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.

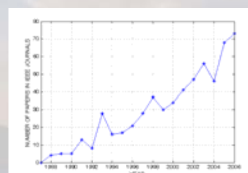
Outline

- 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

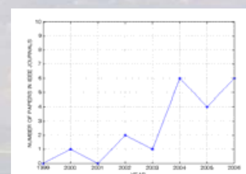
Introduction

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.

Introduction



(a)



(b)

- (a): # of papers /year in IEEE Journals "**EM Algorithm**"
- (b): # of papers /year in IEEE Journals "**Variational Methodology**"

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 6879944 - Variational relevance vector machine, Issued on April 12, 2005
- US Patent 6556960 - Variational inference engine for probabilistic graphical models, Issued on April 29, 2003
- US Patent 20080025627-Removing camera shake from a single photograph
- US Patent 6990447 Method and apparatus for denoising and deverberation using variational inference and Strong Speech Models, Issue date: Jan 24, 2006.
- US Patent 6591146- Method for learning switching linear dynamic system models from data, Issue date: Jul 8, 2003
- US Patent 6931374-Method of speech recognition using variational inference with switching state space models, Issue date: Aug 16, 2005.
- US Patent 2004/0260548-Variational inference and learning for segmental switching state space models of hidden speech dynamics,
- US Patent 2004/0254903 A1-Systems and methods for tractable variational approximation for inference in decision graph Bayesian networks, Issue Feb 27, 2007
- Publication number US 2007/0083928-Data security and intrusion detection,
- Publication number: US 2005/0176057 A1-Diagnostic markers of mood disorders and methods of use thereof
- Publication number: US 2007/0233392 A1-Population sequencing
- Publication number: US 2006/0184260 A1-Player ranking with partial information
- Publication number: US 2007/0265718 A1-Team matching
- Publication number: US 2007/0098254 A1-Detecting humans via their pose

Bayesian Inference Basics

- Estimation => **Parameter**
- Observations: x
- Parameters: θ
- Likelihood Function: $p(x;\theta)$ $(\theta - \text{parameter})$
- Maximum Likelihood Estimation

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

Bayesian Inference Basics

- Inference => θ **Random Variable**
- Find **Posterior** $p(\theta \mid x)$
- Posterior **MORE** information than point estimate
  - $E(\theta \mid x)$ MMSE estimate
  - $Var(\theta \mid x)$ accuracy of estimates

Bayesian Inference Basics

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

Bayesian Inference Basics

- Find Likelihood, **Marginalize** Hidden Variables

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

- Find Posterior

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

- 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 :

$$\hat{\theta}_{MAP} = \operatorname*{argmax}_{\theta} p(\theta \mid x)$$

- Based on Bayes' theorem

$$p(\theta \mid x) = \frac{p(x \mid \theta)\,p(\theta)}{p(x)}$$

- Can be found as:

$$\hat{\theta}_{MAP} = \operatorname*{argmax}_{\theta} p(x \mid \theta)\,p(\theta)$$

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

- <u>Example #1</u>: μ hidden variable

$$p(x \mid \mu;\sigma^2) = N(x;\mu,\sigma^2),\ p(\mu;\mu_0,\sigma_0^2) = N(x;\mu_0,\sigma_0^2)$$

$$p(x;\mu_0,\sigma^2,\sigma_0^2) = \int p(x \mid \mu;\sigma^2)\,p(\mu;\mu_0,\sigma_0^2)\,d\mu$$

Conjugate Priors

- w.r.t. to μ **both** $p(x \mid \mu;\sigma^2)$ and $p(\mu;\mu_0,\sigma_0^2)$ have the **same** form (Gaussian).

$$p(x \mid \mu;\sigma^2) = f(\mu) \propto \exp\left(-\frac{1}{2\sigma^2}(\mu^2 - 2\mu x)\right)$$

$$p(\mu,\mu_0,\sigma_0^2) = g(\mu) \propto \exp\left(-\frac{1}{2\sigma_0^2}(\mu^2 - 2\mu\mu_0)\right)$$

- $p(x \mid \mu;\sigma^2)$, and $p(\mu;\mu_0,\sigma_0^2)$ **conjugate**

Conjugate Priors

- Marginalizing μ possible (Gaussian Integral):

$$p(x;\mu_0,\sigma^2,\sigma_0^2) = \int p(x \mid \mu;\sigma^2)\,p(\mu;\mu_0,\sigma_0^2)\,d\mu$$

$$= N\left(x;\mu_0,\left(\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}\right)^{-1}\right)$$

- Posterior:

$$p(\mu \mid x;\sigma^2,\mu_0,\sigma_0^2) = N\left(\mu;\frac{\sigma_0^2 x + \sigma^2\mu_0}{\sigma^2 + \sigma_0^2},\left(\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}\right)^{-1}\right)$$

Conjugate Priors

- <u>Example #2</u>: a "hidden variable"

$$p(x \mid a) = N(x;0,a^{-1}),$$

$$p(a;b,c) = Gamma(a;b,c) = \frac{c^b a^{b-1}\exp(-ca)}{\Gamma(b)}$$

Conjugate Priors

- w.r.t. to a **both** $p(x|a)$ and $p(a;b,c)$ have the **same** form (Gamma).

$$p(x|a) = f(a) \propto a^{1/2} \exp\left(-\frac{ax^2}{2}\right)$$

$$p(a;b,c) = g(a) \propto a^{b-1} \exp(-ac)$$

- $p(x|a)$ and $p(a;b,c)$ **conjugate**

Conjugate Priors

- Marginalize a possible

$$p(x;b,c) = \frac{\Gamma(b+1/2)}{\Gamma(b)} c^b \left(\frac{1}{2\pi}\right)^{1/2} \left(c + \frac{x^2}{2}\right)^{-b-1/2}$$

- Can write as Student's–t with $\nu = 2b, \lambda = b/c$

$$p(x;\lambda,\nu) = St(x;\lambda,\nu) = \frac{\Gamma(\nu/2+1/2)}{\Gamma(\nu/2)} \left(\frac{\lambda}{\pi\nu}\right)^{1/2} \left(1 + \frac{\lambda x^2}{\nu}\right)^{-\nu/2-1/2}$$

- Posterior

$$p(a|x;b,c) = Gamma\left(a; b + \frac{1}{2}, c + \frac{x^2}{2}\right)$$

Conjugate Priors

Likelihood	Conjugate Prior Distribution	Posterior Distribution
$\mathrm{N}(\mathbf{X}\mid\boldsymbol{\mu},\boldsymbol{\Sigma})$	$\mathrm{N}(\boldsymbol{\mu}\mid\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)$	$\mathrm{N}\left(\boldsymbol{\mu}\mid\left(\boldsymbol{\Sigma}_0^{-1}+n\boldsymbol{\Sigma}^{-1}\right)^{-1}\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0+n\boldsymbol{\Sigma}^{-1}\overline{\mathbf{x}}\right)^{-1},\left(\boldsymbol{\Sigma}_0^{-1}+n\boldsymbol{\Sigma}^{-1}\right)^{-1}\right)$
$\mathrm{N}(\mathbf{x}\mid\mu,\sigma^2)$	$\mathrm{Gamma}(\sigma^{-2}\mid a,b)$	$\mathrm{Gamma}\left(\sigma^{-2}\mid a+n/2, b+\sum_{i=1}^n (x_i-\mu)^2/2\right)$
$\mathrm{N}(\mathbf{X}\mid\boldsymbol{\mu},\boldsymbol{\Sigma})$	$\mathrm{Wishart}\left(\boldsymbol{\Sigma}^{-1}\mid\nu,\mathbf{V}\right)$	$\mathrm{Wishart}\left(\boldsymbol{\Sigma}^{-1}\mid\nu+n,\mathbf{V}+\sum_{i=1}^n(\mathbf{X}_i-\boldsymbol{\mu})^T(\mathbf{X}_i-\boldsymbol{\mu})\right)$
$\mathrm{Multin}(\mathbf{X}\mid\boldsymbol{\pi})$	$\mathrm{Dir}(\boldsymbol{\pi}\mid\mathbf{a})$	$\mathrm{Dir}(\boldsymbol{\pi}\mid\mathbf{a}+\sum_{i=1}^n\mathbf{X}_i)$

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

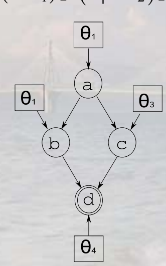
- x_s rv associated node s, $\pi(s)$ parents of s
- $p\left(x_s | x_{\pi(s)}\right)$ conditional pdf
- Joint pdf over all variables

$$p(x) = \prod_s p\left(x_s | x_{\pi(s)}\right)$$

Graphical Models

•Example:

$$p(a,b,c,d;\boldsymbol{\theta}) = p(a;\theta_1)\, p(b|a;\theta_2)\, p(c|a;\theta_3)\, p(d|b,c;\theta_4)$$



EM Algorithm

- x- observations, z -“hidden variables”, θ-parameters.

- Define:

$$Q\left(\theta,\theta^{old}\right)=E\left[\ln p\left(x,z;\theta\right)\right]_{p\left(z|x;\theta^{old}\right)}$$
$$=\int \ln p\left(x,z;\theta\right) p\left(z\mid x;\theta^{old}\right)dz$$

EM Algorithm

1. Initial selection θ^{old}
2. **E-step**: Evaluate $p(z|x;\ \theta^{old})$
3. **M-step:** Evaluate θ^{new}

$$\theta^{new}=\arg\max_{\theta} Q\left(\theta,\theta^{old}\right)$$

4. Check for convergence parameters of log-likelihood, if not satisfied, *go to 2*.

An Alternative View of the EM Algorithm

- Can write:

$$\ln p(x;\theta)=F(q,\theta)+KL(q\|p)$$
$$F(q,\theta)=\int q(z)\ln\left(\frac{p(x,z;\theta)}{q(z)}\right)dz$$
$$KL(q\|p)=-\int q(z)\ln\left(\frac{p(z\mid x;\theta)}{q(z)}\right)dz$$

- $q(z)$ any pdf,
- $KL(q\|p)$ Kullback-Leibler Divergence

An Alternative View of the EM Algorithm ***

- $KL(q\|p)\geq 0$
- $\ln p(x;\theta)\geq F(q,\theta)$
- $\ln p(x;\theta)=F(q,\theta)$ when $KL(q\|p)=0$ or $q(z)=p(z\mid x;\theta)$

An Alternative View of the EM Algorithm ***

- Substitute $q(z)=p(z\mid x;\theta^{OLD})$
- Can write:

$$F(q,\theta)=\int p(z\mid x;\theta^{OLD})\ln p(x,z;\theta)\,dz-$$
$$\int p(z\mid x;\theta^{OLD})\ln p(x\mid z;\theta^{OLD})\,dz$$
$$=Q(\theta,\theta^{OLD})+constant$$

- Same as before:

$$Q(\theta,\theta^{OLD})=E\left[\ln p(x,z;\theta)\right]_{p(z|x;\theta^{OLD})}$$

The Variational EM framework ***

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

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** $F(q,\theta)$ => **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:

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

- with

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

Conjugate-Exponential models

- Prior distributions belong to the exponential family

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

- Graphical model with conjugate priors at each level
  - hidden z, parents $\pi(z)$
  - $p(z \mid \pi(z))$ conjugate to $p(\pi(z) \mid \pi(\pi(z)))$

$$p(z \mid \pi(\pi(z))) = \int p(z \mid \pi(z))\, p(\pi(z) \mid \pi(\pi(z)))\, d\pi(z)$$

