# Probabilistic models which enforce sparsity

Ali Mohammad-Djafari
Laboratoire des Signaux et Systèmes,
UMR8506 CNRS-SUPELEC-UNIV PARIS SUD 11
SUPELEC, 91192 Gif-sur-Yvette, France

Email: djafari@lss.supelec.fr
http://djafari.free.fr

# Summary

In this paper, we propose different prior modeling for signals and images which can be used in a Bayesian inference approach in many inverse problems in signal and image processing where we want to infer on sparse signals or images. The sparsity may be directly on the original space or in a transformed space. Here we consider it directly on the original space (impulsive signals).
These models are either simple heavy tailed or hierarchical mixture models.
Depending on the prior model selected, the Bayesian computations (optimization for the Joint Maximum A Posteriori (MAP) estimate or MCMC or Variational Bayes Approximations (VBA) for Posterior Means (PM) or complete density estimation) may become more complex.
We propose these models and drive the corresponding appropriate algorithms, and discuss on their corresponding relative complexities and performances.

# 1. Bayesian inference for inverse problems

- Inverse problems: $\boldsymbol{g} = \boldsymbol{H} \boldsymbol{f} + \boldsymbol{\epsilon}$
- Bayesian inference:

$$p(\boldsymbol{f}|\boldsymbol{g},\boldsymbol{\theta}) = \frac{p(\boldsymbol{g}|\boldsymbol{f},\boldsymbol{\theta}_1)\, p(\boldsymbol{f}|\boldsymbol{\theta}_2)}{p(\boldsymbol{g}|\boldsymbol{\theta})}$$

  with $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$
- Point estimators:
  Maximum A Posteriori (MAP), Posterior Mean (PM)
- Simple prior models: $p(\boldsymbol{f}|\boldsymbol{\theta}_2)$

$$q(\boldsymbol{f},\boldsymbol{\theta}|\boldsymbol{g}) \propto p(\boldsymbol{g}|\boldsymbol{f},\boldsymbol{\theta}_1)\, p(\boldsymbol{f}|\boldsymbol{\theta}_2)\, p(\boldsymbol{\theta})$$

- Prior models with hidden variables: $p(\boldsymbol{f}|\boldsymbol{z},\boldsymbol{\theta}_2)\; p(\boldsymbol{z}|\boldsymbol{\theta}_3)$

$$q(\boldsymbol{f},\boldsymbol{z},\boldsymbol{\theta}|\boldsymbol{g}) \propto p(\boldsymbol{g}|\boldsymbol{f},\boldsymbol{\theta}_1)\, p(\boldsymbol{f}|\boldsymbol{\theta}_2)\, p(\boldsymbol{z}|\boldsymbol{\theta}_3)\, p(\boldsymbol{\theta})$$

# 2. Sparsity enforcing prior models

- Simple heavy tailed models:
  - Generalized Gaussian, Double Exponential
  - Symmetric Weibull, Symmetric Rayleigh
  - Student-t, Cauchy
  - Generalized hyperbolic
  - Elastic net
- Hierarchical mixture models:
  - Mixture of Gaussians
  - Bernoulli-Gaussian
  - Mixture of Gammas
  - Bernoulli-Gamma
  - Mixture of Dirichlet
  - Bernoulli-Multinomial

# 3. Simple heavy tailed models

- Generalized Gaussian, Double Exponential

$$p(\boldsymbol{f}|\gamma,\beta) = \prod_j \mathcal{GG}(f_j|\gamma,\beta) \propto \exp\left\{-\gamma \sum_j |f_j|^\beta\right\}$$

$\beta = 1$ Double exponential or Laplace.
$0 < \beta \leq 1$ are of great interest for sparsity enforcing.

- Symmetric Weibull

$$p(\boldsymbol{f}|\gamma,\beta) = \prod_j \mathcal{W}(f_j|\gamma,\beta) \propto \exp\left\{-\gamma \sum_j |f_j|^\beta + (\beta-1)\log|f_j|\right\}$$

$\beta = 2$ is the Symmetric Rayleigh distribution.
$\beta = 1$ is the Double exponential and
$0 < \beta \leq 1$ are of great interest for sparsity enforcing.

