Latent Variable Bayesian Models for Promoting Sparsity

David P. Wipf, Bhaskar D. Rao, and Srikantnan Nagarajan

Appeared in IEEE Trans. on Information Theory 2011

Presented by Yan Kaganovsky
Duke University
Overview

- Type I estimation in coefficient space (MAP)
- Type II estimation in latent space (Empirical Bayes)
  - Variational view of sparse Bayesian learning
- Duality and Unification: Transform Type I $\leftrightarrow$ Type II
- Theoretical Analysis of the type II cost function in coefficient space
Motivation

- Transform Type I $\leftrightarrow$ Type II allows direct comparison between two approaches.
- Some constraints are more natural in coefficient space (e.g., positivity) and some in latent space (e.g., constraints on noise variance) $\Rightarrow$ allows to extend the range of problems to which type I and II are applicable.
- Transforming type II into type I allows to utilize the existing theory for type I and analyze performance of type II.
Type I Estimation

We will be concerned with the generative model

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

(1)

\( \Phi \in \mathbb{R}^{n \times m}, \ x \in \mathbb{R}^m, \ y \in \mathbb{R}^n \)

Consider the regularized regression problem

\[ x^{(I)} \triangleq \arg \min_x \| y - \Phi x \|^2 + \lambda \sum_i g(x_i), \]  

(2)

leading to a Type I estimator \( x^{(I)} \)

Focusing on the following class of sparsity promoting penalties

\[ g(z) = h(z^2) \]  

(3)

\( h(z) \) is some concave and non-decreasing function on \([0, \infty)\).
Type I Estimation: “Bayesian” Interpretation

Prior: \[ p(x) \propto \exp \left[ -\frac{1}{2} \sum_i g(x_i) \right] = \exp \left[ -\frac{1}{2} \sum_i h(x_i^2) \right] \]

Likelihood: \[ p(y|x; \lambda) = \mathcal{N}(y; \Phi x, \lambda I) \]

The type I estimation is interpreted as a maximum à posteriori (MAP) solution

\[ x(I) = \arg \max_x p(x|y; \lambda) = \arg \max_x \frac{p(y|x; \lambda)p(x)}{p(y; \lambda)} \] (4)

Here \( p(y; \lambda) \) is not important but for Type II it will be.
Examples for Penalties/Priors

Penalty for each element $g(z) = h(z^2)$
Prior for each element $p(z) \propto \exp(-\frac{1}{2}g(z))$

Examples:

- $h(z) = z \rightarrow \ell_2$ norm penalty (doesn’t promote sparsity)
- $h(z) = \sqrt{z} \rightarrow \ell_1$ norm penalty (Laplacian)
- $h(z) = |z|^{p/2}, \rightarrow \ell_p$ norm penalty $p \in (0, 2)$ (Promotes sparsity stronger than $\ell_1$)
- $g(z) = \log(|z| + \epsilon), \epsilon \geq 0$ (Limiting case of Student’s t)

$p(\sqrt{z})$ is log-convex on $(0, \infty]$, i.e., super-Gaussian
“more concave” $h \rightarrow$ more sparse solutions
Variational Representation of the Prior

Express the prior \( p(x) \) in terms of \( \gamma \equiv [\gamma_1, \ldots, \gamma_m]^T \in \mathbb{R}_+^m \)

\[
p(x) = \prod_{i=1}^m p(x_i), \quad p(x_i) = \max_{\gamma_i \geq 0} \mathcal{N}(x_i; 0, \gamma_i) \varphi(\gamma_i)
\]

(5)

with \( \varphi \geq 0 \). Dropping “max” we get strict lower bounds and a variational representation for the prior.

Any prior \( p(x) \) constructed via

\[
p(x_i) \propto \exp\left(-\frac{1}{2}g(x_i)\right), \quad g(x_i) = h(x_i^2)
\]

(6)

with \( h \) concave and non-decreasing on \([0, \infty)\) can be expressed in this way.

This includes Laplacian, Jeffreys, Student’s t and generalized Gaussian priors.
Matching the Variational Representation to a Given Prior

In the variation representation

\[ p(x_i) = \max_{\gamma_i \geq 0} \mathcal{N}(x_i; 0, \gamma_i) \varphi(\gamma_i) \] (7)

the function \( \varphi \) is found via

\[ \varphi(\gamma_i) = \sqrt{\frac{2\pi}{\gamma_i}} \exp(g^*(\gamma_i/2)) \] (8)

where \( g^* \) is the concave conjugate of \( g \) given by

\[ g^*(\gamma_i) = \max_z [\gamma_i z - g(z)] \] (9)
Example: Sparse Bayesian Learning (RVM)

From “Perspectives on SBL” in NIPS 2004

In the RVM the following prior is considered

\[
p(x_i | \gamma_i) = \int \mathcal{N}(x_i; 0, \gamma_i) \text{Gamma}(\gamma_i; a, b) d\gamma_i = C(b + x_i^2/2)^{-(a+1/2)}
\]

which is a Student’s t Prior.