Conjugate-Exponential models

- $p(x|z_1)$,$p(z_1|z_2)$,$p(z_2)$ exponential distributions
- $p(z_2)$ conjugate to $p(z_1|z_2)$

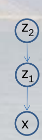
$$p(x \mid z_2) = \int p(x \mid z_1)\, p(z_1 \mid z_2)\, dz_1$$

- $P(x|z_1)$ conjugate to $p(z_1|z_2)$

$$p(z_1) = \int p(z_1 \mid z_2)\, p(z_2)\, dz_2$$

- Cannot evaluate marginal

$$p(x) = \int p(x \mid z_1)\, p(z_1 \mid z_2)\, p(z_2) dz_1 dz_2$$



Conjugate-Exponential models

- Tractable Variational computations

$$q(z_i) \propto \exp\left(\langle \ln p(x,z;\theta)\rangle_{i\neq j}\right)$$

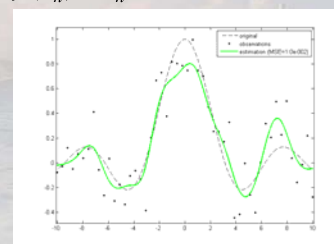
- $\ln q(z_1) = \ln p(x \mid z_1;\theta) + \langle \ln p(z_1 \mid z_2;\theta)\rangle_{q(z_2)} + const$
- $\ln q(z_2) = \langle \ln p(z_1 \mid z_2;\theta)\rangle_{q(z_1)} + \ln p(z_2;\theta) + const$

Examples
Linear Regression

Linear Regression

- Observations at t_n find $y(x)$

$$t_n = y(x_n) + \varepsilon_n,\ n = 1, \ldots, N$$



Linear Regression

- Signal $y(x)$ modeled by

$$y(\mathbf{x};\mathbf{w}) = \sum_{m=1}^{M} w_m \phi_m(\mathbf{x})$$

- Observation model

$$t_n = y(\mathbf{x}_n;\mathbf{w}) + \varepsilon_n, n = 1 \ldots N$$
$$\mathbf{t} = \mathbf{\Phi w} + \boldsymbol{\varepsilon}$$

- $\mathbf{\Phi}$ NXM design matrix

$$\mathbf{\Phi} = (\phi_1, \ldots, \phi_M), \phi_m = (\phi_m(x_1), \ldots, \phi_m(x_N))^T$$

Linear Regression

•Gaussian additive noise

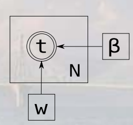
$$p(\boldsymbol{\varepsilon}) = N(\boldsymbol{\varepsilon} \mid 0, \beta^{-1}\mathbf{I})$$

•Likelihood of observations

$$p(\mathbf{t};\mathbf{w},\beta) = N(\mathbf{t} \mid \mathbf{\Phi w}, \beta^{-1}\mathbf{I})$$

Linear Regression: Least Squares ($\boldsymbol{w}$-parameters)

•Graphical Model



•Maximum Likelihood Estimation

$$\mathbf{w}_{ML} = \arg\max_{\mathbf{w}} p(\mathbf{t};\mathbf{w},\beta) = \arg\max_{\mathbf{w}} N(\mathbf{t} \mid \mathbf{\Phi w}, \beta^{-1}\mathbf{I})$$

Linear Regression: Parameter Estimation

- Minimize mean square error

$$E_{LS}(\mathbf{w}) = \|\mathbf{t} - \mathbf{\Phi w}\|^2 = \sum_{n=1}^{N} [t_n - y(\mathbf{x_n};\mathbf{w})]^2$$

- Solution

$$\mathbf{w}_{LS} = (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{t} \qquad \beta^{-1} = \frac{1}{N}\|\mathbf{t} - \mathbf{\Phi w}\|^2$$

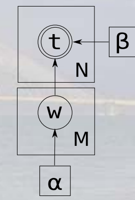
Linear Regression: Problems with Least Squares (*w*-parameters)

- Maximum Likelihood Limitations:
  - N observations M parameters to estimate, $N >> M$
  - Otherwise $\left[\mathbf{\Phi}^T\mathbf{\Phi}\right]^{-1}$ 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

$$p(\mathbf{w};\alpha) = \prod_{m=1}^{M} N(w_m \mid 0, \alpha^{-1})$$



- Why Gaussian?
  - Conjugate to likelihood

Bayesian Linear Regression: Inference

- Posterior given by Bayes's law

$$p(\mathbf{w} \mid \mathbf{t};\alpha,\beta) = \frac{p(\mathbf{t} \mid \mathbf{w};\beta)\,p(\mathbf{w};\alpha)}{p(\mathbf{t};\alpha,\beta)}$$

- Can be found in closed form

$$p(\mathbf{w} \mid \mathbf{t};\alpha,\beta) = N(\mathbf{w} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$$
$$\boldsymbol{\mu} = \beta\boldsymbol{\Sigma}\mathbf{\Phi}^T\mathbf{t}$$
$$\boldsymbol{\Sigma} = (\beta\mathbf{\Phi}^T\mathbf{\Phi} + \alpha\mathbf{I})^{-1}$$

Bayesian Linear Regression: Parameter Estimation

- Maximum Likelihood

$$(\alpha_{ML}, \beta_{ML}) = \underset{\alpha,\beta}{\operatorname{argmin}}\; p(\mathbf{t};\alpha,\beta) = \int p(\mathbf{t} \mid \mathbf{w};\beta)\,p(\mathbf{w};\alpha)\,d\mathbf{w}$$
$$= \underset{\alpha,\beta}{\operatorname{argmin}} \left\{ \log\left|\beta^{-1}\mathbf{I} + \alpha^{-1}\mathbf{\Phi}\mathbf{\Phi}^T\right| + \mathbf{t}^T\left(\beta^{-1}\mathbf{I} + \alpha^{-1}\mathbf{\Phi}\mathbf{\Phi}^T\right)^{-1}\mathbf{t} \right\}$$

- Not straight forward constrained optimization because $\alpha > 0$, $\beta > 0$.
- Resort to EM algorithm!

Bayesian Linear Regression: EM Algorithm

- Observations t, parameters α, β; **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

- $Q^{(t)}(\mathbf{t},\mathbf{w};\alpha,\beta) = \left\langle \ln p(\mathbf{t},\mathbf{w};\alpha,\beta) \right\rangle_{p(\mathbf{w}|\mathbf{t};\alpha^{(t)},\beta^{(t)})} =$
  $\left\langle \ln p(\mathbf{t} \mid \mathbf{w};\alpha,\beta)\,p(\mathbf{w};\alpha,\beta) \right\rangle_{p(\mathbf{w}|\mathbf{t};\alpha^{(t)},\beta^{(t)})}$

$$Q^{(t)}(\mathbf{t},\mathbf{w};\alpha,\beta) = \left\langle \frac{N}{2}\ln\beta - \frac{\beta}{2}\left(\left\|\mathbf{t} - \mathbf{\Phi}\mathbf{w}\right\|^2\right) + \frac{M}{2}\ln\alpha - \frac{\alpha}{2}\left(\left\|\mathbf{w}\right\|^2\right) \right\rangle + \text{const}$$
$$= \frac{N}{2}\ln\beta - \frac{\beta}{2}\left\langle\left\|\mathbf{t} - \mathbf{\Phi}\mathbf{w}\right\|^2\right\rangle + \frac{M}{2}\ln\alpha - \frac{\alpha}{2}\left(\left\langle\left\|\mathbf{w}\right\|^2\right\rangle\right) + \text{const}$$

$$Q^{(t)}(\mathbf{t},\mathbf{w};\alpha,\beta) = \frac{N}{2}\ln\beta - \frac{\beta}{2}\left(\left\|\mathbf{t} - \mathbf{\Phi}\boldsymbol{\mu}^{(t)}\right\|^2 + tr[\mathbf{\Phi}^T\boldsymbol{\Sigma}^{(t)}\mathbf{\Phi}]\right)$$
$$+ \frac{M}{2}\ln\alpha - \frac{\alpha}{2}\left(\left\|\boldsymbol{\mu}^{(t)}\right\|^2 + tr[\boldsymbol{\Sigma}^{(t)}]\right) + \text{const}$$

Linear Regression: EM algorithm, E-step

- E-step: Evaluate $p(\boldsymbol{w}|\boldsymbol{t};\alpha^{(t)},\beta^{(t)})=N(\boldsymbol{\mu}^{(t)},\boldsymbol{\Sigma}^{(t)})$

$$\mathbf{\Sigma}^{(t)} = (\beta^{(t)}\mathbf{\Phi}^T\mathbf{\Phi} + \alpha^{(t)}\mathbf{I})^{-1}$$

$$\boldsymbol{\mu}^{(t)} = \beta^{(t)}\mathbf{\Sigma}^{(t)}\mathbf{\Phi}^T\mathbf{t}$$

Bayesian Linear Regression: Parameter Estimation: M-step

- M-step

$$(\alpha^{(t+1)}, \beta^{(t+1)}) = argmax_{(\alpha,\beta)} Q^{(t)}(\mathbf{t},\mathbf{w};\alpha,\beta)$$

$$\frac{\partial Q^{(t)}(\mathbf{t},\mathbf{w};\alpha,\beta)}{\partial\alpha} = \frac{M}{2\alpha} - \frac{1}{2}\left(\left\|\boldsymbol{\mu}^{(t)}\right\|^2 + tr[\mathbf{\Sigma}^{(t)}]\right)$$

$$\frac{\partial Q^{(t)}(\mathbf{t},\mathbf{w};\alpha,\beta)}{\partial\beta} = \frac{N}{2\beta} - \frac{1}{2}\left(\left\|\mathbf{t} - \mathbf{\Phi}\boldsymbol{\mu}^{(t)}\right\|^2 + tr[\mathbf{\Phi}^T\mathbf{\Sigma}^{(t)}\mathbf{\Phi}]\right)$$

$$\alpha^{(t+1)} = \frac{M}{\left\|\boldsymbol{\mu}^{(t)}\right\|^2 + tr[\mathbf{\Sigma}^{(t)}]} \qquad \beta^{(t+1)} = \frac{N}{\left\|\mathbf{t} - \mathbf{\Phi}\boldsymbol{\mu}^{(t)}\right\|^2 + tr[\mathbf{\Phi}^T\mathbf{\Sigma}^{(t)}\mathbf{\Phi}]}$$

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

- New Prior: w_i not identically distributed

$$p(\mathbf{w}\,|\,\boldsymbol{\alpha}) = \prod_{m=1}^{M} N(w_m\,|\,0,\alpha_m^{-1})$$

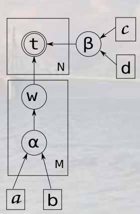
- N-parameters to estimate, N-observations
- Use Conjugate Hyperpriors

$$p(\boldsymbol{\alpha};a,b) = \prod_{m=1}^{M} Gamma(\alpha_m\,|\,a,b)$$

$$p(\beta;c,d) = Gamma(\beta\,|\,c,d)$$

Sparse Bayesian Linear Regression: Prior Distribution

- Graphical Model: $\mathbf{w}$, $\boldsymbol{\alpha}$ **hidden variables**, a,b,c,d, parameters