- Student-t and Cauchy models

$$p(\boldsymbol{f}|\nu) = \prod_j \mathcal{S}t(f_j|\nu) \propto \exp\left\{-\frac{\nu+1}{2}\sum_j \log\left(1+f_j^2/\nu\right)\right\}$$

Cauchy model is obtained when $\nu = 1$.

- Elastic net prior model

$$p(\boldsymbol{f}|\nu) = \prod_j \mathcal{E}\mathcal{N}(f_j|\nu) \propto \exp\left\{-\sum_j (\gamma_1|f_j| + \gamma_2 f_j^2)\right\}$$

- Generalized hyperbolic (GH) models

$$p(\boldsymbol{f}|\delta,\nu,\beta) = \prod_j (\delta^2 + f_j^2)^{(\nu-1/2)/2} \exp\{\beta x)\} K_{\nu-1/2}(\alpha\sqrt{\delta^2 + f_j^2})$$

# 4. Mixture models

- Mixture of two Gaussians (MoG2) model

$$p(\boldsymbol{f}|\lambda, v_1, v_0) = \prod_j \left(\lambda \mathcal{N}(f_j|0, v_1) + (1-\lambda)\mathcal{N}(f_j|0, v_0)\right)$$

- Bernoulli-Gaussian (BG) model

$$p(\boldsymbol{f}|\lambda, v) = \prod_j p(f_j) = \prod_j \left(\lambda \mathcal{N}(f_j|0, v) + (1-\lambda)\delta(f_j)\right)$$

- Mixture of Gammas

$$p(\boldsymbol{f}|\lambda, v_1, v_0) = \prod_j \left(\lambda \mathcal{G}(f_j|\alpha_1, \beta_1) + (1-\lambda)\mathcal{G}(f_j|\alpha_2, \beta_2)\right)$$

- Bernoulli-Gamma model

$$p(\boldsymbol{f}|\lambda, \alpha, \beta) = \prod_j \left[\lambda \mathcal{G}(f_j|\alpha, \beta) + (1-\lambda)\delta(f_j)\right]$$

- Mixture of Dirichlets model

$$p(\boldsymbol{f}|\lambda, \boldsymbol{a}_1, \boldsymbol{\alpha}_1, \boldsymbol{a}_2, \boldsymbol{\alpha}_2) = \prod_j \lambda \mathcal{D}(f_j|\boldsymbol{a}_1, \boldsymbol{\alpha}_1) + (1-\lambda)\mathcal{D}(f_j|\boldsymbol{a}_2, \boldsymbol{\alpha}_2)$$

where

$$\mathcal{D}(f_j|\boldsymbol{a}, \boldsymbol{\alpha}) = \prod_{k=1}^{K} \frac{\Gamma(\alpha)}{\Gamma(\alpha_0)\Gamma(\alpha_K)} a_k^{\alpha_k - 1}, \quad \alpha_k \geq 0, \quad a_k \geq 0$$

where $\boldsymbol{a} = \{a_1, \cdots, a_K\}$ and $\boldsymbol{\alpha} = \{\alpha_1, \cdots, \alpha_K\}$
with $\sum_k \alpha_k = \alpha$ and $\sum_k a_k = 1$.

- Bernoulli-Multinomial (BMultinomial) model

$$p(\boldsymbol{f}|\lambda, \boldsymbol{a}, \boldsymbol{\alpha}) = \prod_j \lambda \delta(f_j) + (1-\lambda)\mathcal{M}ult(f_j|\boldsymbol{a}, \boldsymbol{\alpha})$$

# 5. MAP, Joint MAP

- Inverse problems: $\boldsymbol{g} = \boldsymbol{H}\boldsymbol{f} + \boldsymbol{\epsilon}$
- Posterior law:

$$p(\boldsymbol{f}|\boldsymbol{\theta},\boldsymbol{g}) \propto p(\boldsymbol{g}|\boldsymbol{f},\boldsymbol{\theta}_1)\, p(\boldsymbol{f}|\boldsymbol{\theta}_2)$$

