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:

$$\mathbf{y} = \mathbf{\Phi x} + \boldsymbol{\varepsilon}$$

Observed *m*-dimensional data vector

Matrix of *n* basis vectors

Vector of Unknown Coefficients

Gaussian noise with variance λ

- **Objective**: Estimate the unknown **x** given the following assumptions:

1. Φ 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 ($\varepsilon = 0$):

$$\mathbf{x}_0 \triangleq \arg\min_{\mathbf{x}} \|\mathbf{x}\|_0 \quad \text{s.t.} \quad \mathbf{y} = \Phi\mathbf{x}$$

L_0 quasi-norm = # of nonzeros in **x**

- Noisy case ($\varepsilon > 0$):

$$\begin{aligned}\mathbf{x}_0(\lambda) &\triangleq \arg\min_{\mathbf{x}} \|\mathbf{y} - \Phi\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_0 \\ &= \arg\max_{\mathbf{x}} \underbrace{\exp\left[-\tfrac{1}{2\lambda}\|\mathbf{y} - \Phi\mathbf{x}\|_2^2\right]}_{\text{likelihood}} \underbrace{\exp\left[-\tfrac{1}{2}\|\mathbf{x}\|_0\right]}_{\text{prior}}\end{aligned}$$

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:

$$p(x_i) = \int N\left(x_i | 0, \gamma_i\right) p(\gamma_i) d\gamma_i \propto \exp\left[-\tfrac{1}{2} g\left(x_i^2\right)\right]$$

2. Convexity-based representation:

$$p(x_i) = \sup_{\gamma_i \geq 0} N\left(x_i | 0, \gamma_i\right) \varphi(\gamma_i) \propto \exp\left[-\tfrac{1}{2} g\left(x_i^2\right)\right]$$

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

$$g\left(x_i^2\right) = \log\left(x_i^2 + \varepsilon\right),$$ [Chartrand and Yin, 2008; Tipping, 2001]

$$g\left(x_i^2\right) = \log\left(\left|x_i\right| + \varepsilon\right),$$ [Candes et al., 2008]

$$g\left(x_i^2\right) = \left|x_i\right|^p,$$ [Leahy and Jeffs, 1991; Rao and Kreutz-Delgado, 1999]

Type I (MAP Estimation)

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

$$\begin{aligned}
\mathbf{x}^{(I)} &\triangleq \arg\max_{\mathbf{x}} \int p(\mathbf{y} \mid \mathbf{x}) \prod_i N(x_i \mid 0, \gamma_i)\, p(\gamma_i)\, d\gamma_i \\
&= \arg\min_{\mathbf{x}} -\log p(\mathbf{y} \mid \mathbf{x}) - \log p(\mathbf{x}) \\
&= \arg\min_{\mathbf{x}} \underbrace{\|\mathbf{y} - \Phi \mathbf{x}\|_2^2}_{\text{data fit}} + \lambda \sum_{i=1}^{n} \underbrace{g\left(x_i^2\right)}_{\text{nondecreasing, concave penalty}}
\end{aligned}$$

Type I Cont.

- Convenient Optimization $\Rightarrow$ 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 $g\left(x_i^2\right)$ 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* $\boldsymbol{\gamma}$.

$$\begin{aligned}\boldsymbol{\gamma}^{(II)} &\triangleq \arg\max_{\boldsymbol{\gamma}} \int p(\mathbf{y}\,|\,\mathbf{x})\prod_i N(x_i\,|\,0,\gamma_i)\,p(\gamma_i)\,dx_i \\ &= \arg\min_{\boldsymbol{\gamma}} \log\left|\lambda I + \Phi\Gamma\Phi^T\right| + \mathbf{y}^T\left(\lambda I + \Phi\Gamma\Phi^T\right)^{-1}\mathbf{y} + \sum_i f(\gamma_i)\end{aligned}$$

$$\text{where}\quad \begin{aligned}\Gamma &\triangleq \operatorname{diag}[\boldsymbol{\gamma}] \\ f(\gamma_i) &\triangleq -2\log p(\gamma_i)\end{aligned}$$

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

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

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 γ-space.
  - Type II can be expressed as an equivalent problem in $\mathbf{x}$-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 $\mathbf{x}$-space leads to theoretical insights.

