{old}) = E\left[ \ln p(x, z; \theta) \right]_{p(z|\theta^{old})} \\
  = \int \ln p(x, z; \theta) p(z|x; \theta^{old}) dz
  \]

**An Alternative View of the EM Algorithm**

- Can write:
  \[
  \ln p(x; \theta) = F(q, \theta) + KL(q \| p)
  \]
  \[
  F(q, \theta) = \int q(z) \ln \left( \frac{p(x, z; \theta)}{q(z)} \right) dz
  \]
  \[
  KL(q \| p) = -\int q(z) \ln \left( \frac{p(x, z; \theta)}{q(z)} \right) dz
  \]
- \( q(z) \) any pdf,
- \( KL(q \| p) \) Kullback-Leibler Divergence

**The Variational EM framework**

- Assume \( p(z|x; \theta) \) unknown
- \( F(q, \theta) \) functional in terms of \( q(z) \)
- Variational EM
  - Variational E-step: \( q^{NEW}(z) = \max_{q} F(q, \theta^{OLD}) \)
  - Variational M-step: \( \theta^{NEW} = \arg\max_{\theta} F(q^{NEW}, \theta) \)
The Variational EM framework

• Key issue: maximize $F(q, \theta)$ w.r.t. $q(z)$?
• Assume parametric form for $q(z)$
• $q(z)$ approximates unknown posterior $p(z|x)$

\[ \ln p(x; \theta) = F(q, \theta) + KL(q||p) \]

• $\max F(q, \theta) \Rightarrow \min KL(q||p)$

Mean Field Approximation

• Assumption: $q(z)$ factorizes

\[ q(z) = \prod_{i=1}^{M} q_i(z_i) \]

• “Mean Field” approximation, statistical physics

Mean Field Approximation

• Then optimal factor $q_j(z_j)$ is:

\[ q_j^*(z_j) = \exp\left(\frac{\langle \ln p(x, z; \theta) \rangle_{z_j}}{\int \exp \langle \ln p(x, z; \theta) \rangle_{z_j} dz_j}\right) \]

• with

\[ \langle \ln p(x, z; \theta) \rangle_{z_j} = \int \ln p(x, z; \theta) \prod_{i \neq j} q_i dz_i \]

Conjugate-Exponential models

• Prior distributions belong to the exponential family

\[ p(X|Y) = \exp\left[ \phi(Y)^T u(X) + f(X) + g(Y) \right] \]

• Graphical model with conjugate priors at each level
  – hidden $z$, parents $\pi(z)$
  – $p(z|\pi(z))$ conjugate to $p(\pi(z)|\pi(\pi(z)))$

\[ p(z|\pi(\pi(z))) = \int p(z|\pi(z)) \, p(\pi(z)|\pi(\pi(z))) \, d\pi(z) \]

Conjugate-Exponential models

• Tractable Variational computations

\[ q(z) \propto \exp\left(\langle \ln p(x, z; \theta) \rangle_{z_j}\right) \]

• $\ln q(z) = \ln p(x|z) + \langle \ln p(z_i|z_j; \theta) \rangle_{q(z_j)} + \text{const}$

• $\ln q(z) = \langle \ln p(z_i|z_j; \theta) \rangle_{q(z_j)} + \ln p(z_j; \theta) + \text{const}$
Examples
Linear Regression

Linear Regression
• Observations at $t_n$ find $y(x)$
  $$t_n = y(x_n) + \epsilon_n, \ n = 1, \ldots, N$$

Linear Regression
• Signal $y(x)$ modeled by
  $$y(x; \mathbf{w}) = \sum_{n=1}^N w_n \phi_n(x)$$

• Observation model
  $$t_n = y(x_n; \mathbf{w}) + \epsilon_n, \ n = 1, \ldots, N$$
  $$\mathbf{t} = \mathbf{\Phi} \mathbf{w} + \mathbf{\epsilon}$$

• $\mathbf{\Phi}_{NM}$ design matrix
  $$\mathbf{\Phi} = (\phi_1(x_1), \ldots, \phi_N(x_N))^T$$

Linear Regression: Least Squares (w-parameters)
• Graphical Model
  $$\begin{pmatrix} t \end{pmatrix}_{\times N} \begin{pmatrix} \mathbf{\Phi} \end{pmatrix}_{\times N} \begin{pmatrix} \mathbf{\beta} \end{pmatrix}_{\times 1} \begin{pmatrix} \mathbf{w} \end{pmatrix}_{\times N}$$

• Maximum Likelihood Estimation
  $$\mathbf{w}_{\text{ML}} = \arg \max \ p(\mathbf{t}; \mathbf{w}, \mathbf{\beta}) = \arg \max \ N(\mathbf{t} | \mathbf{\Phi} \mathbf{w}, \mathbf{\beta}^{-1} \mathbf{I})$$

Linear Regression: Parameter Estimation
• Minimize mean square error
  $$E_{ss}(\mathbf{w}) = \mathbf{t} - \mathbf{\Phi} \mathbf{w} \| = \sum_{n=1}^N \| t_n - y(x_n; \mathbf{w}) \|^2$$