Sparse Bayesian Linear Regression: Prior Distribution

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

$$p(\mathbf{w};\mathbf{a},b) = \int p(\mathbf{w}\,|\,\boldsymbol{\alpha})\,p(\boldsymbol{\alpha};a,b)\,d\boldsymbol{\alpha}$$

$$= \int \prod_{m=1}^{M} N(w_m\,|\,0,\alpha_m^{-1})\,Gamma(\alpha_m\,|\,a,b)\,d\alpha_m$$

$$= \prod_{m=1}^{m} St(w_m\,|\,\lambda,\nu)$$

Sparse Bayesian Linear Regression: Variational Inference

- Posterior:

$$p(\mathbf{w},\boldsymbol{\alpha},\beta \mid \mathbf{t}) = \frac{p(\mathbf{t} \mid \mathbf{w},\beta)\,p(\mathbf{w} \mid \boldsymbol{\alpha})\,p(\boldsymbol{\alpha})\,p(\beta)}{p(\mathbf{t})}$$

- Cannot compute

$$p(\mathbf{t}) = \int p(\mathbf{t} \mid \mathbf{w},\beta)\,p(\mathbf{w} \mid \boldsymbol{\alpha})\,p(\boldsymbol{\alpha})\,p(\beta)\,d\mathbf{w}\,d\boldsymbol{\alpha}\,d\beta$$

- Variational Mean Field Approximation

$$p(\mathbf{w},\boldsymbol{\alpha},\beta \mid \mathbf{t}) \approx q(\mathbf{w},\boldsymbol{\alpha},\beta) = q(\mathbf{w})q(\boldsymbol{\alpha})q(\beta)$$

Sparse Bayesian Linear Regression: Variational Inference

$$\begin{aligned}
\ln q(\mathbf{w}) &= \left\langle \ln p(\mathbf{t},\mathbf{w},\boldsymbol{\alpha},\beta) \right\rangle_{q(\boldsymbol{\alpha})q(\beta)} + \text{const} \\
&= \left\langle \ln p(t \mid \mathbf{w},\beta)\,p(\mathbf{w} \mid \boldsymbol{\alpha}) \right\rangle_{q(\boldsymbol{\alpha})q(\beta)} + \text{const} \\
&= \left\langle \ln p(t \mid \mathbf{w},\beta) + \ln p(\mathbf{w} \mid \boldsymbol{\alpha}) \right\rangle_{q(\boldsymbol{\alpha})q(\beta)} + \text{const} \\
&= \left\langle -\frac{\beta}{2}(\mathbf{t}-\boldsymbol{\Phi}\mathbf{w})^T(\mathbf{t}-\boldsymbol{\Phi}\mathbf{w}) - \frac{1}{2}\sum_{m=1}^{M} \alpha_m w_m^2 \right\rangle_{q(\boldsymbol{\alpha})q(\beta)} + \text{const} \\
&= -\frac{\langle\beta\rangle}{2}[\mathbf{t}^T\mathbf{t} - 2\mathbf{t}^T\boldsymbol{\Phi}\mathbf{w} + \mathbf{w}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\mathbf{w}] - \frac{1}{2}\sum_{m=1}^{M}\langle\alpha_m\rangle w_m^2 + \text{const} \\
&= -\frac{1}{2}\mathbf{w}^T\left(\langle\beta\rangle\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \langle\mathbf{A}\rangle\right)\mathbf{w} - \langle\beta\rangle\mathbf{w}^T\boldsymbol{\Phi}^T\mathbf{t} + \text{const} \\
&= -\frac{1}{2}\mathbf{w}^T\boldsymbol{\Sigma}^{-1}\mathbf{w} - \mathbf{w}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \text{const}
\end{aligned}$$

$$q(\mathbf{w}) = N(\mathbf{w} \mid \boldsymbol{\mu},\boldsymbol{\Sigma}) \qquad \boldsymbol{\Sigma} = \left(\langle\beta\rangle\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \langle\mathbf{A}\rangle\right)^{-1} \quad \boldsymbol{\mu} = \langle\beta\rangle\boldsymbol{\Sigma}\boldsymbol{\Phi}^T\mathbf{t}$$

Sparse Bayesian Linear Regression: Variational Inference

$$\begin{aligned}
\ln q(\boldsymbol{\alpha}) &= \left\langle \ln p(\mathbf{t},\mathbf{w},\boldsymbol{\alpha},\beta) \right\rangle_{q(\mathbf{w})q(\beta)} + const \\
&= \left\langle \ln p(\mathbf{w} \mid \boldsymbol{\alpha})\,p(\boldsymbol{\alpha}) \right\rangle_{q(w)} + const \\
&= \frac{1}{2}\sum_{m=1}^{M}\ln\alpha_m - \sum_{m=1}^{M}\alpha_m\left\langle w_m^2\right\rangle + (a-1)\sum_{m=1}^{M}\ln\alpha_m - b\sum_{m=1}^{M}\alpha_m + const \\
&= \left(a - \frac{1}{2}\right)\sum_{m=1}^{M}\ln\alpha_m - \sum_{m=1}^{M}\left(\frac{1}{2}\left\langle w_m^2\right\rangle + b\right)\alpha_m + const \\
&= \tilde{a}\sum_{m=1}^{M}\ln\alpha_m - \sum_{m=1}^{M}\tilde{b}_m\alpha_m + const
\end{aligned}$$

$$q(\boldsymbol{\alpha}) = \prod_{m=1}^{M} Gamma(\alpha_m \mid \tilde{a},\tilde{b}_m) \qquad \tilde{a} = a + 1/2 \qquad \tilde{b}_m = b + \frac{1}{2}\left\langle w_m^2\right\rangle$$

Sparse Bayesian Linear Regression: Variational Inference

$$\begin{aligned}
\ln q(\beta) &= \left\langle \ln p(\mathbf{t},\mathbf{w},\boldsymbol{\alpha},\beta) \right\rangle_{q(\mathbf{w})q(\boldsymbol{\alpha})} + const \\
&= \left\langle \ln p(\mathbf{t} \mid \mathbf{w})\,p(\beta) \right\rangle_{q(\mathbf{w})q(\boldsymbol{\alpha})} + const \\
&= \frac{N}{2}\ln\beta - \frac{1}{2}\beta\left\langle\left\|\mathbf{t}-\boldsymbol{\Phi}\mathbf{w}\right\|^2\right\rangle + (c-1)\ln\beta - d\beta + const \\
&= \left(c + \frac{N}{2} - 1\right)\ln\beta - \left(\frac{1}{2}\left\langle\left\|\mathbf{t}-\boldsymbol{\Phi}\mathbf{w}\right\|^2\right\rangle + d\right)\beta + const \\
&= \tilde{c}\ln\beta - \tilde{d}\beta + const
\end{aligned}$$

$$q(\beta) = Gamma(\beta \mid \tilde{c},\tilde{d}) \qquad \tilde{c} = c + N/2 \qquad \tilde{d} = d + \frac{1}{2}\left\langle\left\|\mathbf{t}-\boldsymbol{\Phi}\mathbf{w}\right\|^2\right\rangle$$

Sparse Bayesian Linear Regression: Variational Inference

- Finding the required expectations

$$q(\beta) = Gamma(\beta \mid \tilde{c},\tilde{d}),\ \langle\beta\rangle = \tilde{c}/\tilde{d}$$

$$q(\alpha_m) = Gamma(\alpha_m \mid \tilde{a},\tilde{b}_m), \langle\alpha_m\rangle = \tilde{a}/\tilde{b}_m$$

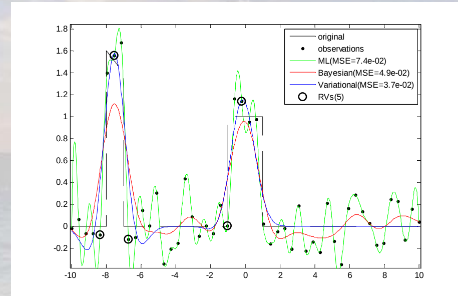
$$q(\mathbf{w}) = N(\mathbf{w} \mid \boldsymbol{\mu},\boldsymbol{\Sigma}), \left\langle w_m^2\right\rangle = \boldsymbol{\mu}_m + \boldsymbol{\Sigma}_{mm}$$

$$\left\langle\left\|\mathbf{t}-\boldsymbol{\Phi}\mathbf{w}\right\|^2\right\rangle = \left\|\mathbf{t}\right\|^2 + tr\left\{\boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Phi}^t\right\} + \boldsymbol{\mu}^t\boldsymbol{\Phi}^t\boldsymbol{\Phi}\boldsymbol{\mu} - 2\mathbf{t}^t\boldsymbol{\Phi}\boldsymbol{\mu}$$

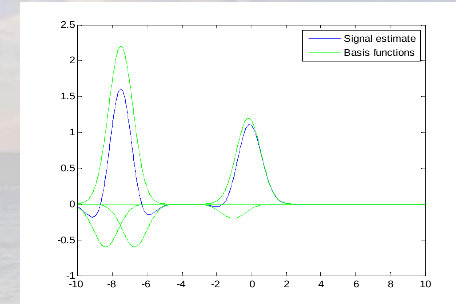
Sparse Bayesian Linear Regression: Parameter Estimation

- Parameters *a,b,c,d*=?
  - Use fixed values that define **uninformative** priors
  - Estimate parameters in Variational M-step
- For fixed *a,b,c,d* iterate only between VE-step for $q(\boldsymbol{\alpha})$, $q(\beta)$ and $q(\boldsymbol{w})$

Linear Regression Example



Linear Regression Example

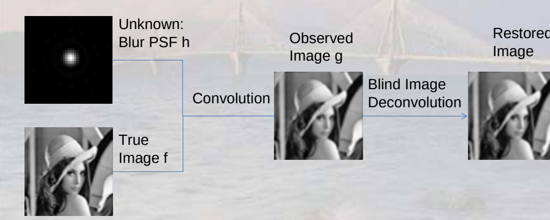


Examples
Blind Image Deconvolution

Blind Image Deconvolution

$$\mathbf{g} = \mathbf{h} * \mathbf{f} + \mathbf{n}$$

$$\mathbf{g} = \mathbf{Hf} + \mathbf{n} = \mathbf{Fh} + \mathbf{n}$$



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

$$p(\mathbf{n} \mid \boldsymbol{\beta}) = \prod_{i=1}^{N} N(n_i \mid 0, \beta_i^{-1}) \qquad p(\boldsymbol{\beta}) = \prod_{i=1}^{N} Gamma(\beta_i \mid a^{\beta}, b^{\beta})$$

$$p(n_i) = \int p(n_i \mid \beta_i) p(\beta_i) d\beta_i = \text{Student's t}$$

Blind Image Deconvolution: PSF Model

- PSF: Sparse Linear model
- Basis functions are Gaussian kernels

$$h(x) = \sum_{i=1}^{N} w_i \phi_i(x) \qquad \mathbf{h} = \mathbf{\Phi w}$$

$$\phi_i(x) = K(x, x_i) \qquad \Phi_{i,j} = K(x_i, x_j)$$

- Sparseness weight prior

$$p(\mathbf{w} \mid \boldsymbol{\alpha}) = \prod_{i=1}^{N} N(w_i \mid 0, \alpha_i^{-1}),\ p(\alpha_i) = Gamma\left(\alpha_i; a^{\alpha}, b^{\alpha}\right)$$

$$p(\mathbf{w}) = \int p(\mathbf{w} \mid \boldsymbol{\alpha}) p(\boldsymbol{\alpha}) d\boldsymbol{\alpha} = \text{Student's t}$$

