Ph.D. Thesis Dissertation

SECOND GENERATION SPARSE MODELS

Ignacio Ramirez
nacho@umn.edu

University of Minnesota
Outline

I Introduction (brief)

II Structured sparse models (brief)

III Information-theoretic approach to sparse modeling (in detail)

IV Conclusions and future work
I. Introduction
Learned Sparse Linear Models

State-of-the-art results in several applications when $D$ is learned from the data.
Learned Sparse Linear Models

**Coding** seeks sparse data representation

\[
A^* = \arg\min_A \sum_{j=1}^{n} \frac{1}{2} \|x_j - Da_j\|_2^2 \quad \text{s.t.} \quad \|a_j\|_r \leq \tau, \ \forall \ j,
\]

**Learning** dictionary encourages sparse representations

\[
(A^*, D^*) = \arg\min_{A,D} \sum_{j=1}^{n} \frac{1}{2} \|x_j - Da_j\|_2^2 \quad \text{s.t.} \quad \|a_j\|_r \leq \tau, \ \forall \ j, \\
\|d_k\|_2 \leq 1, \ \forall k,
\]
Burning questions

Q1 What is the best sparse model for given data?
   ▶ How large $D$, how sparse $A$?

Q2 How do we add robustness to the coding/learning process?

Q3 How does one incorporate more prior information?
II. Structured sparse modeling

Q2: How do we add robustness to the coding/learning process?
Q3: How does one incorporate more prior information?

Structured sparse coding


Structured dictionaries


Structured sparse coding: Collaborative HiLasso

▶ Hypothesis: sparsity at group and factor level, same group sparsity

\[
\min_{\mathbf{A} \in \mathbb{R}^{p \times n}} \frac{1}{2} \| \mathbf{Y} - \mathbf{D} \mathbf{A} \|_F^2 + \lambda_2 \sum_{g=1}^{c} \| \mathbf{A}_g \|_F + \lambda_1 \sum_{j=1}^{n,p} | a_{kj} |
\]

▶ Optimization: proximal method + closed form solution to subproblem

▶ Theory: Improved sparse signal recovery conditions
Results

- Source identification
- Source recovery under missing information
- Texture separation (example below)
Structured dictionaries: atom incoherence

- Incoherence: measure of “orthogonality” within a dictionary
- Affects sparse recovery, sparse coding convergence rate
- Reduced indirectly via Frobenius norm of Gram matrix of $\mathbf{D}$

$$
\min_{(\mathbf{A}, \mathbf{D})} \frac{1}{n} \sum_{j=1}^{n} \left[ \frac{1}{2} \| \mathbf{x}_j - \mathbf{D} \mathbf{a}_j \|_2^2 + \lambda \psi(\mathbf{a}_j) \right] + \eta \| \mathbf{D}^\top \mathbf{D} \|_F^2,
$$

Results (natural image patches)
- **Five-fold ISTA speedup!**
- **Two-fold sparse recovery success!**
Structured dictionaries: Block incoherence

- Measure of “orthogonality” between dictionaries
- Affects block-sparse recovery (block-sparse coding convergence rate?)
- Reduced indirectly via Frobenius norm of cross-Gram matrices of D

\[
\min_{y, \{D^i\}_{i=1}^c} \sum_{i=1}^c \left\{ L(X^i|D^i) + \eta \sum_{j \neq i} \left\| (D^i)^\top D^j \right\|_F^2 \right\}
\]

- Unsupervised classification error significantly reduced!
  Brodatz (x5), ISOLET (x4), USPS (x0.5), MNIST (x2)
III. Information-theoretic sparse modeling

Q1: What is the best sparse model for given data?
Q2: How do we add robustness to the coding/learning process?
Q3: How does one incorporate more prior information?

Universal sparsity-inducing regularizers


MDL-based sparse modeling


Information-theoretic sparse modeling

Idea: Sparse models as compression tools

- The best model compresses the most
  - more regularity captured from the data

- Code length: universal metric to compare different models
  - Beyond Bayesian: including different model formulations!