• Solution
  $$\mathbf{w}_{LS} = (\mathbf{\Phi}^T \mathbf{\Phi})^{-1} \mathbf{\Phi}^T \mathbf{t} \quad \mathbf{\beta}^{-1} = \frac{1}{N} \mathbf{I}$$
Linear Regression: Problems with Least Squares ($w$-parameters)

- Maximum Likelihood Limitations:
  - $N$ observations $M$ parameters to estimate, $N>M$
  - Otherwise $[\Phi^\top \Phi]^{-1}$ ill-conditioned
  - ML estimates large variance
- Assume constraints on the parameters $w$
  - Bayesian: Use prior distribution $p(w)$

Bayesian Linear Regression: Model-2 ($w$-Gaussian iid distributed)

- Gaussian weight prior
  
  \[
  p(w; \alpha) = \prod_{m=1}^{M} N(w_m | 0, \alpha^{-1})
  \]
- Why Gaussian?
  - Conjugate to likelihood

Bayesian Linear Regression: Inference

- Posterior given by Bayes’s law
  \[
  p(w | t; \alpha, \beta) = \frac{p(t | w; \beta)p(w; \alpha)}{p(t; \alpha, \beta)}
  \]
- Can be found in closed form
  \[
  p(w | t; \alpha, \beta) = N(w | \mu, \Sigma)
  \]
  \[
  \mu = \beta \Sigma \Phi^\top t
  \]
  \[
  \Sigma = (\beta \Phi^\top \Phi + \alpha I)^{-1}
  \]

Bayesian Linear Regression: Parameter Estimation

- Maximum Likelihood
  \[
  (\alpha_{ML}, \beta_{ML}) = \arg \min_{\alpha, \beta} \int p(t | w; \beta)p(w; \alpha) \, d\alpha
  \]
  \[
  = \arg \min_{\alpha, \beta} \left\{ \log \left( \beta^{-1} + \alpha^{-1} \Phi \Phi^\top \right) + t^\top \left( \beta^{-1} + \alpha^{-1} \Phi \Phi^\top \right)^{-1} t \right\}
  \]
  - Not straight forward constrained optimization because $\alpha>0$, $\beta>0$.
  - Resort to EM algorithm!

Bayesian Linear Regression: EM Algorithm

- Observations $t$, parameters $\alpha$, $\beta$; hidden variables $w$.
- Key for the application of EM: $p(w|t)$ *explicitly known*
- E-step: $p(w|t; \alpha^0, \beta^0)$ obtained, inference of hidden variables
- M-step: ML estimates of parameters

Bayesian Linear Regression: EM algorithm

- $Q^{(t)}(t, w; \alpha, \beta) = \langle \ln p(t | w; \beta)p(w; \alpha) \rangle_{p(w; \alpha^0, \beta^0)}$
  \[
  = \langle \ln p(t | w; \beta)p(w; \alpha) \rangle_{p(w; \alpha^0, \beta^0)}
  \]
  \[
  = \frac{N}{2} \ln \beta - \frac{\beta}{2} (t - \Phi \mu)^\top (t - \Phi \mu) + \frac{M}{2} \ln \alpha - \frac{\alpha}{2} (\mu - \Phi \Phi^\top) + \text{const}
  \]
  \[
  = \frac{N}{2} \ln \beta - \frac{\beta}{2} (t - \Phi \mu)^\top (t - \Phi \mu) + \frac{M}{2} \ln \alpha - \frac{\alpha}{2} (\mu - \Phi \Phi^\top) + \text{const}
  \]
  \[
  = \frac{N}{2} \ln \beta - \frac{\beta}{2} \left( \Phi \Phi^\top \right) + \frac{M}{2} \ln \alpha - \frac{\alpha}{2} (\Phi \Phi^\top) + \text{const}
  \]
  \[
  + \frac{M}{2} \ln \alpha - \frac{\alpha}{2} (\Phi \Phi^\top) + \text{const}
  \]
  \[
  + \frac{M}{2} \ln \alpha - \frac{\alpha}{2} (\Phi \Phi^\top) + \text{const}
  \]
Linear Regression: EM algorithm, E-step

- E-step: Evaluate $p(w|t; \alpha^{(0)}, \beta^{(0)}) = N(\mu^{(0)}, \Sigma^{(0)})$
  
  $\Sigma^{(t)} = (\beta^T \Phi \Phi + \alpha^{(t)} I)^{-1}$
  
  $\mu^{(t)} = \beta^T \Sigma^{(t)} \Phi^T t$

Bayesian Linear Regression: Parameter Estimation: M-step

- M-step
  
  $\alpha^{(t+1)}, \beta^{(t+1)} = \arg\max_{\alpha, \beta} Q^{(t)}(t, w; \alpha, \beta)$
  
  $\mathcal{Q}^{(t)}(t, w; \alpha, \beta) = \frac{M}{2\alpha} \left( ||w||^2 + \text{tr}(\Sigma_t) \right)$
  
  $\mathcal{Q}^{(t)}(t, w; \alpha, \beta) = \frac{N}{2\beta} \left( ||\Phi w||^2 + \text{tr}(\Phi^T \Sigma_t \Phi) \right)$