Blind Image Deconvolution: Image Model

- Directional image differences

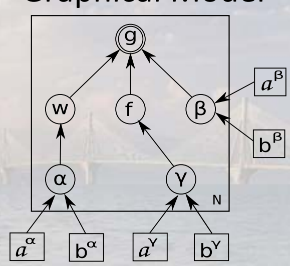
$$\varepsilon^1(x,y) = f(x,y) - f(x+1,y)$$

$$\varepsilon^2(x,y) = f(x,y) - f(x,y+1)$$

$$p(\boldsymbol{\varepsilon}^k \mid \boldsymbol{\gamma}^k) = \prod_{i=1}^{N} N\left(\varepsilon_i^k \mid 0, \left(\gamma_i^k\right)^{-1}\right) \qquad p(\boldsymbol{\gamma}^k) = \prod_{i=1}^{N} Gamma\left(\gamma_i^k \mid a^{\gamma}, b^{\gamma}\right)$$

$$p(\mathbf{f} \mid \tilde{\boldsymbol{\gamma}}) = N\left(\mathbf{f} \mid 0, \left(\tilde{\mathbf{Q}}^T \tilde{\mathbf{\Gamma}} \tilde{\mathbf{Q}}\right)^{-1}\right)$$

Blind Image Deconvolution: Graphical Model



- Joint pdf

$$p(\mathbf{g}, \mathbf{f}, \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma; \theta) = p(\mathbf{g} \mid \mathbf{f}, \mathbf{w}, \boldsymbol{\beta}) p(\mathbf{f} \mid \gamma) p(\mathbf{w} \mid \boldsymbol{\alpha}) p(\boldsymbol{\beta}) p(\gamma) p(\boldsymbol{\alpha})$$

Blind Image Deconvolution: Variational Bound

- Mean field approximation

$$q(\mathbf{f}, \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = q(\mathbf{f}) q(\mathbf{w}) q(\boldsymbol{\alpha}) q(\boldsymbol{\beta}) q(\boldsymbol{\gamma})$$

- Maximization of Variational Bound results in:

$$\log q(\mathbf{f}) = \left\langle \log p(\mathbf{g}, \mathbf{f}, \mathbf{w}, \tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\gamma}}; \theta) \right\rangle_{q(\mathbf{w}) q(\tilde{\boldsymbol{\alpha}}) q(\tilde{\boldsymbol{\beta}}) q(\tilde{\boldsymbol{\gamma}})}$$

$$\log q(\mathbf{w}) = \left\langle \log p(\mathbf{g}, \mathbf{f}, \mathbf{w}, \tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\gamma}}; \theta) \right\rangle_{q(\mathbf{f}) q(\tilde{\boldsymbol{\alpha}}) q(\tilde{\boldsymbol{\beta}}) q(\tilde{\boldsymbol{\gamma}})}$$

$$\log q(\tilde{\boldsymbol{\alpha}}) = \left\langle \log p(\mathbf{g}, \mathbf{f}, \mathbf{w}, \tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\gamma}}; \theta) \right\rangle_{q(\mathbf{w}) q(\mathbf{f}) q(\tilde{\boldsymbol{\beta}}) q(\tilde{\boldsymbol{\gamma}})}$$

$$\log q(\tilde{\boldsymbol{\beta}}) = \left\langle \log p(\mathbf{g}, \mathbf{f}, \mathbf{w}, \tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\gamma}}; \theta) \right\rangle_{q(\mathbf{w}) q(\tilde{\boldsymbol{\alpha}}) q(\mathbf{f}) q(\tilde{\boldsymbol{\gamma}})}$$

$$\log q(\tilde{\boldsymbol{\gamma}}) = \left\langle \log p(\mathbf{g}, \mathbf{f}, \mathbf{w}, \tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\gamma}}; \theta) \right\rangle_{q(\mathbf{w}) q(\tilde{\boldsymbol{\alpha}}) q(\mathbf{f}) q(\tilde{\boldsymbol{\beta}})}$$

Blind Image Deconvolution: Find approximate posterior $q(\mathbf{f})$

- Mean Field Optimization

$$\log q(\mathbf{f}) = \left\langle -\frac{1}{2}(\mathbf{\Phi W f} - \mathbf{g})^T \mathbf{B} (\mathbf{\Phi W f} - \mathbf{g}) - \frac{1}{2} \mathbf{f}^T \tilde{\mathbf{Q}}^T \mathbf{\Gamma} \tilde{\mathbf{Q}} \mathbf{f} \right\rangle_{q(\mathbf{w}) q(\tilde{\boldsymbol{\beta}}) q(\tilde{\boldsymbol{\gamma}})} + const$$

$$= \left\langle -\frac{1}{2} \mathbf{f}^T (\mathbf{\Phi}^T \mathbf{W}^T \mathbf{B W \Phi} + \tilde{\mathbf{Q}}^T \mathbf{\Gamma} \tilde{\mathbf{Q}}) \mathbf{f} + \mathbf{f}^T \mathbf{\Phi}^T \mathbf{W}^T \mathbf{B g} \right\rangle_{q(\mathbf{w}) q(\tilde{\boldsymbol{\beta}}) q(\tilde{\boldsymbol{\gamma}})} + const$$

$$= -\frac{1}{2} \mathbf{f}^T (\mathbf{\Phi}^T \left\langle \mathbf{W}^T \mathbf{B W} \right\rangle \mathbf{\Phi} + \tilde{\mathbf{Q}}^T \left\langle \mathbf{\Gamma} \right\rangle \tilde{\mathbf{Q}}) \mathbf{f} + \mathbf{f}^T \mathbf{\Phi}^T \left\langle \mathbf{W}^T \mathbf{B} \right\rangle \mathbf{g}$$

- Completing Square results in

$$q(\mathbf{f}) = N(\mathbf{f} \mid \boldsymbol{\mu}_f, \mathbf{\Sigma}_f)$$

$$\boldsymbol{\mu}_f = \mathbf{\Sigma}_f \mathbf{\Phi}^T < \mathbf{W}^T \mathbf{B} > \mathbf{g}$$

$$\mathbf{\Sigma}_f = \left(\mathbf{\Phi}^T < \mathbf{W}^T \mathbf{B W} > \mathbf{\Phi} + \tilde{\mathbf{Q}}^T < \tilde{\mathbf{\Gamma}} > \tilde{\mathbf{Q}}\right)^{-1}$$

Blind Image Deconvolution: Approximate posterior $q(\mathbf{w})$

- Similar calculations

$$\log q(\mathbf{w}) = \left\langle -\frac{1}{2}(\mathbf{F \Phi w} - \mathbf{g})^T \mathbf{B} (\mathbf{F \Phi w} - \mathbf{g}) - \frac{1}{2} \mathbf{w}^T \mathbf{A w} \right\rangle_{q(\mathbf{f}) q(\tilde{\boldsymbol{\beta}}) q(\tilde{\boldsymbol{\alpha}})} + const$$

$$= -\frac{1}{2} \mathbf{w}^T (\mathbf{\Phi}^T \left\langle \mathbf{F}^T \mathbf{B F} \right\rangle \mathbf{\Phi} + \left\langle \mathbf{A} \right\rangle) \mathbf{w} + \mathbf{w}^T \mathbf{\Phi}^T \left\langle \mathbf{F}^T \mathbf{B} \right\rangle \mathbf{g} + const$$

$$q(\mathbf{w}) = N(\mathbf{w} \mid \boldsymbol{\mu}_w, \mathbf{\Sigma}_w)$$

$$\boldsymbol{\mu}_w = \mathbf{\Sigma}_w \mathbf{\Phi}^T < \mathbf{F}^T \mathbf{B} > \mathbf{g}$$

$$\mathbf{\Sigma}_w = \left(\mathbf{\Phi}^T < \mathbf{F}^T \mathbf{B F} > \mathbf{\Phi} + < \mathbf{A} >\right)^{-1}$$

Blind Image Deconvolution: Approximate posterior $q(\boldsymbol{\alpha})$

- Based on the mean field approximation

$$\log q(\boldsymbol{\alpha}) = \left\langle \log p(\mathbf{w} \mid \boldsymbol{\alpha}) + \log p(\boldsymbol{\alpha}) \right\rangle_{q(\mathbf{w})} + const$$

$$= \left\langle \frac{1}{2}\sum_{i=1}^{N} \log \alpha_i - \frac{1}{2}\sum_{i=1}^{N} \alpha_i w_i^2 + \left(a^{\alpha} - 1\right)\sum_{i=1}^{N} \log \alpha_i - b^{\alpha} \sum_{i=1}^{N} \alpha_i \right\rangle_{q(\mathbf{w})} + const$$

$$= \sum_{i=1}^{N} (a^{\alpha} - 1 + \frac{1}{2}) \log \alpha_i - \sum_{i=1}^{N} (b^{\alpha} + \left\langle w_i^2 \right\rangle) \alpha_i + const$$

- Which implies

$$q(\boldsymbol{\alpha}) = \prod_{i=1}^{N} Gamma(\alpha_i \mid \tilde{a}^{\alpha}, \tilde{b}_i^{\alpha})$$

$$\tilde{a}^{\alpha} = a^{\alpha} + 1/2$$

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

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

- Similarly we get

$$q(\boldsymbol{\beta}) = \prod_{i=1}^{N} Gamma(\beta_i \mid \tilde{a}^{\beta}, \tilde{b}_i^{\beta})$$

$$\tilde{a}^{\gamma} = a^{\gamma} + 1/2$$

$$\tilde{b}_i^{\gamma} = b^{\gamma} + \frac{1}{2}\left((\mathbf{Q}^k)^T < \mathbf{f}\mathbf{f}^T > \mathbf{Q}^k\right)_{ii}$$

$$q(\boldsymbol{\gamma}) = \prod_{k=1}^{K} \prod_{i=1}^{N} Gamma(\gamma_i^k \mid \tilde{a}^{\gamma}, \tilde{b}_i^{\gamma k})$$

$$\tilde{a}^{\beta} = a^{\beta} + 1/2$$

$$\tilde{b}_i^{\beta} = b^{\beta} + \frac{1}{2} < \mathbf{n}\mathbf{n}^T >_{ii}$$

$$\mathbf{n} = \mathbf{F}\boldsymbol{\Phi}\mathbf{w} - \mathbf{g}$$

Blind Image Deconvolution: Statistics of Approximate Posteriors

$$\begin{aligned}
\langle \boldsymbol{w} \rangle &= \boldsymbol{\mu}_w, \\
\langle w_i^2 \rangle &= \mu_{w_i}^2 + \Sigma_{w_{ii}}, \\
\langle \boldsymbol{f} \rangle &= \boldsymbol{\mu}_f, \\
\langle \boldsymbol{f}\boldsymbol{f}^T \rangle &= \boldsymbol{\mu}_f \boldsymbol{\mu}_f^T + \boldsymbol{\Sigma}_f, \\
\langle \alpha_i \rangle &= \tilde{a}^{\alpha} / \tilde{b}_i^{\alpha}, \\
\langle \beta_i \rangle &= \tilde{a}^{\beta} / \tilde{b}_i^{\beta}, \\
\langle \gamma_i^k \rangle &= \tilde{a}^{\gamma} / \tilde{b}_i^{\gamma k},
\end{aligned}$$

$$\langle \boldsymbol{n}\boldsymbol{n}^T \rangle = \boldsymbol{g}\boldsymbol{g}^T - 2\boldsymbol{\Phi}\langle \boldsymbol{F}\boldsymbol{w} \rangle \boldsymbol{g}^T + \boldsymbol{\Phi}\langle \boldsymbol{F}\boldsymbol{w}\boldsymbol{w}^T\boldsymbol{F}^T \rangle \boldsymbol{\Phi}^T.$$

