Latent Variable Bayesian Models for Promoting Sparsity

David Wipf

Biomagnetic Imaging Lab
University of California, San Francisco

Compressive Sensing Workshop, Duke University, Feb. 2009
Overview

♦ Sparse inverse problems

♦ Latent variable representations of sparse priors

♦ Two approaches to sparse estimation
  ♦ **Type I**: Integrate out latent variables, maximize over coefficients (MAP)
  ♦ **Type II**: Integrate out coefficients, maximize over latent variables (empirical Bayes)

♦ Duality

♦ Properties of Type II cost function

♦ Optimization strategies, e.g., reweighted $L_1$ and $L_2$

♦ Empirical Results
Sparse Inverse Problem

- Linear generative model:

\[ y = \Phi x + \varepsilon \]

- **Objective**: Estimate the unknown \( x \) given the following assumptions:

1. \( \Phi \) is *overcomplete*, meaning the number of features (columns) \( n \) is greater than the signal dimension \( m \).

2. \( x \) is *maximally sparse*, i.e., many elements equal zero.
Sparse Inverse Problem Cont.

- Noiseless case ($\varepsilon = 0$):
  \[
  x_0 \triangleq \arg \min_x \|x\|_0 \quad \text{s.t.} \quad y = \Phi x
  \]
  
  $L_0$ quasi-norm = # of nonzeros in $x$

- Noisy case ($\varepsilon > 0$):
  \[
  x_0(\lambda) \triangleq \arg \min_x \|y - \Phi x\|_2^2 + \lambda \|x\|_0
  \]
  \[
  = \arg \max_x \exp\left[-\frac{1}{2\lambda} \|y - \Phi x\|_2^2\right] \exp\left[-\frac{1}{2} \|x\|_0\right]
  \]

  likelihood prior
Sparse Inverse Problem Cont.

- Forward model is linear, the inverse problem is very difficult to solve for two reasons:
  1. Combinatorial number of local minima
  2. Objective is discontinuous

- A variety of approximate methods can be viewed in Bayesian terms using a flexible class of sparse priors.
Latent Variable Representations of Sparse Priors

1. Gaussian scale mixture:
   \[ p(x_i) = \int N(x_i|0, \gamma_i) p(\gamma_i) d\gamma_i \propto \exp\left[-\frac{1}{2} g\left(x_i^2\right)\right] \]

2. Convexity-based representation:
   \[ p(x_i) = \sup_{\gamma_i \geq 0} N(x_i|0, \gamma_i) \varphi(\gamma_i) \propto \exp\left[-\frac{1}{2} g\left(x_i^2\right)\right] \]

Properties

- Essentially all sparse priors can be represented in both forms [Palmer et al., 2006].
- For non-negative functions \( p(\gamma_i) \) and \( \varphi(\gamma_i) \), resulting \( g\left(x_i^2\right) \) will be non-decreasing, concave (favors sparsity).
Examples

\[ g(x_i^2) = \log(x_i^2 + \varepsilon), \quad \text{[Chartrand and Yin, 2008; Tipping, 2001]} \]
\[ g(x_i^2) = \log(|x_i| + \varepsilon), \quad \text{[Candes et al., 2008]} \]
\[ g(x_i^2) = |x_i|^p, \quad \text{[Leahy and Jeffs, 1991; Rao and Kreutz-Delgado, 1999]} \]
Type I (MAP Estimation)

Integrate out the *latent variables* $\gamma$, maximize over the *coefficients* $x$.

$$ x^{(1)} \triangleq \arg \max_x \int p(y \mid x) \prod_i N(x_i \mid 0, \gamma_i) p(\gamma_i) d\gamma_i $$

$$ = \arg \min_x -\log p(y \mid x) - \log p(x) $$

$$ = \arg \min_x \|y - \Phi x\|_2^2 + \lambda \sum_{i=1}^n g(x_i^2) $$

*data fit* \hspace{1cm} *nondecreasing, concave penalty*
Type I Cont.

♦ Convenient Optimization ⇒ Iterative reweighted $L_1, L_2$.

♦ Examples:
  ♦ Candes et al. (2008)
  ♦ Chartrand and Yin (2008)
  ♦ Figueiredo et al. (2007)
  ♦ Rao et al. (2003)

♦ Potential Limitations:
  ♦ If $g(x_i^2)$ is too sparse, problem is not convex.
  ♦ If it is not sparse enough, then global minimum may not be sparse enough.
Type II (Empirical Bayes)

Integrate out the *coefficients* $\mathbf{x}$, maximize over the *latent variables* $\gamma$.