  $\alpha^{(t+1)} = \frac{M}{\|w\|^2 + \text{tr}(\Sigma_t)}$
  
  $\beta^{(t+1)} = \frac{N}{\|\Phi w\|^2 + \text{tr}(\Phi^T \Sigma_t \Phi)}$

Sparse Bayesian Linear Regression

- Limited model: $w_i$ iid
- How to select basis functions?
- Sparse Linear Model
  
  - Consider many basis functions
  
  - Estimations use only few basis functions
- Advantages
  
  - Small variance (good generalization)
  
  - Fast Evaluation of estimation

Sparse Bayesian Linear Regression: Prior Distribution

- "True" weight prior is Student's-t

  $p(w; a, b) = \int p(w | a, b) p(a, b) da$

  $= \int \prod_{n=1}^{N} N(w_n | 0, \alpha_n^{-1}) Gamma(\alpha_n | a, b) da$

  $+ \int \prod_{n=1}^{N} \text{St}(w_n | \lambda, \nu) da$

Sparse Bayesian Linear Regression: Prior Distribution

- "True" weight prior is Student's-t

  $p(w; a, b) = \int p(w | a, b) p(a, b) da$

  $= \int \prod_{n=1}^{N} N(w_n | 0, \alpha_n^{-1}) Gamma(\alpha_n | a, b) da$

  $+ \int \prod_{n=1}^{N} \text{St}(w_n | \lambda, \nu) da$
Sparse Bayesian Linear Regression: Variational Inference

- Posterior:
  \[ p(w, a, \beta | t) = \frac{p(t | w, \beta) p(w | a) p(\alpha) p(\beta)}{p(t)} \]

- Cannot compute
  \[ p(t) = \int p(t | w, \beta) p(w | a) p(\alpha) p(\beta) \, dw \, da \, d\beta \]

- Variational Mean Field Approximation
  \[ p(w, a, \beta | t) = q(w, a, \beta) = q(\alpha) q(\beta) q(\omega) \]

Sparse Bayesian Linear Regression: Variational Inference

**Finding the required expectations**

- \( q(\beta) = \Gamma(\beta | \tilde{c}, \tilde{a}) \), \( \langle \beta \rangle = \tilde{c} / \tilde{a} \)
- \( q(\alpha_n) = \Gamma(\alpha_n | \tilde{\alpha}_n, \tilde{\beta}_n) \), \( \langle \alpha_n \rangle = \tilde{\alpha}_n / \tilde{\beta}_n \)
- \( q(w) = N(w | \mu, \Sigma) \), \( \langle w \rangle = \mu + \Sigma \mu \)
- \( \langle \|\Phi \omega\|_F^2 \rangle = \|\tilde{\mu}^2 + \nu \langle \Phi \Sigma \Phi \rangle + \mu \langle \Phi \mu \rangle = 2 \|\Phi \mu\|^2 \)

Sparse Bayesian Linear Regression: Parameter Estimation

- Parameters \( a, b, c, d \)?
  - Use fixed values that define **uninformative** priors
  - Estimate parameters in Variational M-step
- For fixed \( a, b, c, d \) iterates only between VE-step for \( q(\alpha), q(\beta) \) and \( q(w) \)
Linear Regression Example

Examples Blind Image Deconvolution

Blind Image Deconvolution

- Unknown quantities \((f, h)\) twice than known \((g)\)
- Properties that model should impose:
  - PSF: smooth, limited support
  - Image: smooth, preserve edges
  - Noise: robustness

Blind Image Deconvolution: Noise Model

- Robust noise model
  - Student’s t pdf

\[
p(n | \beta) = \prod_{i=1}^{N} N(n_i | 0, \beta_i) \quad p(\beta) = \prod_{i=1}^{N} \text{Gamma}(\beta_i | a^{\beta_i}, b^{\beta_i})
\]

\[
p(n) = \int p(n | \beta) p(\beta) d\beta = \text{Student's t}
\]
Blind Image Deconvolution: PSF Model

- **PSF**: Sparse Linear model
- **Basis functions** are Gaussian kernels
  \[
  h(x) = \sum_j w_j \phi_j(x) \quad \text{and} \quad h = \Phi w
  \]
- **Sparseness weight prior**
  \[
  p(w | \alpha) = \prod_i N(w_i | 0, \alpha^{-1}), \quad p(\alpha) = \text{Gamma}(\alpha, a^*, b^*)
  \]
  \[
  p(w) = \int p(w | \alpha)p(\alpha) d\alpha = \text{Student's t}
  \]

Blind Image Deconvolution: Image Model

- **Directional image differences**
  \[
  \epsilon^{x+1}(x, y) = f(x, y) - f(x, y+1)
  \]
  \[
  \epsilon^{x}(x, y) = f(x, y) - f(x+1, y)
  \]
  \[
  p(\epsilon' | \gamma') = \prod_i N(\epsilon'_i | 0, \gamma'^{-1}_{\epsilon'}), \quad p(\gamma') = \Gamma\gamma^*, \quad p(\epsilon') = N(0, \Phi' \Phi')
  \]