- Unifying framework for understanding sparse models, in all its variants, from a common perspective.
Universal sparsity-inducing priors

Goal: bypass critical choice of regularization parameter

\[ a^* = \arg \min_a \left\{ \frac{1}{2} \| x - Da \|^2 + \lambda \| a \|_1 \right\} \]

\[ \lambda = \frac{2 \sigma^2}{\theta} \text{ after factorization, } \sigma^2 \text{ known, } \theta \text{ unknown} \]
Universal sparsity-inducing priors

**Goal:** bypass critical choice of regularization parameter

\[
a^* = \arg\min_a \left\{ \frac{1}{2} \|x - Da\|^2 + \lambda \|a\|_1 \right\}
\]

Reinterpreted as a **codelength minimization problem**

\[
= \arg\min_a \left\{ -\log P(x|a, \sigma^2) - \log P(a|\theta) \right\}
\]

\[
= \arg\min_a \left\{ e^{-\frac{1}{2\sigma^2} \|x - Da\|^2} \right\} \left\{ e^{-\theta \|a\|_1} \right\}
\]

\[
\lambda = 2\sigma^2 \theta \text{ after factorization, } \sigma^2 \text{ known, } \theta \text{ unknown}
\]
Universal sparsity-inducing priors

**Goal:** bypass critical choice of regularization parameter

\[ a^* = \arg \min_a \left\{ \frac{1}{2} \| x - Da \|^2 + \lambda \| a \|_1 \right\} \]

Reinterpreted as a codelength minimization problem

\[ = \arg \min_a \left\{ - \log P(x \mid a, \sigma^2) - \log P(a \mid \theta) \right\} \]

\[ = \frac{L(x \mid a)}{L(a)} \approx \text{arg min}_{\theta} \left\{ - \log P(a \mid \theta) \right\} , \quad \forall a \]

\( \lambda = 2\sigma^2\theta \) after factorization, \( \sigma^2 \) known, \( \theta \) unknown

**Universal probability models:** replace \( P(a \mid \theta) \) for \( Q(a) \) so that

\[ - \log Q(a) \approx \arg \min_{\theta} \left\{ - \log P(a \mid \theta) \right\} , \quad \forall a \]
Universal sparsity-inducing priors

Goal: bypass critical choice of regularization parameter

\[ a^* = \arg \min_a \{ \frac{1}{2} \| x - Da \|^2 + \lambda \| a \|_1 \} \]

Reinterpreted as a code-length minimization problem

\[ \begin{aligned} &\arg \min_a \left\{ - \log P(x|a, \sigma^2) - \log P(a|\theta) \right\} \\ &= \arg \min_a \left\{ - \log \frac{L(x|a)}{L(a)} \right\} \end{aligned} \]

\[ \begin{aligned} &e^{-\frac{1}{2\sigma^2} \| x - Da \|^2} e^{-\theta \| a \|_1} \\ &\lambda = 2\sigma^2 \theta \text{ after factorization, } \sigma^2 \text{ known, } \theta \text{ unknown} \]

Our choice (among possible dictated by theory): convex mixture

\[ Q(a) = \int_{\Theta} P(a|\theta)w(\theta)d\theta , \quad w(\theta) \geq 0, \int_{\theta} w(\theta)d\theta = 1 \]
Universal sparsity-inducing priors

Goal: bypass critical choice of regularization parameter

\[ a^* = \arg\min_a \left\{ \frac{1}{2} \| x - Da \|^2 + \lambda \| a \|_1 \right\} \]

Reinterpreted as a code-length minimization problem

\[ = \arg\min_a \left\{ -\log P(x \mid a, \sigma^2) - \log P(a \mid \theta) \right\} \]

\[ = \arg\min_a \left\{ -\log \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2} \| x - Da \|^2_2} - \theta \| a \|_1 \right\} \]

\[ \lambda = 2\sigma^2\theta \text{ after factorization, } \sigma^2 \text{ known, } \theta \text{ unknown} \]