$$
\gamma^{(II)} \triangleq \underset{\gamma}{\text{arg max}} \; \int p(\mathbf{y} \mid \mathbf{x}) \prod_{i} N(x_{i} \mid 0, \gamma_{i}) p(\gamma_{i}) d\mathbf{x}_{i}
$$

$$
= \underset{\gamma}{\text{arg min}} \; \log |\lambda I + \Phi \Gamma \Phi^{T}| + \mathbf{y}^{T} \left(\lambda I + \Phi \Gamma \Phi^{T}\right)^{-1} \mathbf{y} + \sum_{i} f(\gamma_{i})
$$

where

$$
\Gamma \triangleq \text{diag}[\gamma]
$$

$$
f(\gamma_{i}) \triangleq -2\log p(\gamma_{i})
$$

Given $\gamma^{(II)}$, can easily get a point estimate for $\mathbf{x}$ using

$$
\mathbf{x}^{(II)} \triangleq \mathbf{E}[\mathbf{x} \mid \mathbf{y}, \gamma^{(II)}] = \Gamma^{(II)} \Phi^{T} \left(\lambda I + \Phi \Gamma^{(II)} \Phi^{T}\right)^{-1} \mathbf{y}
$$
Type II Cont.

- Convenient Optimization $\Rightarrow$ Iterative reweighted $L_1, L_2$.

- Examples:
  - Bishop and Tipping (2000)
  - Girolami (2001)
  - Neal (1996)
  - Sato et al. (2004)
  - Tipping (2001)
  - Wipf and Nagarajan (2008)

- Potential Limitations:
  - Typically non-convex cost function.
  - Unclear how to choose $f(\gamma_i)$ to get maximal sparsity.
Outstanding Issues

- What is the exact relationship between Type I and Type II?

- Duality:
  - Type I can be expressed as an equivalent problem in $\gamma$-space.
  - Type II can be expressed as an equivalent problem in $x$-space.

- So direct comparisons are possible by evaluating in an equivalent space
  - E.g., Type I is a special limiting case of Type II.

- Viewing Type II in $x$-space leads to theoretical insights.
Type II Cost Function in x-Space

Theorem 1

\[ x^{(II)} = E \left[ x \mid y, \gamma^{(II)} \right] = \arg \min_x \| y - \Phi x \|_2^2 + \lambda g^{(II)} \left( x^2 \right) \]

where \( x^2 \triangleq \left[ x_1^2, \ldots, x_n^2 \right]^T \)

\[ g^{(II)} \left( x^2 \right) \triangleq \text{Concave conjugate of } -\log \left| \lambda I + \Phi \Gamma \Phi^T \right| + \sum_i f(\gamma_i) \text{ w.r.t. } \Gamma^{-1} \]
Properties of Type II Penalty

Assume simplest case where $f(\gamma_i) = 0$ [Tipping, 2001].

1. Concave in $|x_i|$ for all $i$ $\Rightarrow$ sparsity-inducing

2. Non-factorial, meaning

$$g^{(\text{III})}(x^2) \neq \sum_i g_i^{(\text{III})}(x_i^2) \Rightarrow \text{better approx. to } L_0 \text{ quasi-norm}$$
Advantages of Non-Factorial Penalty

Theorem 2

In the low noise limit ($\lambda \to 0$), and assuming $\|x_0\| < \text{spark}[\Phi] - 1$, then Type II penalty satisfies:

$$x_0 = \arg \min_x g^{(II)}(x^2) \quad \text{s.t.} \quad y = \Phi x$$

No factorial penalty $g(x^2) = \sum_i g(x_i^2)$ satisfies this condition and has fewer minima than the Type II penalty $g^{(II)}(x^2)$ in the feasible region.
Example of Local Minima Smoothing

♦ Consider when \( y = \Phi x \) has a 1-D feasible region, i.e.,
\[
n = m + 1
\]

♦ Any feasible solution \( x \) will satisfy:
\[
x = x' + \alpha v
\]

\[
v \in \text{Null}(\Phi)
\]

where \( \alpha \) is a scalar
\( x' \) is a fixed solution

♦ Can plot penalty functions vs. \( \alpha \) to view local minima profile
over the 1-D feasible region.
Local Minima Smoothing Example

\[ g^{(III)}(x^2) \]

\[
\sum_i |x_i|^{0.01} \approx \|x\|_0
\]

maximally sparse solution
Conditions For a Single Minimum

Theorem 3

- Assume $\|x_0\| < \text{spark}[\Phi] - 1$. If the magnitudes of the non-zero elements in $x_0$ are sufficiently scaled, then the Type II cost function ($\lambda=0$) has a *single minimum* which is located at $x_0$.

- No possible factorial penalty satisfies this condition.

**Diagram:**

- **Scaled Weights (easy):**
  - $x_0$ with points indicating different magnitudes.

- **Uniform Weights (hard):**
  - $x_0$ with uniform magnitudes.