Blind Image Deconvolution: Graphical Model

- **Joint pdf**
  \[
  q(f, w, \alpha, \beta, \gamma) = q(f)q(w)q(\alpha)q(\beta)q(\gamma)
  \]
  \[
  \log q(f, w, \alpha, \beta, \gamma) = \log q(f) + \log q(w) + \log q(\alpha) + \log q(\beta) + \log q(\gamma)
  \]

Blind Image Deconvolution: Variational Bound

- **Mean field approximation**
  \[
  q(f, w, \alpha, \beta, \gamma) = q(f)q(w)q(\alpha)q(\beta)q(\gamma)
  \]
  \[
  \log q(f, w, \alpha, \beta, \gamma) = \log q(f) + \log q(w) + \log q(\alpha) + \log q(\beta) + \log q(\gamma)
  \]

- **Maximization of Variational Bound results in**
  \[
  log q(f) = q(f) p(f | w, \alpha, \beta, \gamma) + \text{const}
  \]
  \[
  log q(w) = q(w) p(w | f, \alpha, \beta, \gamma) + \text{const}
  \]
  \[
  log q(\alpha) = q(\alpha) p(\alpha | f, w, \beta, \gamma) + \text{const}
  \]
  \[
  log q(\beta) = q(\beta) p(\beta | f, w, \alpha, \gamma) + \text{const}
  \]
  \[
  log q(\gamma) = q(\gamma) p(\gamma | f, w, \alpha, \beta) + \text{const}
  \]

Blind Image Deconvolution: Find approximate posterior \( q(f) \)

- **Mean Field Optimization**
  \[
  \log q(f) = \log N(f | \mu, \Sigma_f) + \text{const}
  \]
  \[
  \mu_f = \Phi^T w + \mu_f
  \]
  \[
  \Sigma_f = \text{cov}(f)
  \]

- **Completing Square results in**
  \[
  q(f) = N(f | \mu_f, \Sigma_f)
  \]
  \[
  p_\mu = \Sigma_f^T \Phi^T + \Sigma_f^T \Phi^T + \Phi^T \Phi^T g + \text{const}
  \]
  \[
  p_\Sigma = \Phi^T \Phi^T + \Phi^T \Phi^T + \text{const}
  \]
  \[
  p_\Sigma = \Phi^T \Phi^T + \Phi^T \Phi^T + \text{const}
  \]

Blind Image Deconvolution: Approximate posterior \( q(w) \)

- **Similar calculations**
  \[
  \log q(w) = \log N(w | \mu, \Sigma) + \text{const}
  \]
  \[
  p_\mu = \Sigma_f^T \Phi^T + \Phi^T \Phi^T + \Phi^T \Phi^T g + \text{const}
  \]
  \[
  p_\Sigma = \Phi^T \Phi^T + \Phi^T \Phi^T + \text{const}
  \]
  \[
  p_\Sigma = \Phi^T \Phi^T + \Phi^T \Phi^T + \text{const}
  \]
  \[
  q(w) = N(w | \mu, \Sigma)
  \]
Blind Image Deconvolution: Approximate posterior $q(\alpha)$

Based on the mean field approximation

$$\log q(\alpha) = \log p(w, \alpha) + \log p(\alpha)_{\text{prior}} + \text{const}$$

$$= \frac{1}{2} \sum_i \log \alpha_i - \frac{1}{2} \sum_i \alpha_i n_i^2 + (\alpha^0 - 1) \sum_i \log \alpha_i - b^2 \sum_i \alpha_i + \text{const}$$

$$= \frac{1}{2} (\alpha - 1 + \frac{1}{2}) \log \alpha - \frac{1}{2} (b + \gamma \alpha) d_{\text{prior}} + \text{const}$$

Which implies

$$q(\alpha) = \prod_i \text{Gamma}(\alpha_i, \tilde{a}^\alpha, \tilde{b}^\alpha) \quad \tilde{a}^\alpha = \alpha^0 + 1/2$$

$$\tilde{b}^\alpha = b + \frac{1}{2} < n_i^2 >$$

Blind Image Deconvolution: Approximate posteriors $q(\beta)$ and $q(\gamma)$

Similarly we get

$$q(\beta) = \prod_i \text{Gamma}(\beta_i, \tilde{a}^\beta, \tilde{b}^\beta) \quad \tilde{a}^\beta = \alpha^0 + 1/2$$

$$\tilde{b}^\beta = b + \frac{1}{2} < n_i^2 >$$

$$\text{nnl} = \Phi g - W$$

Blind Image Deconvolution: Statistics of Approximate Posteriors

$$\langle w_i \rangle = \mu_{w_i},$$

$$\langle w_i^2 \rangle = \mu_{w_i}^2 + \Sigma_{w_i}$$

$$\langle f \rangle = \mu_{f},$$

$$\langle f^2 \rangle = \mu_{f}^2 + \Sigma_{f}.$$ 

$$\langle \alpha \rangle = \tilde{a}^\alpha / \tilde{b}^\alpha$$

$$\langle \beta \rangle = \tilde{a}^\beta / \tilde{b}^\beta$$

$$\langle \gamma \rangle = \frac{1}{\gamma} \tilde{b}^\gamma$$

$$\langle \text{nnl}^2 \rangle = \langle w_i^2 \rangle - 2 \Phi g \Phi g^T + \Phi (\mu_{\text{mean}} \Phi)^T \Phi.$$

Blind Image Deconvolution: Overall V-EM Algorithm

- Fix parameters $\{a^*, b^*, a^0, b^0, a^*, b^*\}$ to yield uninformative hyperpriors (no Variational M-step).