Example \( w(\theta) = \Gamma(\theta; \kappa, \beta) \) (conjugate prior of Laplacian)

\[ Q(a; \kappa, \beta) = \frac{1}{2} \kappa \beta^\kappa (|a| + \beta)^{-(\kappa+1)} \]

\( (\kappa, \beta) \) are non-informative parameters
Universal sparsity-inducing priors

**Goal:** bypass critical choice of regularization parameter

\[
\mathbf{a}^* = \arg \min_{\mathbf{a}} \left\{ \frac{1}{2} \| \mathbf{x} - D\mathbf{a} \|^2 + \lambda \| \mathbf{a} \|_1 \right\}
\]

Reinterpreted as a codelength minimization problem

\[
= \arg \min_{\mathbf{a}} \left\{ - \log P(\mathbf{x} \mid \mathbf{a}, \sigma^2) - \log P(\mathbf{a} \mid \theta) \right\}
\]

\[
e^{-\frac{1}{2\sigma^2} \| \mathbf{x} - D\mathbf{a} \|^2} e^{-\theta \| \mathbf{a} \|_1}
\]

\[\lambda = 2\sigma^2\theta\] after factorization, \(\sigma^2\) known, \(\theta\) unknown

Resulting sparse coding problem

\[
\mathbf{a}^* = \arg \min_{\mathbf{a}} \frac{1}{2\sigma^2} \| \mathbf{x} - D\mathbf{a} \|^2 + (\kappa + 1) \sum_{k=1}^{p} \log (|a_k| + \beta)
\]

\(-\log Q(\mathbf{a})\) (non-convex!)
Universal sparsity-inducing priors

Goal: bypass critical choice of regularization parameter

\[ \mathbf{a}^* = \arg \min_{\mathbf{a}} \left\{ \frac{1}{2} \| \mathbf{x} - \mathbf{D} \mathbf{a} \|_2^2 + \lambda \| \mathbf{a} \|_1 \right\} \]

Reinterpreted as a codelength minimization problem

\[ \mathbf{a}^* = \arg \min_{\mathbf{a}} \left\{ -\log P(\mathbf{x} | \mathbf{a}, \sigma^2) - \log P(\mathbf{a} | \theta) \right\} \]

\[ \lambda = 2\sigma^2 \theta \text{ after factorization, } \sigma^2 \text{ known, } \theta \text{ unknown} \]

Results: significant improvements over LASSO formulation:

- 3x sparse recovery success rate
- Up to 4dB improvement in image denoising
Q1: What is the best sparse model for given data?

**Model selection**: find best model $M^* \in \mathcal{M}$ for given data

- **In general**: Cost function = goodness-of-fit + model complexity
- **Traditional model selection**: AIC [Akaike 1974], BIC [Schwartz 1978]
  - risk-based: heavy assumptions, only valid for models of the same type
- **Minimum Description Length (MDL) model selection**
  - compression-based: milder assumptions, can compare models of any type
The MDL principle for model selection [Rissanen, 1978, 1984]

- Capture regularity = compress
- cost function = codellength $L(X)$
MDL in a nutshell

The MDL principle for model selection [Rissanen, 1978, 1984]
- Capture regularity = compress
- cost function = codelength $L(X)$

Main result

\[
\hat{L}(X) = L(X|M) + L(M)
\]

- **Stochastic complexity**: depends on particularities of the data $X$.
- **Parametric complexity**: inevitable cost due to the flexibility of the model class $M$. Data independent.
Framework components

▶ Quantization: \( L(X) \) must be **finite**, shorter than trivial desc.