---

**Explanation:**

- The theorem states that under certain conditions, the Type II cost function with $\lambda=0$ has a single minimum at $x_0$.
- The conditions involve the norm of $x_0$ and the spark of the matrix $\Phi$.
- Scaled weights are easier to satisfy, while uniform weights are harder to satisfy this condition.
Reweighted $L_2$ Implementation of Type II

\[ \mathbf{x}^{(k+1)} \rightarrow \arg \min_{\mathbf{x}} \| \mathbf{y} - \Phi \mathbf{x} \|^2 + \lambda \sum_{i} \frac{x_i^2}{w_i^{(k)}} \]

\[ = W^{(k)} \Phi^T \left( \lambda I + \Phi W^{(k)} \Phi^T \right)^{-1} \mathbf{y} \]

\[ w_i^{(k+1)} \rightarrow (x_i^{(k+1)})^2 + w_i^{(k)} \left[ 1 - w_i^{(k)} \phi_i^T \left( \lambda I + \Phi W^{(k)} \Phi^T \right)^{-1} \phi_i \right] \]

\[ \mathcal{E}_i \]

Many other variants are possible using different majorization-minimization algorithms
Connection with Type I

Equivalent to solving

\[ x^{(II)} = \arg \min_x \|y - \Phi x\|_2^2 + \lambda \sum_i \log(x_i^2 + \epsilon_i) \]
Reweighted $L_1$ Implementation of Type II

$$\mathbf{x}^{(k+1)} \rightarrow \underset{\mathbf{x}}{\text{arg min}} \left\| \mathbf{y} - \Phi \mathbf{x} \right\|_2^2 + \lambda \sum_{i} \frac{|x_i|}{w_i^{(k)}}$$

$$w_i^{(k+1)} \rightarrow \left[ \phi_i^T \left( \lambda \mathbf{I} + \Phi W^{(k)} \text{diag}[\mathbf{x}^{(k+1)}] \Phi^T \right)^{-1} \phi_i \right]^{-\frac{1}{2}}$$
Properties of Reweighted $L_1$ Minimization

- **Globally convergent** [Zangwill 1969]: Guaranteed to locally minimize the Type II objective.

- **Sparsity will not increase** ($\lambda=0$): $\|\hat{x}^{(k+1)}\|_0 \leq \|\hat{x}^{(k)}\|_0 \leq \|x^\text{BP}\|_0$

- **Extensible**: Easy to extend to more general cases by adding constraints to the $x$-update step, e.g., non-negative sparse inverse problems, alternative likelihood models, etc.

- **Fast, Robust**: Even one or two iterations greatly improves upon the performance of the minimum $L_1$-norm solution.
Always Room for Improvement

Theorem 4

- Assume \( \text{spark}(\Phi) = m + 1 \).

- Let \( \mathbf{x}^* \) be any coefficient vector drawn from support \( S \) with cardinality \( |S| < (m + 1)/2 \) such that standard \( L_1 \) minimization fails.

- Then there exists a set of signals \( \mathbf{y} = \Phi \mathbf{x} \), with \( \mathbf{x} \) having support \( S \), such that
  1. Type II reweighted \( L_1 \) always succeeds
  2. Standard \( L_1 \) minimization always fails.
## Empirical Example

<table>
<thead>
<tr>
<th>Penalty</th>
<th>Updates</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sum_i \log \left( x_i^2 + \epsilon \right)$, $L_2$ iters</td>
<td>[Chartrand and Yin, 2008]</td>
</tr>
<tr>
<td>$\sum_i \log \left(</td>
<td>x_i</td>
</tr>
<tr>
<td>$\sum_i \log \left(</td>
<td>x_i</td>
</tr>
<tr>
<td>$g^{(II)} \left( x^2 \right)$, $L_1$ iters</td>
<td>[Wipf and Nagarajan, 2008]</td>
</tr>
<tr>
<td>$g^{(II)} \left( x^2 \right)$, $L_2$ iters</td>
<td>[Bishop and Tipping, 2000; Wipf, 2006]</td>
</tr>
</tbody>
</table>
Empirical Example

- Generate data via $Y = \Phi X_0$:
  - $\Phi$ is 50-by-100 with Gaussian iid entries
  - $X_0$ is 100-by-5 with random nonzero rows, i.e., simultaneous sparse approximation problem [Cotter et al., 2005; Tropp, 2006; Wipf and Rao, 2007].

- Run each algorithm and check if $X_0$ is recovered.
Results \((m = 50, n = 100)\)
Non-Negative Sparse Recovery Example Using Iterative Reweighted $L_1$ (Type II)

$m = 50, n = 100$
30 nonzeros
Results With Different Nonzero Coefficient Distributions ($m = 50, n = 100$)

Approx. Jeffreys Weights

Unit Weights

OMP
BP
Type II
Summary

- Type II methods motivate some non-traditional means of solving underdetermined sparse linear inverse problems.
- Non-factorial penalty functions have some very desirable properties.
- Reweighted $L_1$ and $L_2$ updates reveal
  - some connections between algorithms,
  - often lead to performance improvements, and
  - remove some of the stigma of non-convex cost functions.
Thank You