- Iterate between estimates of statistics of $q(f)$, $q(w)$, $q(\alpha)$, $q(\beta)$ and $q(\gamma)$ (only Variational E-step).

Blind Image Deconvolution: Example

- PSF: Square 7x7, 40db noise
Blind Image Deconvolution: Examples

Example: Image Restoration (Constrained Variational Inference*)


Image Restoration

Problem Definition
- Imaging Model (N-pixels)
  \[ g = h \ast f + n = Hf + n \]
  - \( g \in \mathbb{R}^{N \times 1} \): Degraded (Observations)
  - \( h \in \mathbb{R}^{N \times 1} \), \( H \in \mathbb{R}^{N \times N} \): Point Spread Function (known)
  - \( f \in \mathbb{R}^{N \times 1} \): Original Image (unknown)
  - \( n \sim \mathcal{N}(0, \beta^{-1} I) \): Noise (unknown)

Image Restoration: \( f \) Parameter

- Likelihood of Observations
  \[ p(g; f, \beta) = \mathcal{N}(Hf, \beta^{-1} I) \]
  \[ \hat{f} = \arg \max \{ p(g; f, \beta) \} = \arg \min \| g - Hf \| \]
  - Graphical Model
  \[ \hat{f} = (H' H)^{-1} H' g \]
- Problematic too many parameters

Image Restoration: \( f \) Parameter - Example

\[ \text{SNR} = 20 \log \frac{\| Hf \|}{\| n \|} \]

Double circled observed \( g \)
Image Restoration: Bayesian Inference (f: hidden r.v.),

- **Error from Local Linear Predictor**
  \[ f(i) - \frac{1}{2}(f(i-1) + f(i+1)) = \epsilon(i) \]

- **Assume Gaussian i.i.d. Prediction Errors**
  \[ p(\epsilon(i)) = N(0, \alpha^{-1}) \quad p(\epsilon) = \prod_i p(\epsilon(i)) \]

Image Restoration: Bayesian Inference (f: hidden r.v.)

- **Observation Likelihood** *(computed analytically)*
  \[ p(g; \alpha, \beta) = \int p(g | f) p(f; \alpha, \beta) df \]

- **Posterior of Hidden** *(computed analytically)*
  \[ p(f | g; \alpha, \beta) = N(\mu_f, \Sigma_f) \]

- **Bayesian Inference via EM**

Image Restoration: Spatially-Varying Bayesian Model

- **Spatially Varying**
  \[ p(\epsilon(i)) = N(0, \alpha^{-1}) \]

- **α**, “hidden variables” Bayesian Inference

- **Conjugate pdf**
  \[ p(\alpha_i) = \text{Gamma}(\alpha, \alpha, \beta) \]

Image Restoration: Spatially-Varying, Bayesian Inference

- **Use many** \( \nu_i = Q_i f, \quad i = 1, 2, \ldots, P \) in prior

- **“Product” Prior**
  \[ p(f | \nu) = \frac{1}{Z(\nu)} \prod_i p(f | \nu_i), \quad p(f | \nu_i) \sim \exp \left( -\frac{1}{2} f^T Q_i A_i Q_i f \right) \]

- **Prior:** Enforces many properties **simultaneously**
Image Restoration: Spatially Varying Bayesian Inference

• Change Observations Domain
  \[ Q_{g} = Q_{g}Hf + Q_{n}, \]
  \[ y_{k} = H\varepsilon_{k} + n_{k}, \]  \[ k = 1, \ldots, P \]

• Hidden Variables:
  \[ \varepsilon = [\varepsilon_{1}, \varepsilon_{2}, \ldots, \varepsilon_{N}], \alpha = [\alpha_{1}, \alpha_{2}, \ldots, \alpha_{N}] \]

• Prior on \( \varepsilon \) **NO difficulty** normalizing
  \[ p(\varepsilon | \alpha) = \prod_{k=1}^{P} p(\varepsilon_{k} | \alpha_{k}(i)) \]

• Cannot marginalize hidden variables
• Resort to Variational Methodology

Image Restoration: Spatially Varying, Bayesian Inference

• Difficulty: Compute normalization of prior
  \[ Z(\mathbf{a}) \propto \det \left( \sum_{k=1}^{P} Q_{k}^{T}QA_{k} \right) \]
  \[ Q_{k}A_{k} (N \times N), N = 10^{5} - 10^{6} \]

Image Restoration: Spatially Varying, Bayesian Inference

• "Posteriors" Mean Field Approximation
  \[ q(\varepsilon_{k}, \alpha_{k}) = q(\varepsilon_{k})q(\alpha_{k}), \]  \[ k = 1, \ldots, P \]

• Maximize Var. Bound
  \[ q(\varepsilon_{k}) = N(\mu_{k}, \Sigma_{k}), \]
  \[ \mu_{k} = \beta H^{T}g, \]
  \[ \Sigma_{k} = (\beta H^{T} + \lambda Q_{k}^{T}QA_{k})^{-1} \]