Blind Image Deconvolution: Statistics of Approximate Posteriors

- Make diagonal and circulant approximations

$$\begin{aligned}
\bar{\boldsymbol{\Sigma}}_w &= \left(\mathrm{diag}\{\boldsymbol{\Phi}^T \langle \boldsymbol{F}^T \boldsymbol{B} \boldsymbol{F} \rangle \boldsymbol{\Phi}\} + \langle \boldsymbol{A} \rangle\right)^{-1}, \\
\bar{\boldsymbol{\Sigma}}_f &= \left(\langle \bar{\beta} \rangle \boldsymbol{\Phi}^T \langle \boldsymbol{W}^T \boldsymbol{W} \rangle \boldsymbol{\Phi} + \langle \bar{\gamma} \rangle \bar{\boldsymbol{Q}}^T \bar{\boldsymbol{Q}}\right)^{-1},
\end{aligned}$$

$$\bar{\gamma} = \frac{1}{MK}\sum_{k=1}^{M}\sum_{i=1}^{M}\gamma_i^k, \quad \bar{\beta} = \frac{1}{M}\sum_{i=1}^{M}\beta_i$$

- This results in

$$\begin{aligned}
\langle \boldsymbol{W}^T \boldsymbol{W} \rangle &= \langle \boldsymbol{W}^T \rangle \langle \boldsymbol{W} \rangle + I \sum_{i=1}^{M} \langle \bar{\Sigma}_{w_{ii}} \rangle, \\
\langle \boldsymbol{F}^T \boldsymbol{B} \boldsymbol{F} \rangle &= \langle \boldsymbol{F}^T \rangle \langle \boldsymbol{B} \rangle \langle \boldsymbol{F} \rangle + \bar{\boldsymbol{\Sigma}}_f \sum_{i=1}^{M} \langle \beta_i \rangle.
\end{aligned}$$

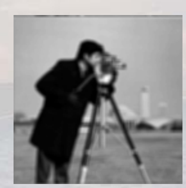
Blind Image Deconvolution: Overall V-EM Algorithm

- Fix parameters $\left\{\mathbf{a}^{\alpha}, \mathbf{b}^{\alpha}, \mathbf{a}^{\beta}, \mathbf{b}.^{\beta} \mathbf{a}^{\gamma}, \mathbf{b}^{\gamma}\right\}$ to yield **uninformative hyperpriors** (no Variational M-step).

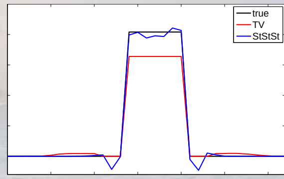
- Iterate between estimates of statistics of $q(\mathbf{f})$, $q(\mathbf{w})$, $q(\boldsymbol{\alpha})$, $q(\boldsymbol{\beta})$ and $q(\boldsymbol{\gamma})$ (only Variational E-step).

Blind Image Deconvolution: Example

- PSF: Square 7x7, 40db noise



Blind Image Deconvolution: Examples



Example: Image Restoration (Constrained Variational Inference*)

*G. Chantas, N. Galatsanos, A. Likas and M. Saunders, "Variational Bayesian Image Restoration Based on a Product of T-Distributions Image Prior", *IEEE Trans. on Image Processing*, to appear. (available on line)

Image Restoration

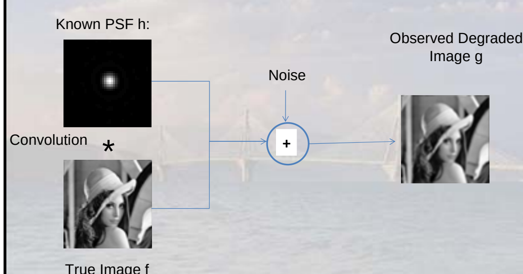


Image Restoration: Problem Definition

- Imaging Model (N-pixels)

$$\mathbf{g} = \mathbf{h} * \mathbf{f} + \mathbf{n} = \mathbf{H}\mathbf{f} + \mathbf{n}$$

- $\mathbf{g}: (N \times 1)$: Degraded (*Observations*)
- $\mathbf{h}: (N \times 1)$, $\mathbf{H}: (N \times N)$: Point Spread Function (*known*)
- $\mathbf{f}: (N \times 1)$: Original Image (***unknown***)
- $\mathbf{n} \sim N\left(\mathbf{0}, \beta^{-1}\mathbf{I}\right)$: Noise (***unknown***)

Image Restoration: $\mathbf{f}$ Parameter

- Likelihood of Observations

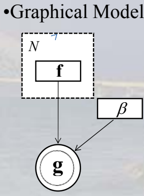
$$p(\mathbf{g}; \mathbf{f}, \beta) = N\left(\mathbf{H}\mathbf{f}, \beta^{-1}\mathbf{I}\right)$$

$$\hat{\mathbf{f}} = \arg\max\left\{p\left(\mathbf{g}; \mathbf{f}, \beta\right)\right\} = \arg\min \left\|\mathbf{g} - \mathbf{H}\mathbf{f}\right\|^2$$

$$\hat{\mathbf{f}} = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{g}$$

- Problematic **too many** parameters

- Graphical Model



Double circled observed r.v.

Image Restoration: $\mathbf{f}$ Parameter-Example

$$SNR = 20\log_{10}\frac{\|\mathbf{f}\|}{\|\mathbf{n}\|}$$

SNR~80dB *SNR~120dB*

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

- Error from Local Linear Predictor

$$f(i)-\overbrace{\frac{1}{2}\left(f(i-1)+f(i+1)\right)}^{\hat{f}(i)}=\varepsilon(i)$$

- Assume Gaussian i.i.d. Prediction Errors

$$p\left(\varepsilon(i)\right)=N\left(0,\alpha^{-1}\right) \quad p(\boldsymbol{\varepsilon})=\prod_{i=1}^{N}p\left(\varepsilon(i)\right)$$

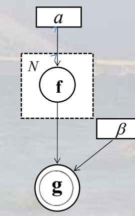
Image Restoration: Bayesian Inference (f: *hidden r.v.*)

- Simultaneously Autoregressive (SAR) Prior

$\mathbf{Qf}=\boldsymbol{\varepsilon},\ \mathbf{Q}$ Operator that describes $\boldsymbol{\varepsilon}=\mathbf{f}-\hat{\mathbf{f}}$

$$p(\mathbf{f};a)\propto\alpha^{\frac{N-1}{2}}\exp\left(-\alpha\frac{\mathbf{f}^T\mathbf{Q}^T\mathbf{Q}\mathbf{f}}{2}\right)$$

Graphical Model



Double circled observed r.v.

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

- Observation Likelihood (*computed analytically*)

$$p(\mathbf{g};\beta,a)=\int p(\mathbf{g}\,|\,\mathbf{f};\beta)p(\mathbf{f};a)d\mathbf{f}$$

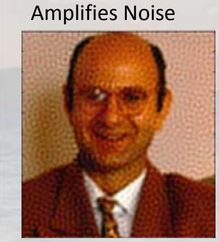
- Posterior of Hidden (*computed analytically*)

$$p(\mathbf{f}\,|\,\mathbf{g};a,\beta)=N\left(\boldsymbol{\mu}_{f|g},\boldsymbol{\Sigma}_{f|g}\right)$$

- Bayesian Inference via **EM**

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

"Small" α/β
Amplifies Noise



"Large" α/β
Smoothes Edges

Image Restoration: Spatially-Varying Bayesian Model

- **Spatially Varying** $p\left(\varepsilon(i)\right)=N\left(0,a_i^{-1}\right)$
- α_i " hidden variables" Bayesian Inference
- Conjugate pdf $p(\alpha_i)=Gamma\left(a_i;\alpha,\beta\right)$

Image Restoration: Spatially-Varying, Bayesian Inference

- Use **many** $\boldsymbol{\varepsilon}_k=\mathbf{Q}_k\mathbf{f},\ k=1,2\ldots P$ in prior
- "Product" Prior

$$p(\mathbf{f}\,|\,\tilde{\mathbf{a}})=\frac{1}{Z(\tilde{\mathbf{a}})}\prod_{k=1}^{P}p_k(\mathbf{f}\,|\,\mathbf{a}_k),\ p(\mathbf{f}\,|\,\mathbf{a}_k)\propto\exp\left(-\frac{1}{2}\mathbf{f}^T\mathbf{Q}_k^T\mathbf{A}_k\mathbf{Q}_k\mathbf{f}\right)$$

$$\mathbf{A}_k=diag\left(a_k(1),a_k(2)\ldots a_k(N)\right)$$

- Prior: Enforces many properties **simultaneously**

Image Restoration: Spatially Varying Bayesian Inference

Graphical Model

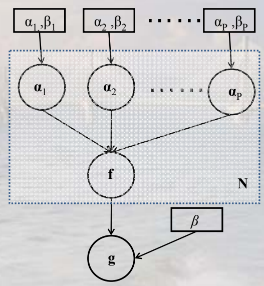


Image Restoration: Spatially Varying, Bayesian Inference

- Difficulty: Compute normalization of prior

$$Z(\tilde{\mathbf{a}}) \propto \det\left\{\sum_{k=1}^{K} \mathbf{Q}_k^T \mathbf{A}_k \mathbf{Q}_k\right\} \quad \mathbf{Q}_k, \mathbf{A}_k \ (N\text{x}N),\ N=10^5\text{-}10^6$$

Image Restoration: Spatially-Varying, Bayesian Inference

•Change Observations Domain

$$\mathbf{Q}_k\mathbf{g} = \mathbf{Q}_k\mathbf{H}\mathbf{f} + \mathbf{Q}_k\mathbf{n},$$
$$\mathbf{y}_k = \mathbf{H}\boldsymbol{\varepsilon}_k + \mathbf{n}_k, k = 1,\ldots,P$$

•Hidden Variables:

$$\tilde{\boldsymbol{\varepsilon}} = [\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2 \ldots \boldsymbol{\varepsilon}_P], \tilde{\mathbf{a}} = [\mathbf{a}_1, \mathbf{a}_2 \ldots \mathbf{a}_P]$$
$$\boldsymbol{\varepsilon}_l = [\varepsilon_l(1), \varepsilon_l(2) \ldots \varepsilon_l(N)], \mathbf{a}_l = [a_l(1), a_l(2) \ldots a_l(N)]$$

Image Restoration: Spatially-varying, Bayesian Inference

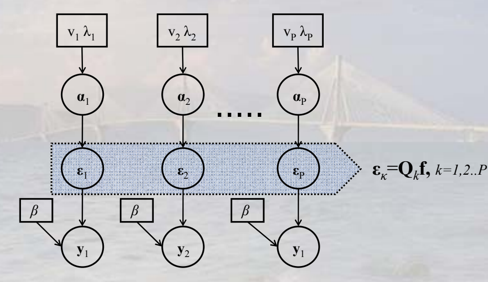


Image Restoration: Spatially Varying Bayesian Inference

•Prior on $\tilde{\boldsymbol{\varepsilon}}$ **NO difficulty** normalizing

$$p(\tilde{\boldsymbol{\varepsilon}} \mid \tilde{\mathbf{a}}) = \prod_{k=1}^{P}\prod_{i=1}^{N} p(\varepsilon_k(i) \mid a_k(i)),$$
$$p(\varepsilon_k(i) \mid a_k(i); \lambda_k) = N\left(\varepsilon_k(i); 0, \lambda_k a_k(i)^{-1}\right)$$
$$p(a_k(i); \nu_k) = Gamma\left(a_k(i); \frac{\nu_k}{2}, \frac{\nu_k}{2}\right)$$