- Example: Gaussian noise, Double Exponential prior and MAP:

$$\widehat{\boldsymbol{f}} = \arg\min_{\boldsymbol{f}} \{J(\boldsymbol{f})\} \text{ with } J(\boldsymbol{f}) = \|\boldsymbol{g} - \boldsymbol{H}\boldsymbol{f}\|_2^2 + \lambda\|\boldsymbol{f}\|_1$$

- Full Bayesian: Joint Posterior:

$$p(\boldsymbol{f},\boldsymbol{\theta}|\boldsymbol{g}) \propto p(\boldsymbol{g}|\boldsymbol{f},\boldsymbol{\theta}_1)\, p(\boldsymbol{f}|\boldsymbol{\theta}_2)\, p(\boldsymbol{\theta})$$

- Joint MAP:

$$(\widehat{\boldsymbol{f}},\widehat{\boldsymbol{\theta}}) = \arg\max_{(\boldsymbol{f},\boldsymbol{\theta})} \{p(\boldsymbol{f},\boldsymbol{\theta}|\boldsymbol{g})\}$$

# 6. Marginal MAP and PM estimates

- Marginal MAP: $\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \{p(\boldsymbol{\theta}|\boldsymbol{g})\}$ where

$$p(\boldsymbol{\theta}|\boldsymbol{g}) = \int p(\boldsymbol{f}, \boldsymbol{\theta}|\boldsymbol{g}) \, \mathrm{d}\boldsymbol{f} = \int p(\boldsymbol{g}|\boldsymbol{f}, \boldsymbol{\theta}_1) \, p(\boldsymbol{f}|\boldsymbol{\theta}_2) \, \mathrm{d}\boldsymbol{f}$$

  and then $\widehat{\boldsymbol{f}} = \arg\max_{\boldsymbol{f}} \left\{p(\boldsymbol{f}|\widehat{\boldsymbol{\theta}}, \boldsymbol{g})\right\}$
- Posterior Mean: $\widehat{\boldsymbol{f}} = \int \boldsymbol{f} \, p(\boldsymbol{f}|\widehat{\boldsymbol{\theta}}, \boldsymbol{g}) \, \mathrm{d}\boldsymbol{f}$

- EM and GEM Algorithms

- Variational Bayesian Approximation:
  Approximate $p(\boldsymbol{f}, \boldsymbol{\theta}|\boldsymbol{g})$ by $q(\boldsymbol{f}, \boldsymbol{\theta}|\boldsymbol{g}) = q_1(\boldsymbol{f}|\boldsymbol{g}) \, q_2(\boldsymbol{\theta}|\boldsymbol{g})$
  and then continue computations.

# 7. Hierarchical models and hidden variables

- All the mixture models and some of simple models can be modeled via hidden variables $\boldsymbol{z}$.
- Example 1: MoG model:

$$\left\{\begin{array}{l} p(\boldsymbol{f}|\boldsymbol{z}) = \prod_j p(f_j|z_j) = \prod_j \mathcal{N}\left(f_j|0, v_{z_j}\right) \propto \exp\left\{-\frac{1}{2}\sum_j \frac{f_j^2}{v_{z_j}}\right\} \\ P(z_j = 1) = \lambda, \qquad P(z_j = 0) = 1 - \lambda \end{array}\right.$$

- Example 2: Student-t model

$$\left\{\begin{array}{l} p(\boldsymbol{f}|\boldsymbol{z}) = \prod_j p(f_j|z_j) = \prod_j \mathcal{N}(f_j|0, 1/z_j) \propto \exp\left\{-\frac{1}{2}\sum_j z_j f_j^2\right\} \\ p(z_j|a,b) = \mathcal{G}(z_j|a,b) \propto z_j^{(a-1)} \exp\left\{-b z_j\right\} \text{ with } a = b = \nu/2 \end{array}\right.$$