Type II Cost Function in x-Space

Theorem 1

$$\mathbf{x}^{(II)} \;=\; \mathrm{E}\left[\mathbf{x} \,|\, \mathbf{y}, \boldsymbol{\gamma}^{(II)}\right] \;=\; \arg\min_{\mathbf{x}} \left\|\mathbf{y} - \Phi\mathbf{x}\right\|_2^2 + \lambda g^{(II)}\left(\mathbf{x}^2\right)$$

where $\mathbf{x}^2 \triangleq \left[x_1^2, \ldots, x_n^2\right]^T$

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

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

$$g^{(II)}\left(\mathbf{x}^2\right) \neq \sum_i g_i^{(II)}\left(x_i^2\right) \quad \Rightarrow$$

better approx. to L_0 quasi-norm

Advantages of Non-Factorial Penalty

Theorem 2

- In the low noise limit ($\lambda \rightarrow 0$), and assuming $\|\mathbf{x}_0\| < \text{spark}[\Phi]-1$, then Type II penalty satisfies:

$$\mathbf{x}_0 = \arg\min_{\mathbf{x}} \; g^{(II)}\left(\mathbf{x}^2\right) \quad \text{s.t.} \quad \mathbf{y} = \Phi\mathbf{x}$$

- No factorial penalty $g\left(\mathbf{x}^2\right) = \sum_i g\left(x_i^2\right)$ satisfies this condition *and* has fewer minima than the Type II penalty $g^{(II)}\left(\mathbf{x}^2\right)$ in the feasible region.

Example of Local Minima Smoothing

- Consider when $\mathbf{y} = \Phi\mathbf{x}$ has a 1-D feasible region, i.e.,

$$n = m+1$$

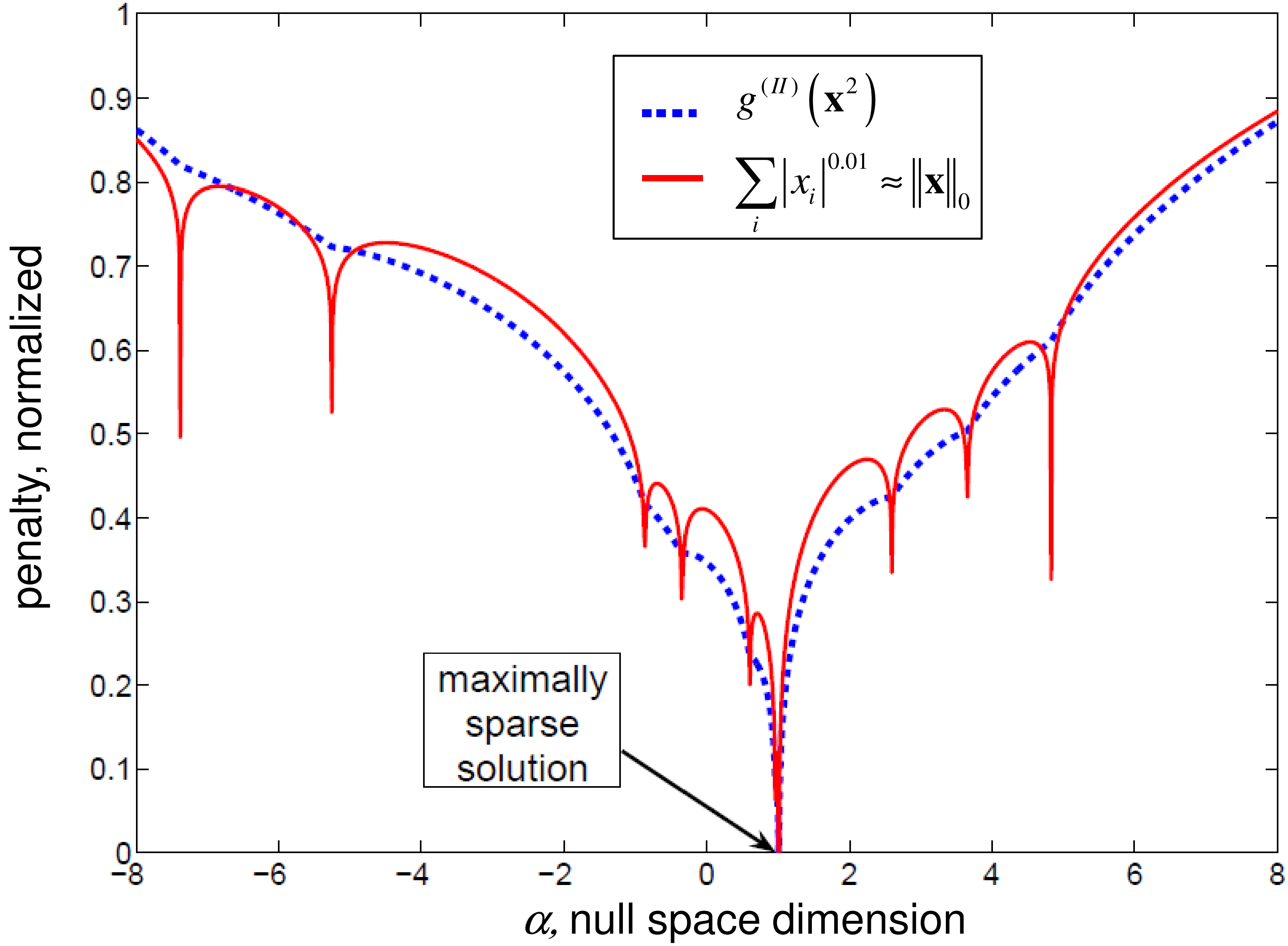
- Any feasible solution $\mathbf{x}$ will satisfy:

$$\mathbf{x} = \mathbf{x}' + \alpha\mathbf{v}$$

$$\text{where} \quad \begin{array}{l} \mathbf{v} \in \mathcal{N}\text{ull}(\Phi) \\ \alpha \text{ is a scalar} \\ \mathbf{x}' \text{ is a fixed solution} \end{array}$$

- Can plot penalty functions vs. α to view local minima profile over the 1-D feasible region.

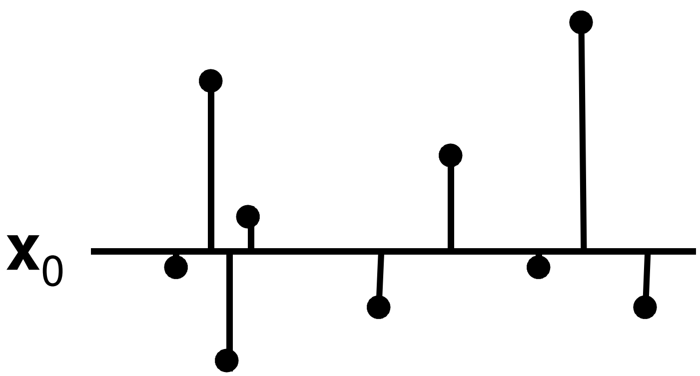
Local Minima Smoothing Example



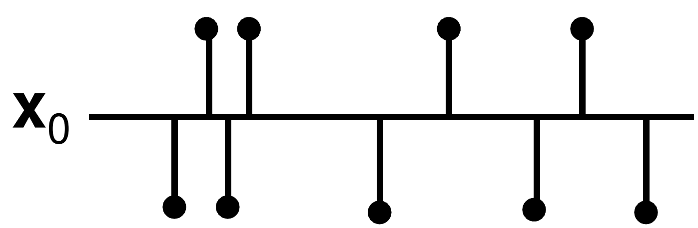
Conditions For a Single Minimum

Theorem 3

- Assume $\|\mathbf{x}_0\| < \text{spark}[\Phi]-1$. If the magnitudes of the non-zero elements in $\mathbf{x}_0$ are sufficiently scaled, then the Type II cost function (λ=0) has a *single minimum* which is located at $\mathbf{x}_0$.



Scaled Weights (easy)



Uniform Weights (hard)

- No possible factorial penalty satisfies this condition.

Reweighted L_2 Implementation of Type II

$$\mathbf{x}^{(k+1)} \rightarrow \arg\min_{\mathbf{x}} \;\|\mathbf{y}-\Phi\mathbf{x}\|_2^2 + \lambda\sum_i \frac{x_i^2}{w_i^{(k)}}$$

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

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

Many other variants are possible using different majorization-minimization algorithms

Connection with Type I

Equivalent to solving

$$\mathbf{x}^{(II)} = \arg\min_{\mathbf{x}} \left\| \mathbf{y} - \Phi \mathbf{x} \right\|_2^2 + \lambda \sum_i \log\left(x_i^2 + \varepsilon_i \right)$$

adaptive

Reweighted L_1 Implementation of Type II

$$\mathbf{x}^{(k+1)} \rightarrow \arg\min_{\mathbf{x}} \; \left\|\mathbf{y} - \Phi\mathbf{x}\right\|_2^2 + \lambda \sum_i \frac{|x_i|}{w_i^{(k)}}$$

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

Properties of Reweighted L_1 Minimization

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

- ***Sparsity will not increase (λ=0)***: $\left\| \hat{\mathbf{x}}^{(k+1)} \right\|_0 \leq \left\| \hat{\mathbf{x}}^{(k)} \right\|_0 \leq \left\| \mathbf{x}^{\mathrm{BP}} \right\|_0$

- ***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 spark(Φ) = $m + 1$.
- 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.