•Cannot marginalize hidden variables

•Resort to Variational Methodology

Image Restoration: Spatially-Varying, Bayesian Inference

•"Posteriors" Mean Field Approximation

$$q(\boldsymbol{\varepsilon}_k, \mathbf{a}_k) = q(\boldsymbol{\varepsilon}_k)\, q(\mathbf{a}_k),\ k = 1,\ldots,P$$

•Maximize Var. Bound

$$q(\boldsymbol{\varepsilon}_k) = N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$
$$\boldsymbol{\mu}_k = \beta\boldsymbol{\Sigma}_k\mathbf{H}^T\mathbf{g},$$
$$\boldsymbol{\Sigma}_k = \left(\beta\mathbf{H}^T\mathbf{H} + \lambda_k\mathbf{Q}_k^T\mathbf{A}_k\mathbf{Q}_k\right)^{-1}$$

•New problem: **Different** $\mathbf{f}$ for each $\boldsymbol{\mu}_k$

Image Restoration: Constrained Variational Inference

- Define "**Constrained Posterior**"

$$q(\boldsymbol{\varepsilon}_k) = \boxed{N\left(\mathbf{Q}_k\mathbf{m}, \mathbf{Q}_k\mathbf{R}\mathbf{Q}_k^T\right)}$$

- Consistent with

$$\boldsymbol{\varepsilon}_k = \mathbf{Q}_k\mathbf{f},\ \mathbf{m} = E(\mathbf{f}),\ \mathbf{R} = E\left((\mathbf{f}-\mathbf{m})^T(\mathbf{f}-\mathbf{m})\right)$$

- Maximize the **Variational Bound** w.r.t. $\boxed{\mathbf{m} \text{ and } \mathbf{R}}$

Image Restoration: Constrained Variational Inference

- VE-step: $\left[q^{(t+1)}(\tilde{\mathbf{a}}), \theta_1^{(t+1)}\right] = \arg\max_{[q(\tilde{\mathbf{a}}),\theta_1]}\left(F\left(q(\tilde{\mathbf{a}}), \theta_1, \theta_2^{(t)}\right)\right)$

- VM-step: $\theta_2^{(t+1)} = \arg\max_{\theta_2}\left(F\left(q^{(t+1)}(\tilde{\mathbf{a}}), \theta_1^{(t+1)}, \theta_2\right)\right)$

$$\theta_1 = [\mathbf{R}, \mathbf{m}],\quad \theta_2 = [\beta, \lambda_1, \ldots, \lambda_P, \nu_1, \ldots, \nu_P]^{\mathrm{T}}$$

Image Restoration: Constrained Variational Inference

- <u>VE-Step</u>

$$F(q,\theta) = \int\prod_{k=1}^{P} q(\boldsymbol{\varepsilon}_k;\theta_1)\, q(\mathbf{a}_k) \log p(\tilde{\mathbf{y}}, \tilde{\boldsymbol{\varepsilon}}, \tilde{\mathbf{a}}; \theta_2)\, d\tilde{\boldsymbol{\varepsilon}} d\tilde{\mathbf{a}} -$$

$$-\int\prod_{k=1}^{P} q(\boldsymbol{\varepsilon}_k;\theta_1)\, q(\mathbf{a}_k) \log \prod_{k=1}^{P} q(\boldsymbol{\varepsilon}_k;\theta_1)\, q(\mathbf{a}_k)\, d\tilde{\boldsymbol{\varepsilon}} d\tilde{\mathbf{a}},$$

$$F \propto \tilde{F}(\theta_1)$$

$$\tilde{F}(\theta_1) = \sum_{k=1}^{P}\int q(\boldsymbol{\varepsilon}_k;\theta_1) q(\mathbf{a}_k) \log p(\mathbf{y}_k \mid \boldsymbol{\varepsilon}_k; \theta_2)\, p(\boldsymbol{\varepsilon}_k \mid \mathbf{a}_k; \theta_2)\, d\boldsymbol{\varepsilon}_k d\mathbf{a}_k -$$

$$-\sum_{k=1}^{P}\int q(\boldsymbol{\varepsilon}_k;\theta_1) \log q(\boldsymbol{\varepsilon}_k;\theta_1)\, d\boldsymbol{\varepsilon}_k.$$

Image Restoration: Constrained Variational Inference

$$\sum_{k=1}^{P}\int q(\boldsymbol{\varepsilon}_k;\theta_1) q(\mathbf{a}_k) \log p(\mathbf{y}_k \mid \boldsymbol{\varepsilon}_k; \theta_2)\, p(\boldsymbol{\varepsilon}_k \mid \mathbf{a}_k; \theta_2)\, d\boldsymbol{\varepsilon}_k d\mathbf{a}_k$$

$$\propto \sum_{k=1}^{P}\left\langle -\beta(\mathbf{H}\boldsymbol{\varepsilon}_k - \mathbf{y}_k)^{\mathrm{T}} \mathbf{Q}_k^{-\mathrm{T}} \mathbf{Q}_k^{-1} (\mathbf{H}\boldsymbol{\varepsilon}_k - \mathbf{y}_k) - \lambda_k \boldsymbol{\varepsilon}_k^{\mathrm{T}} \mathbf{A}_k \boldsymbol{\varepsilon}_k \right\rangle_{q(\boldsymbol{\varepsilon}_k;\theta_1) q(\mathbf{a}_k)} =$$

$$-\beta P\|\mathbf{H}\mathbf{m} - \mathbf{g}\|_2^2 - \sum_{k=1}^{P}\lambda_k \mathbf{m}^{\mathrm{T}} \mathbf{Q}_k^{\mathrm{T}} \langle \mathbf{A}_k \rangle \mathbf{Q}_k \mathbf{m} - trace\left\{\left(\beta P \mathbf{H}^{\mathrm{T}}\mathbf{H} + \sum_{k=1}^{P}\lambda_k \mathbf{Q}_k^{\mathrm{T}} \langle \mathbf{A}_k \rangle \mathbf{Q}_k\right)\mathbf{R}\right\}$$

$$\int q(\boldsymbol{\varepsilon}_k;\theta_1) \log q(\boldsymbol{\varepsilon}_k;\theta_1)\, d\boldsymbol{\varepsilon}_k \propto \frac{1}{2}\log\det|\mathbf{R}|$$

Image Restoration: Constrained Variational Inference

$$\frac{\partial \tilde{F}(\theta_1)}{\partial \mathbf{R}} = 0 \Rightarrow \frac{\partial trace\left\{\beta P \mathbf{H}^{\mathrm{T}}\mathbf{H}\mathbf{R} + \sum_{k=1}^{P}\lambda_k \mathbf{Q}_k^{\mathrm{T}} \hat{\mathbf{A}}_k \mathbf{Q}_k \mathbf{R}\right\} - P\partial \log\det|\mathbf{R}|}{\partial \mathbf{R}} = 0$$

$$\Rightarrow \beta P \mathbf{H}^{\mathrm{T}}\mathbf{H} + \sum_{k=1}^{P}\lambda_k \mathbf{Q}_k^{\mathrm{T}} \hat{\mathbf{A}}_k \mathbf{Q}_k - P\mathbf{R}^{-1} = 0 \Rightarrow \mathbf{R} = \left(\beta \mathbf{H}^{\mathrm{T}}\mathbf{H} + \frac{1}{P}\sum_{k=1}^{P}\lambda_k \mathbf{Q}_k^{\mathrm{T}} \hat{\mathbf{A}}_k \mathbf{Q}_k\right)^{-1}$$

$$\frac{\partial \tilde{F}(\theta_1)}{\partial \mathbf{m}} = 0 \Rightarrow \mathbf{m} = \beta \mathbf{R}\mathbf{H}^{\mathrm{T}}\mathbf{g}$$

Image Restoration: Constrained Variational Inference

$$q(\tilde{\mathbf{a}}) = \exp\left\langle \log p(\tilde{\mathbf{y}}, \tilde{\boldsymbol{\varepsilon}}, \tilde{\mathbf{a}}) \right\rangle_{q(\tilde{\boldsymbol{\varepsilon}})} \propto$$

$$\prod_{k=1}^{P}\prod_{i=1}^{N} (a_k(i))^{\frac{\nu_k}{2}+\frac{1}{2}-1} \exp\left\{-\frac{\nu_k}{2} a_k(i) - \frac{1}{2}\lambda_k\left((\mathbf{m}_k(i))^2 + \mathbf{C}_k(i,i)\right) a_k(i)\right\}$$

$$q(a_k(i)) = Gamma\left(a_k(i); \frac{\nu_k}{2} + \frac{1}{2}, \frac{\nu_k}{2} + \frac{1}{2}\lambda_k\left((\mathbf{m}_k(i))^2 + \mathbf{C}_k(i,i)\right)\right)$$

$$\mathbf{m}_k = \mathbf{Q}_k\mathbf{m},\ \text{and}\ \mathbf{C}_k = \mathbf{Q}_k\mathbf{R}\mathbf{Q}_k^{\mathrm{T}}$$

Image Restoration: Constrained Variational Inference

- VM-step

$$\frac{dF\left(q^{(t+1)}(\tilde{\mathbf{a}}),\theta_1^{(t+1)},\theta_2\right)}{d\theta_2}=\frac{d\left\langle \log p\left(\tilde{\mathbf{y}},\tilde{\boldsymbol{\varepsilon}},\tilde{\mathbf{a}};\theta_2\right)\right\rangle_{q\left(\tilde{\boldsymbol{\varepsilon}};\theta_1^{(t+1)}\right),q^{(t+1)}(\tilde{\mathbf{a}})}}{d\theta_2}0$$

$$\Rightarrow$$

$$\beta^{(t+1)}=N\left(\left\|\mathbf{Hm}^{(t+1)}-\mathbf{g}\right\|_2^2+trace\left\{\mathbf{H}^{\mathrm{T}}\mathbf{HR}^{(t+1)}\right\}\right)^{-1}$$

$$\lambda_k^{(t+1)}=N\left(\sum_{i=1}^{N}\left(\left(\mathbf{m}_k^{(t+1)}(i)\right)^2+\mathbf{C}_k^{(t+1)}(i,i)\right)\left\langle a_k(i)\right\rangle_{q^{(t+1)}(a_k(i))}\right)^{-1}$$

$$\frac{1}{N}\left(\sum_{i=1}^{N}\log\left\langle a_k(i)\right\rangle_{q^{(t+1)}(\tilde{\mathbf{a}})}-\sum_{i=1}^{N}\left\langle a_k(i)\right\rangle_{q^{(t+1)}(\tilde{\mathbf{a}})}\right)+\psi\left(\nu_k^{(t)}\frac{1}{2}+\frac{1}{2}\right)-$$

$$\log\left(\nu_k^{(t)}\frac{1}{2}+\frac{1}{2}\right)-\psi\left(\frac{\nu_k}{2}\right)+\log\left(\frac{\nu_k}{2}\right)+1=0 \qquad \psi(x)=\frac{d}{dx}\log\Gamma(x)$$

Image Restoration: Constrained Variational Inference

1. Initialize, $\mathbf{m}$ stationary model.
2. Repeat until convergence:
   VE-step:
   Update, $\mathbf{m}$ and $\mathbf{R}$, calculate $\mathbf{m}_k$ and $\mathbf{C}_k$. Calculate expected value w.r.t. $q(a_k(i))$, needed for VM-step and the next VE-step.
   VM-step:
   Update β, λ_k, ν_k, $k=1,2...P$
3. Use $\mathbf{m}$ as restored image estimate.