- With these models we have:

$$p(\boldsymbol{f}, \boldsymbol{z}, \boldsymbol{\theta}|\boldsymbol{g}) \propto p(\boldsymbol{g}|\boldsymbol{f}, \boldsymbol{\theta}_1)\, p(\boldsymbol{f}|\boldsymbol{z}, \boldsymbol{\theta}_2)\, p(\boldsymbol{z}|\boldsymbol{\theta}_3)\, p(\boldsymbol{\theta})$$

# 8. Bayesian Computation and Algorithms

- Often, the expression of $p(\boldsymbol{f}, \boldsymbol{z}, \boldsymbol{\theta} | \boldsymbol{g})$ is complex.
- Its optimization (for Joint MAP) or
  its marginalization or integration (for Marginal MAP or PM)
  is not easy
- Two main techniques:
  MCMC and Variational Bayesian Approximation (VBA)
- MCMC:
  Needs the expressions of the conditionals
  $p(\boldsymbol{f} | \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{g})$, $p(\boldsymbol{z} | \boldsymbol{f}, \boldsymbol{\theta}, \boldsymbol{g})$, and $p(\boldsymbol{\theta} | \boldsymbol{f}, \boldsymbol{z}, \boldsymbol{g})$
- VBA: Approximate $p(\boldsymbol{f}, \boldsymbol{z}, \boldsymbol{\theta} | \boldsymbol{g})$ by a separable one

$$q(\boldsymbol{f}, \boldsymbol{z}, \boldsymbol{\theta} | \boldsymbol{g}) = q_1(\boldsymbol{f})\, q_2(\boldsymbol{f})\, q_3(\boldsymbol{\theta})$$

  and do any computations with these separable ones.

# 9. Conclusions and Perspectives

- We proposed a list of different probabilistic prior models which can be used for sparsity enforcing.
- We classified these models in two categories: simple heavy tails and hierarchical mixture models
- We showed how to use these models for inverse problems where the desired solutions are sparse
- Different algorithms have been developed and their relative performances are compared.
- We use these models for inverse problems in different signal and image processing applications such as:
  - Synthetic Aperture Radar (SAR) Imaging
  - Signal deconvolution in Proteomic and molecular imaging
  - X ray Computed Tomography,
    Diffraction Optical Tomography,
    Microwave Imaging, ...

# 10. Main references


- A. Doucet and P. Duvaut, "Bayesian estimation of state-space models applied to deconvolution of Bernoulli-Gaussian processes," *Signal Processing*, vol. 57, no. 2, 1997.
- T. Park and G. Casella., "The Bayesian Lasso," *Journal of the American Statistical Association*, 2008.
- M. Tipping, "Sparse Bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, 2001.
- C. Févotte and S. Godsill, "A Bayesian aproach for blind separation of sparse source," *IEEE Transactions on Audio, Speech, and Language processing*, 2006.
- J. Griffin and P. Brown, "Inference with normal-gamma prior distributions in regression problems," *Bayesian Analysis*, 2010.
- N. Polson and J. Scott., "Shrink globally, act locally: sparse Bayesian regularization and prediction," *Bayesian Statistics 9*, 2010.
- H. Snoussi and J. Idier., "Bayesian blind separation of generalized hyperbolic processes in noisy and underdeterminate mixtures," *IEEE Trans. on Signal Processing*, 2006.
- P. Williams, "Bayesian regularization and pruning using a Laplace prior," *Neural Computation*, 1995.
- T. Mitchell and J. Beauchamp, "Bayesian variable selection in linear regression," *Journal of the American Statistical Association*, 1988.
- J. R. H. Ishwaran, "Spike and slab variable selection: Frequentist and bayesian strategies," *Annals of Statistics*, 2005.
- J. J. Kormylo and J. M. Mendel, "Maximum-likelihood detection and estimation of Bernoulli-Gaussian processes," vol. 28, pp. 482–488, 1982.
- M. Lavielle, "Bayesian deconvolution of Bernoulli-Gaussian processes," *Signal Processing*, vol. 33, pp. 67–79, 1993.