▶ Probability assignment: deal with unknown parameters

\[
P(E, A, D) = P(E|A, D) \, P(A|D) \, P(D)
\]

▶ Codelength assignment: must be efficient to compute

\[
L(X) = L(E, A, D) = L(E|A, D) + L(A|D) + L(D)
\]

▶ Algorithms:
  - Efficient model search within \( M \)
  - Efficient computation of \( L(X) \)
Residual model

Random error variable $E$ sum of two effects:

- Random noise: Gaussian $\text{Gaussian}(0, \sigma_e^2)$ ($\sigma_e$ known)
- Model deviation: Laplacian $\text{Laplace}(\theta_a)$ ($\theta_e$ unknown)

Result: Laplacian+Gaussian convolution distribution, $E \sim \text{LG}(\sigma_e, \theta_e)$

- $L(E) = -\log \text{LG}(E)$ $M$-type estimator

Unknown parameters: universal mixture/two-parts code $Q(E)$

- $L(E) = -\log Q(E)$ non-convex $M$-type estimator
Coefficients model

- Coeff. vector $a$ quantized to precision $\delta_a$ ($\delta_a = 1$ in the figure)
- Decomposed into three parts: $a \rightarrow (z, s, v)$
  - Sparsity: Bernoulli support vector $z$ (unknown parameter)
  - Symmetry: Symmetric vector $s$
  - Laplacian magnitude vector $v$ (unknown parameter)
Dictionary model

- Piecewise smooth → predictable
- Atom prediction residuals $b_k = Wd_k$ are encoded
- Causal bilinear prediction operator $W$ maps atoms to residuals

$$L(D) = \approx \theta_d \sum_{k=1}^{p} \|Wd_k\|_1.$$
Collaborative sequential encoding

- **Assumption:** same unknown model parameters for every col. in \( X \)
- **Sequential encoding:** parameters learned from past samples
  - \( L(e), L(v) \) become convex, faster to compute
  - This is also a Universal model (plug-in sequential)!
- **Spatial correlation:** Markov dependency on supports

\[
\begin{align*}
\rho^k(4) &= \frac{z_{k_1} + z_{k_2} + z_{k_3} + 1}{3 + 1} \\
\theta^k_\alpha(4) &= \frac{v_{k_1} + v_{k_2} + v_{k_3}}{z_{k_1} + z_{k_2} + z_{k_3}} \\
\theta^k_c(4) &= \sqrt{\frac{1}{2} \left( \frac{e_{k_1}^2 + e_{k_2}^2 + e_{k_3}^2}{3} - \sigma_c^2 \right)}
\end{align*}
\]
Coding algorithms

- Fixed dictionary $D$, data sample $a$.
- Goal: minimize $L(x|M)$, $M \in \mathcal{M}$,

$$\mathcal{M} = \bigcup_{\gamma} \mathcal{M}(\gamma), \quad \mathcal{M}(\gamma) = \{a \in \mathbb{R}^p, \|a\|_0 \leq \gamma\}, \gamma = 0, \ldots, p$$
Coding algorithms

- Fixed dictionary \( D \), data sample \( a \).
- Goal: minimize \( L(x|M) \), \( M \in \mathcal{M} \),

\[
\mathcal{M} = \bigcup_\gamma \mathcal{M}(\gamma), \quad \mathcal{M}(\gamma) = \{ a \in \mathbb{R}^p, \|a\|_0 \leq \gamma \}, \gamma = 0, \ldots, p
\]

- COdelengt-MInimizing Pursuit Algorithm (COMPA):
  - Forward-stepwise selection: direction that maximizes \( \Delta L(x|a) \)

- COdelengt-MInimizing Continuation Algorithm (COMICA):
  - Convex relaxation of \( L(x|M) \), solved using CD
  - Candidate supports \( z \) estimated via homotopy/continuation
  - Candidate coefficient values \( v \) estimated via OLS
Coding algorithms

- Fixed dictionary $\mathbf{D}$, data sample $\mathbf{a}$.
- Goal: minimize $L(\mathbf{x}|M)$, $M \in \mathcal{M}$,

\[
\mathcal{M} = \bigcup_{\gamma} \mathcal{M}(\gamma), \quad \mathcal{M}(\gamma) = \{ \mathbf{a} \in \mathbb{R}^p, \|\mathbf{a}\|_0 \leq \gamma \}, \quad \gamma = 0, \ldots, p
\]