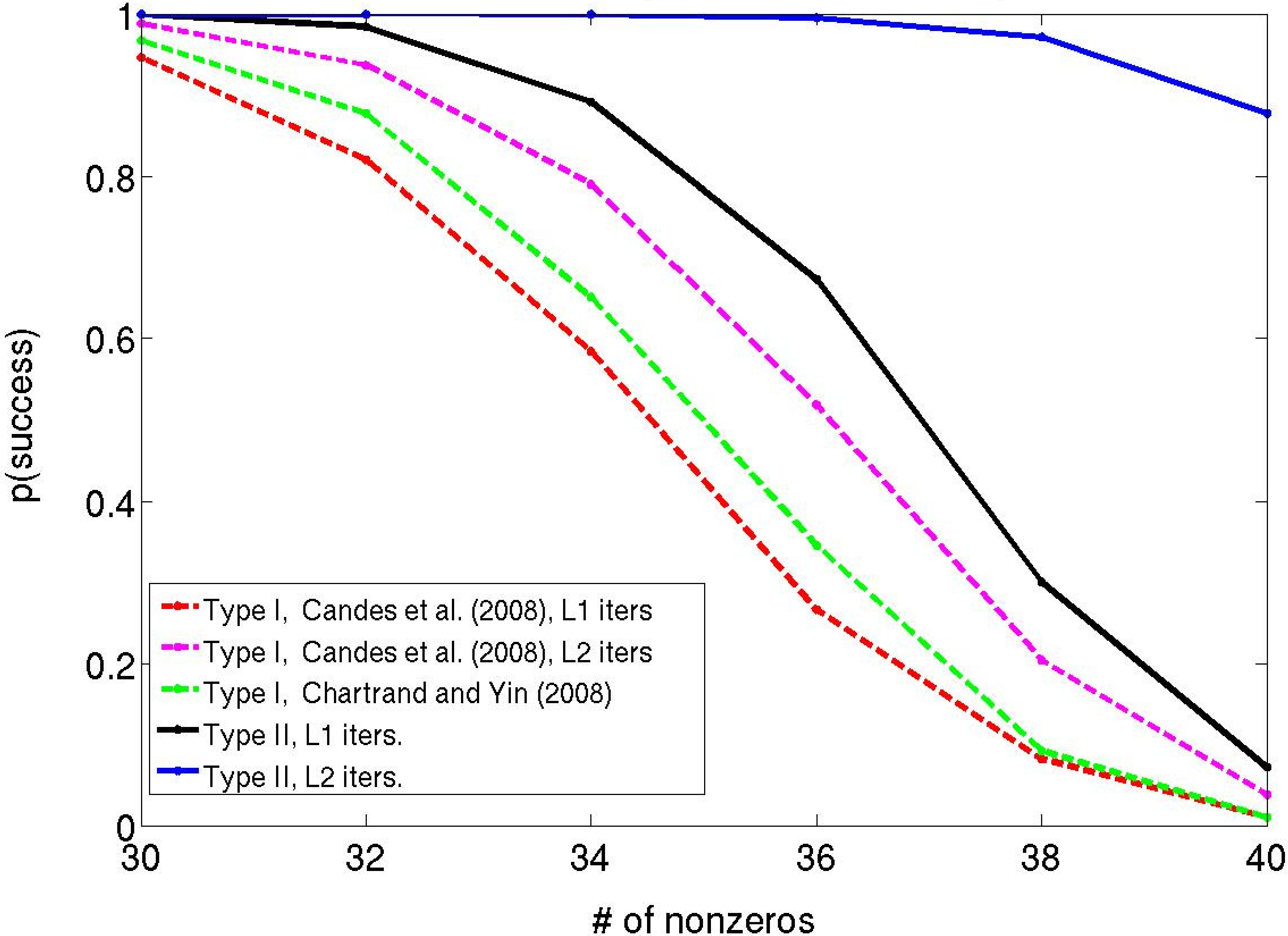
Empirical Example

Penalty	Updates	
$\sum_i \log\left(x_i^2+\varepsilon\right)$,	L_2 iters	[Chartrand and Yin, 2008]
$\sum_i \log\left(\left\lvert x_i\right\rvert+\varepsilon\right)$,	L_1 iters	[Candes et al., 2008]
$\sum_i \log\left(\left\lvert x_i\right\rvert+\varepsilon\right)$,	L_2 iters	
$g^{(II)}\left(\mathbf{x}^2\right)$,	L_1 iters	[Wipf and Nagarajan, 2008]
$g^{(II)}\left(\mathbf{x}^2\right)$,	L_2 iters	[Bishop and Tipping, 2000; Wipf, 2006]