• New problem: Different \( f \) for each \( \mu_{k} \)
Image Restoration: Constrained Variational Inference

• Define “Constrained Posterior”

\[ q(\varepsilon_i) = \mathcal{N}(Q_m, Q, RQ) \]

• Consistent with

\[ \varepsilon_i = Q_f, \quad m = E(t), \quad R = E[(f - m)^2 (t - m)] \]

• Maximize the Variational Bound w.r.t. \( m \) and \( R \)

Image Restoration: Constrained Variational Inference

- VE-step:

\[ \hat{q}^{(i+1)}(\hat{\mu}), \hat{d}^{(i+1)}(\hat{\theta}) = \arg \max_{q(\mu), \theta} \{ F(q(\mu), \theta, \hat{d}^{(i+1)}) \} \]

- VM-step:

\[ \hat{d}_i^{(i+1)} = \arg \max_{\theta} \{ F(q^{(i)}(\mu), \hat{d}^{(i+1)}, \theta) \} \]

\[ \hat{\theta}_i = [R, m], \quad \hat{\theta}_j = [\theta, \lambda, \ldots, \lambda, \nu_1, \ldots, \nu_j]^T \]

Image Restoration: Constrained Variational Inference

- VE-Step

\[ F(q, \theta) = \int \int q(\varepsilon_i; \theta) q(\alpha) \log p(y | \varepsilon_i; \theta) p(\varepsilon_i | \alpha; \theta) d\varepsilon_i d\alpha - \int \int q(\varepsilon_i; \theta) q(\alpha) \log \int q(\varepsilon_i; \theta) q(\alpha) f(\varepsilon_i; \alpha) d\varepsilon_i d\alpha, \]

\[ F = F(\theta) = \sum \int q(\varepsilon_i; \theta) q(\alpha) \log p(y | \varepsilon_i; \theta) p(\varepsilon_i | \alpha; \theta) d\varepsilon_i d\alpha - \sum \int q(\varepsilon_i; \theta) \log q(\varepsilon_i; \theta) d\varepsilon_i. \]

Image Restoration: Constrained Variational Inference

- VM-Step

\[ \delta F(q) \to 0 \Rightarrow \delta \text{trace} \left[ \beta \text{PH}^H R + \sum \lambda_i Q_i^T \Lambda_i Q_i R \right] - 2 \log \det[R] - 0 \]

\[ \Rightarrow \beta \text{PH}^H + \sum \lambda_i Q_i^T \Lambda_i Q_i - P^H \gamma - 0 \Rightarrow R = \left( \beta \text{PH}^H + \sum \lambda_i Q_i^T \Lambda_i Q_i \right)^{-1}. \]

\[ \delta F(\theta) \to 0 \Rightarrow m = \beta \text{RH}^H g \]

Image Restoration: Constrained Variational Inference

\[ q(\mu) = \exp \{ \log p(\hat{y}, \hat{z}, \hat{\alpha}) \}_{\theta = \theta} \times \prod \int q(\varepsilon_i; \theta) \\exp \left\{ -\frac{1}{2} \lambda_i \left( \left( m_i(i) \right)^2 + C_i(i) \right) \right\} \]

\[ q(\varepsilon_i; \theta) = \text{Gamma} \left[ \varepsilon_i; \theta, \frac{1}{2} \lambda_i \left( \left( m_i(i) \right)^2 + C_i(i) \right) \right] \]

\[ m_i = Q_m, \quad C_i = Q R Q_i^T \]
Image Restoration: Constrained Variational Inference

1. Initialize: \( m \) stationary model.
2. Repeat until convergence:
   - **VE-step:**
     - Update: \( m \) and \( R \), calculate \( m_k \) and \( C_k \).
     - Calculate expected value w.r.t. \( q(\tilde{a}_k(i)) \), needed for VM-step and the next VE-step.
   - **VM-step:**
     - Update: \( \beta, \lambda, \nu, k=1,2,...,P \)
3. Use \( m \) as restored image estimate.

---

Example: Image Restoration (Bounded Variational Inference*)


---

Image Restoration: Bounded Variational Inference

- **Imaging Model (N-pixels)**
  \[ g = h \ast f + n = Hf + n \]
- **Observations Likelihood**
  \[ p(g | f, \beta) \sim N(Hf, \beta^{-1}I) \]
Image Restoration: Bounded Variational Inference

Graphical Model

\( u, b \)

\( w \)

\( r \)

\( p \)

\( g \)

Image Restoration: Bounded Variational Inference

• Difficulty in VE-step

\[ \log q(\alpha) = \left\{ \log p(g, f; \alpha, \beta, a, b) \right\}_{\alpha, \beta} + \frac{1}{2} \beta \| f - g \|_2^2 - \alpha \sum_i \left( (q(\alpha, \cdot))_i + ((q(\alpha, \cdot))_i^2) \right) + \text{const} \]