- COdelength-MInimizing Pursuit Algorithm (COMPA):
  - Forward-stepwise selection: direction that maximizes $\Delta L(\mathbf{x}|\mathbf{a})$

- COdelength-MInimizing Continuation Algorithm (COMICA):
  - Convex relaxation of $L(\mathbf{x}|\mathbf{a})$, solved using CD
  - Candidate supports $\mathbf{z}$ estimated via homotopy/continuation
  - Candidate coefficient values $\mathbf{v}$ estimated via OLS
Coding results

Evolution of COMPA for a given sample

- Efficient coding speed similar to OMP/LARS/CD
- Compression: 4.1bpp for 8bpp grayscale images. Rivals PNG!
Learning algorithm

- For given data set $X$, choose model $M = (A, D)$ in

$$M = \bigcup_{p=0}^{+\infty} M(p), \quad M(p) = \{ (A, D) : D \in \mathbb{R}^{m \times p} \}$$

that minimizes $L(x|M) = L(x|a) + L(a|D) + L(D)$

- Forward-selection of atoms. Start with null dictionary $D = []$.

- New atom = principal component

- Train using alternate minimization:
  - Coding stage done using COMPA/COMICA
  - Dictionary update: $\ell_1$-regularized update, solved using FISTA

$$D^t = \arg \min_D L(X - DA^t|\theta_e^t, \sigma_e^2) + \theta_d^t \sum_{k=1}^p \|W_d k\|_1.$$  

- Plug-in $(\theta_e, \theta_d)$ estimated from prev. iteration (EM like)
Denoising results

(L) Noisy “Peppers”, AWGN with $\sigma = 20$. (R) Recovered image (31.5dB). Reference: 30.8dB (K-SVD [Aharon et al., 2006]), 29.4dB (MDL denoising [Roos et al., 2009])

State of the art denoising. Parameter free!
Texture segmentation

- A dictionary for each texture, $D^i$.
- Classification: patch $x$ assigned to class $\hat{i}$,

$$\hat{i} = \arg\min_i \{L(x, D^i) + L(D^i)\}$$

95.4% correct. Parameter free!
Low-rank matrix decomposition

- MDL for low-rank model selection: choose best approximation rank
  - Candidate model \((U, \Sigma, V, E)\): RPCA + continuation on \(\lambda\)
    \[
    U\Sigma V^T = \arg \min_W \|W\|_* + \lambda \|Y - W\|_1, \quad E = X - U\Sigma V^T
    \]

- Task: dynamic background recovery in video sequences

Videos: Shopping Mall | Lobby

State of the art results, parameter-free.
IV. Conclusions and future work
Conclusions and future work

Conclusions

▶ Novel structured sparse modeling formulations:
  ▶ efficient optimization
  ▶ theoretical guarantees
  ▶ good results in practice

▶ Information-theoretic sparse modeling framework:
  ▶ New perspective to understand sparse models
  ▶ Solves problem of choosing critical model parameters
  ▶ State of the art (or close) in tested applications
  ▶ Computationally efficient, compares to traditional methods

Future work

▶ Extend MDL framework to structured sparse models
▶ Enlarge the models by including other variables (patch width?)
▶ Extend the concept of sparsity
Acknowledgements

▶ Boss: Guillermo Sapiro
▶ Committee: Tom Luo, Nikos Papanikolopoulos, Yousef Saad
▶ Colleagues (and friends): Fefo Lecumberry, Gonchi Mateos, Pablo Sprechmann, Alexey Castrodad, Gadiel Seroussi, Juan Bazerque
▶ Friends: Julia Demasi, Ana Inés Torres, Heng Chooi, Tal & Niki Makovsky, Ilan & Yasmin Sagiv
▶ Family: my wife Pati Vaz, my kids Joaquin & Violeta, my parents Mario and Beatriz
▶ Dedicated to the memory of my mother-in-law, Amalia Jauri
Bibliography