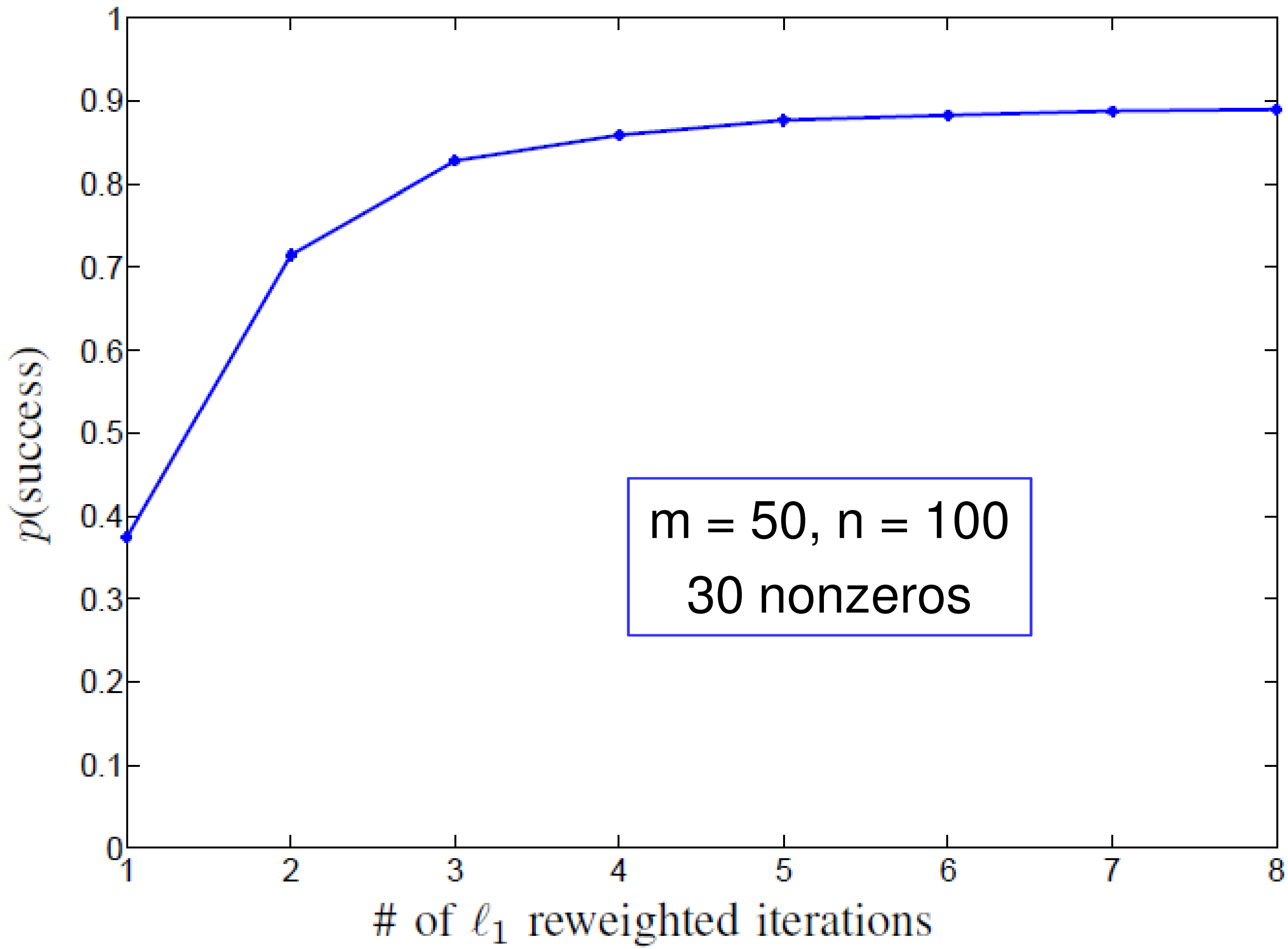
Empirical Example

- Generate data via $Y = \Phi X_0$:
  - Φ 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.