• Due to \( \sqrt{\alpha} \) cannot Compute Expectation

Image Restoration: Bounded Variational Inference

• Bypass difficulty: Maximize a Lower Bound of Variational Bound

• Use Upper Bound: \( f(w) = \sqrt{w} \leq \frac{u + u^*}{2\sqrt{u}} = g(u, w), \forall u > 0 \)

• Bound gets “tight”: \( f(w) = g(u^*, w), u^* = w \)

Image Restoration: Bounded Variational Inference

• Define Function

\[ p(f|\alpha) \geq M(f, \alpha, a) = \alpha \exp \left( -\frac{1}{2} \sum_i \left( (q(\alpha, \cdot))_i + ((q(\alpha, \cdot))_i^2) \right) + u \right) \]

\[ p(g, f; \alpha, \beta) \geq p(f|\alpha) M(f, \alpha, a) p(\alpha) = M(g, f, \alpha, \beta) \]

• Lower Bound of Variational Bound

\[ F(q, \beta) \geq \int q(f, \alpha) \ln \frac{M(g, f, \alpha, \beta) d\alpha d\beta}{q(f, \alpha)} = F^*(q, \beta) \]

Image Restoration: Bounded Variational Inference

• “Tightest” Bound

\[ f(w) = g(u^*, w), u^* = w \Rightarrow \]

\[ u^{(k)} = \left( \left( Q_{\alpha} \right)^i + \left( Q_{\alpha} \right)^i \right)_{i=1}^{\alpha} \]

\[ = \left[ Q_{\alpha} \left( f_{\alpha} \right) M \right] + \left[ Q_{\alpha} \left( f_{\alpha} \right) M \right] + \left[ Q_{\alpha} \left( f_{\alpha} \right) M \right] + \left[ Q_{\alpha} \left( f_{\alpha} \right) M \right] \]

• \( u \), Captures Local Spatial Activity
Example: Gaussian Mixture Models

Gaussian Mixture Models:

- Model any pdf
- Soft clustering

Parameters

\[ \theta = \{ \pi_j, \mu_j, \Sigma_j \}_{j=1}^{M} \]

Maximum Likelihood Estimation Difficult

\[ \theta_{\text{ML}} = \arg \max_\theta \sum_{i=1}^{N} \log \pi_j N(x_i; \mu_j, \Sigma_j) \]

Gaussian Mixture Models: Data Generation Mechanism

- Introduce binary hidden variable \( z \)
  1. Select component \( z_j = 1, z = (0, \ldots, 0, 1, 0, \ldots, 0) \)
  2. Generate sample from selected component \( x \sim p(x) = N(x; \mu_j, \Sigma_j) \)

Gaussian Mixture Models: Posterior

- Posterior (responsibility) can be computed analytically

\[
p(z_j = 1 | x) = \frac{p(x | z_j = 1) p(z_j = 1)}{p(x)} = \frac{\pi_j N(x; \mu_j, \Sigma_j)}{\sum_{j=1}^{M} \pi_j N(x; \mu_j, \Sigma_j)}
\]

\[
p(z | x; \theta) = \prod_{j=1}^{M} \left[ \pi_j N(x; \mu_j, \Sigma_j) \right] / \left[ \sum_{j=1}^{M} \pi_j N(x; \mu_j, \Sigma_j) \right]
\]

Gaussian Mixture Models: Parameter Estimation

- Maximum likelihood

\[ \theta_{\text{ML}} = \arg \max_\theta \log p(X; \theta) = \arg \max_\theta \sum_{i=1}^{N} \log \pi_j N(x_i; \mu_j, \Sigma_j) \]

- Use EM
  - Simplifies optimization
  - Proved convergence
  - Satisfies positivity constraints

\[ \pi_j > 0, \sum_{j=1}^{M} \pi_j = 1 \]
Gaussian Mixture Models: Parameter Estimation

EM

• E-step

\[ Q(\theta^{(t)}) = \log p(X, Z | \theta^{(t)}) \]

\[ = \sum_{i=1}^{N} \left( c_{i}^{(t)} \right) \log \pi_{j} + \sum_{j=1}^{M} \left( c_{i}^{(t)} \right) \log N(z_{i} | \mu_{j}, \Sigma_{j}) \]

• M-step

\[ \pi_{j}^{(t+1)} = \frac{1}{N} \sum_{i=1}^{N} c_{i}^{(t)} \]

\[ \mu_{j}^{(t+1)} = \frac{\sum_{i=1}^{N} c_{i}^{(t)} z_{i}}{\sum_{i=1}^{N} c_{i}^{(t)}} \]

\[ \Sigma_{j}^{(t+1)} = \frac{\sum_{i=1}^{N} c_{i}^{(t)} (z_{i} - \mu_{j})^{T} (z_{i} - \mu_{j})}{\sum_{i=1}^{N} c_{i}^{(t)}} \]