Image Restoration: Constrained Variational Inference

Example: Image Restoration (Bounded Variational Inference*)

*Babacan, S.D.; Molina, R.; Katsaggelos, A.K.; "Parameter Estimation in TV Image Restoration Using Variational Distribution Approximation" , *IEEE Trans. on Image Processing*, Volume 17, Issue 3, March 2008 Page(s):326 - 339

Image Restoration: Bounded Variational Inference

- Imaging Model (N-pixels)

$$\mathbf{g}=\mathbf{h}*\mathbf{f}+\mathbf{n}=\mathbf{Hf}+\mathbf{n}$$

- Observations Likelihood

$$p(\mathbf{g}/\mathbf{f};\beta)\sim N\left(\mathbf{Hf},\beta^{-1}\mathbf{I}\right)$$

Image Restoration: Bounded Variational Inference

- TV Based Image Prior

$$p(\mathbf{f}\,|\,a)=\alpha^{\frac{N^2}{2}}\exp\left(-\alpha\mathrm{TV}(\mathbf{f})\right)$$

$$\mathrm{TV}(\mathbf{f})=\sum_{i,j=1}^{N}\sqrt{\left(\left[\mathbf{Q}_1\mathbf{f}\right]_{ij}\right)^2+\left(\left[\mathbf{Q}_2\mathbf{f}\right]_{ij}\right)^2}$$

$$\left[\mathbf{Q}_1\mathbf{f}\right]_{ij}=\mathbf{f}(i,j)-\mathbf{f}(i,j-1),\ \left[\mathbf{Q}_2\mathbf{f}\right]_{ij}=\mathbf{f}(i,j)-\mathbf{f}(i-1,j)$$

- Conjugate Hyperprior

$$p\left(\alpha;a,b\right)=Gamma(\alpha;a,b)$$

Image Restoration: Bounded Variational Inference

Graphical Model

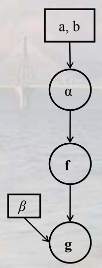


Image Restoration: Bounded Variational Inference

- Difficulty in VE-step

$$\log q(\alpha) = \left\langle \log p(\mathbf{g},\mathbf{f},\alpha;\beta,a,b) \right\rangle_{q(\mathbf{f})} =$$

$$\left\langle -\frac{1}{2}\beta\left\|Hf-g\right\|^2 - \alpha\sum_{i,j=1}^{N}\sqrt{\left([\mathbf{Q}_1\mathbf{f}]_{ij}\right)^2+\left([\mathbf{Q}_2\mathbf{f}]_{ij}\right)^2} \right\rangle_{q(\mathbf{f})} + const$$

- Due to $\sqrt{\ }$ <u>cannot</u> Compute Expectation

Image Restoration: Bounded Variational Inference

- Bypass difficulty: Maximize a <u>Lower Bound</u> of Variational Bound

- Use Upper Bound: $f(w)=\sqrt{w} \le \dfrac{w+u}{2\sqrt{u}} = g(u,w), \forall u>0$

- Bound gets "tight": $f(w)=g(u^*,w),\ u^*=w$

Image Restoration: Bounded Variational Inference

- Define Function

$$p(\mathbf{f}\mid\alpha) \ge M(\mathbf{f},\alpha,\mathbf{u}) = \alpha^{N/2}\exp\left\{-a\sum_{i=1}^{N}\frac{\overbrace{\left([\mathbf{Q}_1\mathbf{f}]_i\right)^2+\left([\mathbf{Q}_2\mathbf{f}]_i\right)^2}^{w}+u_i}{2\sqrt{u_i}}\right\}$$

$$p(\mathbf{g},\mathbf{f},\alpha;\theta) \ge p(\mathbf{g}/\mathbf{f})M(\mathbf{f},\alpha,\mathbf{u})p(\alpha) = \tilde{M}(\mathbf{g},\mathbf{f},\mathbf{u},\alpha;\theta)$$

- <u>Lower Bound</u> of Variational Bound

$$F(q,\theta) \ge \int q(\mathbf{f},\alpha)\ln\frac{\tilde{M}(\mathbf{g},\mathbf{f},\mathbf{u},\alpha;\theta)}{q(\mathbf{f},\alpha)}d\mathbf{f}d\alpha = F^b(q,\mathbf{u},\theta)$$

Image Restoration: Bounded Variational Inference

- <u>VE-Step</u>

$$\left[q^{(t+1)}(\mathbf{f}),q^{(t+1)}(\alpha)\right] = \arg\max_{q(\mathbf{f}),q(\alpha)}\left[F^b(q(\mathbf{f}),q(\alpha),\mathbf{u}^{(t)},\theta^{(t)})\right]$$

- <u>VM-Step</u>

"Bound Tightening"

$$\mathbf{u}^{(t+1)} = \arg\max_{\mathbf{u}}\left[F^b(q^{(t+1)}(\mathbf{f}),q^{(t+1)}(\alpha),\mathbf{u},\theta^{(t)})\right]$$

$$\theta^{(t+1)} = \arg\max_{\theta}\left[F^b(q^{(t+1)}(\mathbf{f}),q^{(t+1)}(\alpha),\mathbf{u}^{(t+1)},\theta)\right]$$

Image Restoration: Bounded Variational Inference

- "Tightest" Bound

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

$$\mathbf{u}_i^{(t+1)} = \left\langle\left([\mathbf{Q}_1\mathbf{f}]_i\right)^2+\left([\mathbf{Q}_2\mathbf{f}]_i\right)^2\right\rangle_{q^{(t+1)}(\mathbf{f})}$$

$$=\left[\mathbf{Q}_1\boldsymbol{\mu}^{(t+1)}\right]_i^2+\left[\mathbf{Q}_2\boldsymbol{\mu}^{(t+1)}\right]_i^2+\left[\mathbf{Q}_1\boldsymbol{\Sigma}^{(t+1)}\mathbf{Q}_1^T\right]_{ii}+\left[\mathbf{Q}_2\boldsymbol{\Sigma}^{(t+1)}\mathbf{Q}_2^T\right]_{ii}, i=1,2\ldots N$$

with $\quad q^{(t+1)}(\mathbf{f}) = N\left(f;\boldsymbol{\mu}^{(t+1)},\boldsymbol{\Sigma}^{(t+1)}\right)$

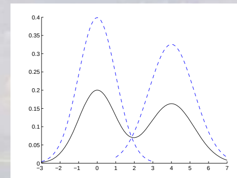
- $\mathbf{u}_i$ Captures Local <u>Spatial Activity</u>

Example:
Gaussian Mixture Models

Gaussian Mixture Models

- $p(\mathbf{x};\boldsymbol{\theta}) = \sum_{j=1}^{M} \pi_j \, N(\mathbf{x};\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)$
  - Model any pdf
  - Soft clustering
- Parameters

  $\boldsymbol{\theta} = \{\pi_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\}_{j=1,\ldots,M}$



- Maximum Likelihood Estimation Difficult

$$\boldsymbol{\theta}_{ML} = \arg\max_{\boldsymbol{\theta}} \log \sum_{j=1}^{M} \pi_j \, N(\mathbf{x};\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)$$

Gaussian Mixture Models: Data Generation Mechanism

- Introduce binary hidden variable $\mathbf{z}$
  1. Select component j

     $z_j = 1, \mathbf{z} = (0,\ldots,0,1,0,\ldots,0)$

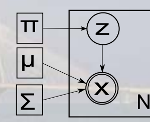
     $\mathbf{z} \sim p(\mathbf{z};\boldsymbol{\pi}) = \prod_{j=1}^{M} \pi_j^{z_j}$ multinomial
  2. Generate sample from selected component

     $x \sim p_j(\mathbf{x}) = N(\mathbf{x};\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)$

Gaussian Mixture Models: Data Generation Mechanism

- Joint pdf

$$p(\mathbf{x},\mathbf{z};\boldsymbol{\theta}) = p(\mathbf{x}\,|\,\mathbf{z})p(\mathbf{z}) = \prod_{j=1}^{M} N(\mathbf{x};\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)^{z_j} \left(\pi_j\right)^{z_j}$$



- Marginal pdf

$$p(\mathbf{x};\boldsymbol{\theta}) = \sum_{\mathbf{z}} p(\mathbf{x},\mathbf{z};\boldsymbol{\theta}) = \sum_{j=1}^{M} \pi_j N\left(\mathbf{x};\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j\right)$$

Gaussian Mixture Models: Posterior

- Posterior (responsibility) can be computed analytically

$$P(z_j = 1\,|\,\mathbf{x}) = \frac{p(\mathbf{x}\,|\,z_j = 1)p(z_j = 1)}{p(\mathbf{x})} = \frac{\pi_j \, N(\mathbf{x}\,|\,\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)}{\sum_{l=1}^{M} \pi_l \, N(\mathbf{x}\,|\,\boldsymbol{\mu}_l,\boldsymbol{\Sigma}_l)}$$

$$p(\mathbf{z}\,|\,\mathbf{x};\boldsymbol{\theta}) = \prod_{j=1}^{M} \left[\frac{\pi_j \, N(\mathbf{x}\,|\,\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)}{\sum_{l=1}^{M} \pi_l \, N(\mathbf{x}\,|\,\boldsymbol{\mu}_l,\boldsymbol{\Sigma}_l)} \right]^{z_j}$$

Gaussian Mixture Models: Parameter Estimation

- Maximum likelihood

$$\boldsymbol{\theta}_{ML} = argmax_{\boldsymbol{\theta}} \log p(\mathbf{X};\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \log \sum_{j=1}^{M} \pi_j \, N(\mathbf{x};\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)$$

- 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

- $p(\mathbf{X},\mathbf{Z};\boldsymbol{\theta}) = \prod_{n=1}^{N}\prod_{j=1}^{M} N(\mathbf{x}_n;\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)^{z_{jn}}\left(\pi_j\right)^{z_{jn}}$
- $Q(\boldsymbol{\theta},\boldsymbol{\theta}^{(t)}) = \left\langle \log p(\mathbf{X},\mathbf{Z};\boldsymbol{\theta})\right\rangle_{p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}^{(t)})}$
  $= \sum_{n=1}^{N}\sum_{j=1}^{M}\left\langle z_{jn}^{(t)}\right\rangle \log \pi_j + \sum_{n=1}^{N}\sum_{j=1}^{M}\left\langle z_{jn}^{(t)}\right\rangle \log N(\mathbf{x}_n;\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)$