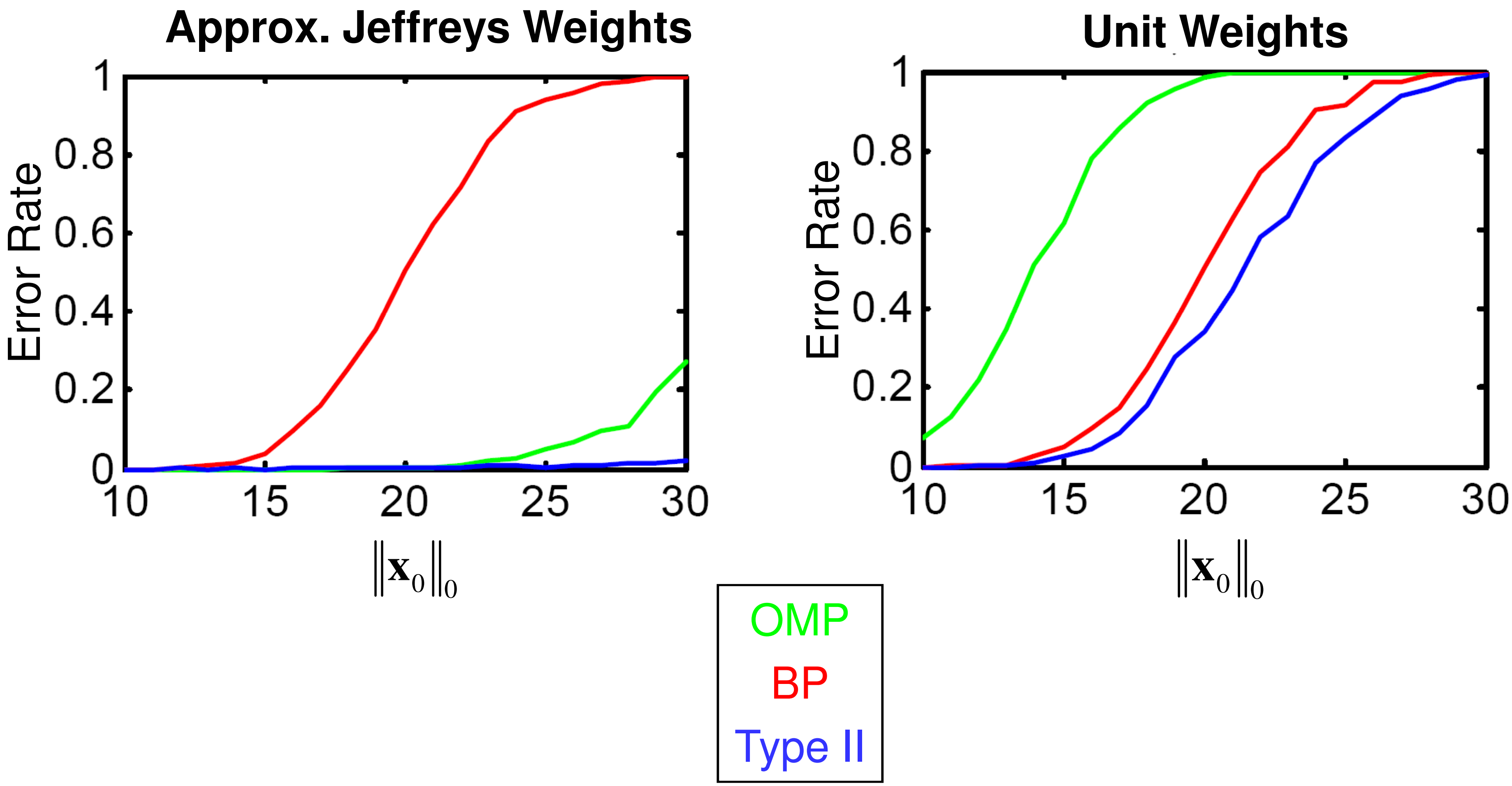
Results (m = 50, n = 100)



Non-Negative Sparse Recovery Example Using Iterative Reweighted L_1 (Type II)



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



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

Wipf and Nagarajan, "Iterative Reweighted L_1 and L_2 Methods for Finding Sparse Solutions," *UCSF Tech Report*, 2009.

Wipf and Nagarajan, "Latent Variable Models for Promoting Sparsity," *UCSF Tech Report*, 2009.