Gaussian Mixture Models: Limitations

• How many components?
• Ill-conditioned covariance matrices

Variational Bayesian Gaussian Mixture Models


Variational Bayesian Gaussian Mixture Models: Prior Distribution

• Treat parameters as hidden variables \( h = \{ Z, \pi, \mu, \Sigma \} \)
• Introduce conjugate priors

\[ p(\pi | a) = \text{Dir}(\pi | a_{1}, \ldots, a_{M}) = \frac{\prod_{j=1}^{M} \Gamma(a_{j})}{\Gamma(\sum_{j=1}^{M} a_{j})} \pi_{1}^{a_{1} - 1} \cdots \pi_{M}^{a_{M} - 1} \]

\[ p(\mu_{j} | \beta_{j}) = N(\mu_{j} | \mu_{0}, \beta_{j}^{-1}) \]

\[ p(\pi | \pi) = \text{Dir}(\pi | a_{1}, \ldots, a_{M}) \]

\[ p(\mu_{j} | \Sigma_{j}) = \text{Norm}(\mu_{j} | \mu_{0}, \Sigma_{j}) \]

\[ p(\Sigma_{j}) = \text{Inv-Wish}(\Sigma_{j} | V, \nu) = \frac{| \Sigma_{j} |^{\nu/2} \exp [- \nu/2 \text{tr}(V \Sigma_{j})]}{2^{\nu/2} \pi^{\nu M/2} | V |^{-\nu/2} \Gamma(\nu + M - 1)/4^{\nu/2}} \]

Variational Bayesian Gaussian Mixture Models: Mean-Field Approximation

• Exact Bayesian Inference Intractable
• Variational Mean Field Approximation

\[ q(h) = q(Z)q(\pi)q(\mu, \Sigma) \]

\[ q_{\pi}(\pi, \mu, \Sigma) \]
Variational Bayesian Gaussian Mixture Models: Approximate Posteriors

\[
\mathbb{E}_q(Z) = \prod_{i=1}^M q_i(z_i)
\]

\[
\log q_i = \langle \log \pi_i \rangle + \langle \log |\Sigma_i| \rangle - \frac{1}{2} (y_i - \mu_i)' (y_i - \mu_i) - \frac{d}{2} \log 2\pi + \text{const}
\]

\[
\pi_i = \text{Dir}(\alpha_i | \{\lambda_i\})
\]

\[
\lambda_i = \pi_i + \alpha \quad \sum \lambda_i = \sum \pi_i
\]

Variational Bayesian Gaussian Mixture Models: Discussion

- Select Parameters that define uninformative priors (e.g. \(\alpha_i = 1/M\))
- Advantages
  - Disallow singular covariance matrix
  - Bayesian model selection
- Dirichlet distribution for mixing coefficients \(\pi\) disallows pruning of unnecessary components

Variational Bayesian Gaussian Mixture Models: Approximate Posteriors

\[
\mathbb{E}_q(\pi | T) = \prod_{i=1}^M \pi_i (\pi_i m_i + \beta_i T_i)
\]

\[
\bar{m}_i = (\bar{\pi}_i m_i + \beta_i m_i') / (\bar{\pi}_i + \beta_i')
\]

\[
\bar{\lambda}_i = \bar{\pi}_i + \beta_i + \lambda_i (\bar{m}_i - m_i') / (\bar{\pi}_i + \beta_i') + T_i
\]

\[
\pi_i = \text{Dir}(\pi_i | \lambda_i)
\]

\[
\sum \pi_i = \frac{1}{M} \sum \pi_i (y_i - \bar{m}_i)'(y_i - \bar{m}_i)
\]

Variational Bayesian GMM: Removing the prior from the mixing weights

- Treat \(\pi\) as parameter
- Include M-step to update \(\pi\)

\[
\pi_j = \frac{\sum_{r=1}^R f_{jr}}{\sum_{j=1}^J \sum_{r=1}^R f_{jr}}
\]

\[
\sum_{j=1}^J \sum_{r=1}^R f_{jr} = \sum_{r=1}^R \exp \left\{ \frac{1}{2} \left[ (\log |\Sigma|) + \frac{1}{2} (\bar{m} - \Sigma^{-1} \bar{m} + (\Sigma^{-1/2} x)' \Sigma^{-1/2} x) \right] \right\}
\]

- Advantage
  - Eliminates irrelevant components

Variational Bayesian GMM: Example

Incremental Variational Bayesian Gaussian Mixture Models

- Solutions depend on:
  - maximum initial number of components
  - initialization of component parameters
  - specification of the scale matrix $V$ of $p(T_j) = \text{Wishart}(v, V)$

Divide components as ‘fixed’ and ‘free’

Restrict competition among ‘free’ components only

\[
p(\pi | \pi) = \left(1 - \sum_{\pi_i} \right) \prod_{i=1}^{\pi} \Gamma(\sigma_j) \prod_{\pi_i} \left(1 - \sum_{\pi_i} \right)^{-1}
\]

We start by training a GMM with two components

At each step:
  - Select a component $j$
  - Set $V = d \lambda I$, where $\lambda$ the max eigenvalue of $\bar{\Gamma}_j$
  - Split the component in two subcomponents
  - Apply VB learning considering the two components as free

If the data in the region of component $j$ suggest the existence of more than two components then the two components will be retained

Otherwise one of the two components will be removed from the model
Conclusions

**Variational Approximation Pros:**
1. Very Flexible Tool
2. Nice Theoretical Properties
3. Gives Tractable Algorithms
4. Applied to Many Problems*

**Variational Approximation Cons:**
1. Tightness of Bound
2. Sometimes Difficult Calculations