Gaussian Mixture Models: Parameter Estimation EM

- <u>E-step</u>

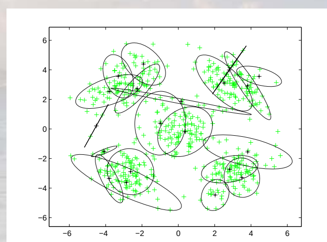
$$\left\langle z_{jn}^{(t)}\right\rangle = \sum_{z_{jn}}\left[\frac{\pi_j^{(t)} N(\mathbf{x}_n;\boldsymbol{\mu}_j^{(t)},\boldsymbol{\Sigma}_j^{(t)})}{\sum_{j=1}^{M}\pi_j^{(t)} N(\mathbf{x}_n;\boldsymbol{\mu}_j^{(t)},\boldsymbol{\Sigma}_j^{(t)})}\right]^{z_{jn}} = \frac{\pi_j^{(t)} N(\mathbf{x}_n;\boldsymbol{\mu}_j^{(t)},\boldsymbol{\Sigma}_j^{(t)})}{\sum_{j=1}^{M}\pi_j^{(t)} N(\mathbf{x}_n;\boldsymbol{\mu}_j^{(t)},\boldsymbol{\Sigma}_j^{(t)})}$$

- <u>M-step</u>

$$\pi_j^{(t+1)} = \frac{1}{N}\sum_{n=1}^{N}\left\langle z_{jn}^{(t)}\right\rangle \qquad \boldsymbol{\mu}_j^{(t+1)} = \frac{\sum_{n=1}^{N}\left\langle z_{jn}^{(t)}\right\rangle \mathbf{x}_n}{\sum_{n=1}^{N}\left\langle z_{jn}^{(t)}\right\rangle} \qquad \boldsymbol{\Sigma}_j^{(t+1)} = \frac{\sum_{n=1}^{N}\left\langle z_{jn}^{(t)}\right\rangle(\mathbf{x}_n-\boldsymbol{\mu}_j^{(t)})(\mathbf{x}_n-\boldsymbol{\mu}_j^{(t)})^T}{\sum_{n=1}^{N}\left\langle z_{jn}^{(t)}\right\rangle}$$

Gaussian Mixture Models: Limitations

- How many components?
- Ill-conditioned covariance matrices

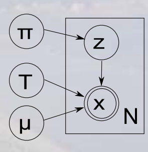


Variational Bayesian Gaussian Mixture Models

- H. Attias, "A Variational Bayesian Framework for Graphical Models", *Proc. NIPS 12*, pp. 209-216, MIT Press, 2000.

Variational Bayesian Gaussian Mixture Models: Prior Distribution

- Treat parameters as hidden variables $\mathbf{h} = \{\mathbf{Z},\boldsymbol{\pi},\boldsymbol{\mu},\mathbf{T}\}$
- Introduce conjugate priors



  - $p(\boldsymbol{\pi}\,|\,\boldsymbol{\alpha}) = \mathrm{Dir}(\boldsymbol{\pi}\,|\,\alpha_1,\ldots,\alpha_M) = \frac{\Gamma(\sum_{j=1}^{M}\alpha_j)}{\prod_{j=1}^{M}\Gamma(\alpha_j)}\prod_{j=1}^{M}\pi_j^{\alpha_j-1}$
  - $p(\boldsymbol{\mu}_j\,|\,\mathbf{T}_j) = N(\boldsymbol{\mu}_j;\boldsymbol{\mu}_0,\beta_0\mathbf{T}_j)$
  - $\mathrm{p}(\mathbf{T}_j\,|\,\nu,\mathbf{V}) = \mathrm{W}(\mathbf{T}_j\,|\,\nu,\mathbf{V}) = \frac{|\mathbf{T}_j|^{(\nu-d-1)/2}\exp tr\left\{-\frac{1}{2}\mathbf{V}\mathbf{T}_j\right\}}{2^{\nu d/2}\pi^{d(d-1)/4}|\mathbf{V}|^{-n/2}\prod_{i=1}^{d}\Gamma((\nu+1-i)/2)}$

Variational Bayesian Gaussian Mixture Models: Mean-Field Approximation

- Exact Bayesian Inference Intractable
  - Variational Mean Field Approximation
- $q(\mathbf{h}) = q_Z(\mathbf{Z})\,q_\pi(\boldsymbol{\pi})\,q_{\mu T}(\boldsymbol{\mu},\mathbf{T})$

  $q_{\mu T}(\boldsymbol{\mu},\mathbf{T}) = \prod_{j=1}^{M} q_\mu(\boldsymbol{\mu}_j\,|\,\mathbf{T}_j)\,q_T(\mathbf{T}_j)$

Variational Bayesian Gaussian Mixture Models: Approximate Posteriors

- $q_Z(\mathbf{Z}) = \prod_{n=1}^{N}\prod_{j=1}^{M} r_{jn}^{z_{jn}}$

$$\log r_{jn} = \langle \log \pi_j \rangle + \langle \log |\mathbf{T}_j| \rangle - \frac{1}{2}(\mathbf{y}_n - \boldsymbol{\mu}_j)^T (\mathbf{y}_n - \boldsymbol{\mu}_j) - \frac{d}{2\beta_j} + const$$

- $q_\pi(\boldsymbol{\pi}) = Dir(\boldsymbol{\pi} \mid \{\lambda_j\})$

$$\lambda_j = \bar{N}_j + \alpha \qquad \bar{N}_j = \sum_{n=1}^{N} z_{jn}$$

Variational Bayesian Gaussian Mixture Models: Approximate Posteriors

- $q_\mu(\boldsymbol{\mu}_j \mid \mathbf{T}) = \prod_{j=1}^{M} N(\boldsymbol{\mu}_j; \mathbf{m}_j, \beta_j \mathbf{T}_j)$

$$\mathbf{m}_j = \left(\bar{N}_j \bar{\boldsymbol{\mu}}_j + \beta^0 \boldsymbol{\mu}^0\right) / \left(\bar{N}_j + \beta^0\right) \qquad \bar{\boldsymbol{\mu}}_j = \frac{1}{\bar{N}_j}\sum_{n=1}^{N} \mathbf{z}_{jn}\mathbf{y}_n$$

$$\beta_j = \bar{N}_j + \beta^0$$

- $q_T(\mathbf{T}) = \prod_{j=1}^{M} W(\mathbf{T}_j; \eta_j, \mathbf{U}_j)$

$$\mathbf{U}_j = \bar{N}_j \boldsymbol{\Sigma}_j + \bar{N}_j \beta^0 (\boldsymbol{\mu}_j - \boldsymbol{\mu}^0)(\boldsymbol{\mu}_j - \boldsymbol{\mu}^0)^T / \left(\bar{N}_j + \beta^0\right) + \mathbf{T}^0$$

$$\eta_j = \bar{N}_j + \nu \qquad \bar{\boldsymbol{\Sigma}}_j = \frac{1}{\bar{N}_j}\sum_{n=1}^{N} \mathbf{z}_{jn} (\mathbf{y}_n - \bar{\boldsymbol{\mu}}_j)(\mathbf{y}_n - \bar{\boldsymbol{\mu}}_j)^T$$

Variational Bayesian Gaussian Mixture Models: Discussion

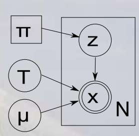
- Select Parameters that define uninformative priors (e.g. $\alpha_j = 1/M$)
- Advantages
  - Disallow singular covariance matrice
  - Bayesian model selection
- Dirichlet distribution for mixing coefficients π disallows pruning of unnecessary components

Variational Bayesian Gaussian Mixture Models: Removing the prior from the mixing weights

- A. Corduneanu and C. Bishop, "Variational Bayesian Model Selection for Mixture Distributions", *Proc. AI and Statistics Conference,* January 2001

Variational Bayesian GMM: Removing the prior from the mixing weights

- Treat π as parameter
- Include <u>M-step</u> to update π



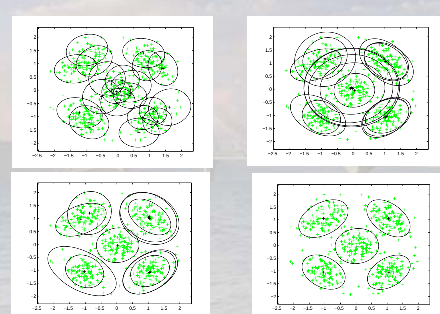
$$\pi_j = \frac{\sum_{n=1}^{N} r_{jn}}{\sum_{k=1}^{M}\sum_{n=1}^{N} r_{kn}}$$

$$r_{jn} = \frac{\tilde{r}_{jn}}{\sum_{k=1}^{M} \tilde{r}_{kn}}$$

$$\tilde{r}_{jn} = \pi_j \exp\left\{\frac{1}{2}\langle \log |\mathbf{T}_j| \rangle - \frac{1}{2} tr\left\{\langle \mathbf{T}_j \rangle \left\{\mathbf{x}_n \mathbf{x}_n^T - \mathbf{x}_n \langle \boldsymbol{\mu}_j \rangle^T + \langle \boldsymbol{\mu}_j \rangle \mathbf{x}_n^T + \langle \boldsymbol{\mu}_j \boldsymbol{\mu}_j^T \rangle\right\}\right\}\right\}$$

- Advantage
  - Eliminates irrelevant components

Variational Bayesian GMM: Example

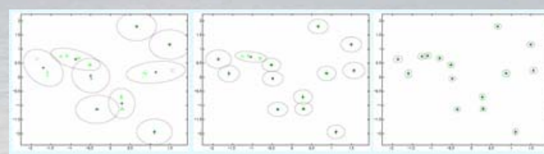


Incremental Variational Bayesian Gaussian Mixture Models

Constantinopoulos C. and Likas, A., "Unsupervised Learning of Gaussian Mixtures Based on Variational Component Splitting", IEEE Trans. on Neural Networks, vol. 18, no. 3, pp. 745-755, 2007.

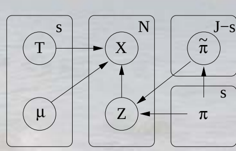
Incremental Variational Bayesian Gaussian Mixture Models

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



Incremental Variational Bayesian Gaussian Mixture Models

- Divide components as ‘fixed’ and ‘free’
- Restrict competition among ‘free’ components only
- $p(\tilde{\boldsymbol{\pi}} \mid \boldsymbol{\pi}) = \left(1-\sum_{j=1}^{s}\pi_j\right)^{-M+s} \frac{\Gamma(\sum_{j=s+1}^{M}\alpha_j)}{\prod_{j=s+1}^{M}\Gamma(\alpha_j)} \prod_{j=s+1}^{M}\left(\frac{\tilde{\pi}_j}{1-\sum_{k=1}^{s}\pi_k}\right)^{\alpha_j-1}$



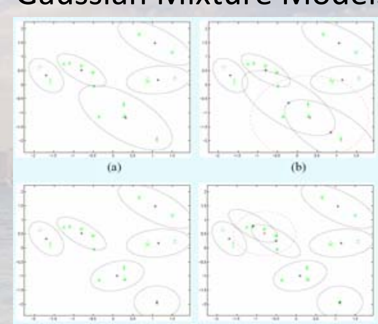
Incremental Variational Bayesian Gaussian Mixture Models

- We start by training a GMM with two components
- At each step:
  - Select a component j
  - Set , $V = d\lambda I$, where λ the max eigenvalue of $\langle T_j^{-1} \rangle$
  - Split the component in two subcomponents
  - Apply VB learning considering the two components as free

Incremental Variational Bayesian Gaussian Mixture Models

- 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

Incremental Variational Bayesian Gaussian Mixture Models



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