The corresponding variational expression is

\[
p(x_i | \gamma_i) = \max_{\gamma_i} \mathcal{N}(x_i; 0, \gamma_i) \varphi(\gamma_i; a, b), \tag{10}
\]

\[
\varphi(\gamma_i; a, b) = \exp \left( -\frac{b}{\gamma_i} \right) \gamma_i^{-a} \tag{11}
\]

For comparison, \( \text{Gamma}(\gamma_i; a, b) \propto \exp(-b/\gamma_i)\gamma_i^{1-a} \)
For fixed $\gamma$ we obtain the *unnormalized* approximate prior

$$\hat{p}_{\gamma}(x) = \prod_{i} \mathcal{N}(x_{i}; 0, \gamma_{i}) \varphi(\gamma_{i})$$  \hspace{1cm} (12)$$

For which the approximate *normalized* posterior

$$\hat{p}_{\gamma}(x|y) = \frac{p(y|x)\hat{p}_{\gamma}(x)}{\int p(y|x)\hat{p}_{\gamma}(x)dx} = \mathcal{N}(x; \mu_{x}, \Sigma_{x})$$ \hspace{1cm} (13)$$

with

$$\mu_{x} = \Gamma \Phi^{T}(\lambda I + \Phi \Gamma \Phi^{T})^{-1}y$$ \hspace{1cm} (14)$$

$$\Sigma_{x} = \Gamma - \Gamma \Phi^{T}(\lambda I + \Phi \Gamma \Phi^{T})^{-1}\Phi$$ \hspace{1cm} (15)$$

where $\Gamma = \text{diag}(\gamma)$
Variational View of Bayesian Type II Estimation

One criterion for selecting the hyperparameters $\gamma$ is

\begin{align}
\gamma_{(II)} \triangleq & \arg \min_{\gamma} \int p(y|x) \left| p(x) - \hat{p}_\gamma(x) \right| \, dx \\
= & \arg \max_{\gamma} \int p(y|x) \hat{p}_\gamma(x) \, dx
\end{align}

Minimize misaligned mass between true and approximate priors in regions where the likelihood is significant.

A common point estimate is then

\begin{equation}
x_{(II)} \triangleq \Gamma_{(II)} \Phi^T (\lambda I + \Phi \Gamma_{(II)} \Phi^T)^{-1} y
\end{equation}

where $\Gamma_{(II)} = \text{diag}(\gamma_{(II)})$.
Type I vs. Type II

**Type I**

\[
\mathcal{L}^x_{(I)} \triangleq -2 \log p(y|x) p(x) \equiv \|y - \Phi x\|_2^2 + \lambda \sum_i g(x_i)
\]

\[
x_{(I)} = \arg \min_x \mathcal{L}^x_{(I)}
\]

**Type II**

\[
\mathcal{L}^\gamma_{(II)} \triangleq y^T \Sigma_y y - \log |\Sigma_y| - 2 \sum_{i=1}^m \log \varphi(\gamma_i)
\]

\[
\Sigma_y = (\lambda I + \Phi \Gamma \Phi^T)^{-1}
\]

\[
\gamma_{(II)} = \arg \min_\gamma \mathcal{L}^\gamma_{(II)} \quad \rightarrow \quad x_{(II)} = F(\gamma_{(II)})
\]
Previous Work

Relationships between type I and type II have been considered before in:


That work focused on similarities and differences between the EM updates not on the relations between cost functions and the duality.
Transforming Type I into Type II-like Problems

Cost function for Type I problem

\[ \mathcal{L}^x_{(I)} \triangleq \| y - \Phi x \|^2_2 + \lambda \sum_i g(x_i) \] (19)

**Theorem 1:** Define the \( \gamma \)-space cost function

\[ \mathcal{L}^\gamma_{(I)} \triangleq y^T \Sigma_y y + \sum_{i=1}^m f_{(I)}(\gamma_i), \quad \gamma \geq 0 \] (20)

\[ \Sigma_y = (\lambda I + \Phi \Gamma \Phi^T)^{-1}, \quad -f_{(I)}(\gamma_i) = \min_{z \geq 0} \gamma_i^{-1} z - h(z) \]

Then \( \gamma_{(I)} \) is a global (or local) minimum of Eq. (20) iff

\[ x_{(I)} \triangleq \Gamma_{(I)} \Phi^T (\lambda I + \Phi \Gamma_{(I)} \Phi^T)^{-1} y \]

\[ \Gamma_{(I)} = \text{diag}[\gamma_{(I)}] \]

is a global (or local) minimum of Eq. (19).
Example: Transforming to a Type II

Using $g(z) = |z|^p$ gives the $\ell_p \ (p \leq 1)$ quasi-norm penalty and the corresponding type II problem is

$$\gamma(I) = \arg \min_{\gamma} \mathbf{y}^T \Sigma \mathbf{y} + \frac{2-p}{p} \left( \frac{p}{2} \right)^{\frac{p}{2-p}} \sum_{i=1}^{m} \gamma_i^{\frac{p}{2-p}} \quad (21)$$

- In D. Wipf, “Dual-Space Analysis of the Sparse Linear Model” (NIPS 2012), the advantages of transforming type I into type II are discussed (even when the “error-bars” are not required)
Transforming Type II into Type I-like Problems

Cost function for Type II problem

\[ \mathcal{L}^{\gamma}_{(II)} \triangleq y^T \Sigma_y y - \log |\Sigma_y| - 2 \sum_{i=1}^{m} \log \varphi(\gamma_i) \] (22)

**Theorem 2:** Define the x-space cost function

\[ \mathcal{L}^{x}_{(II)} \triangleq \|y - \Phi x\|^2_2 + \lambda \mathcal{G}_{(II)}(x) \] (23)

\[ \mathcal{G}_{(II)}(x) \triangleq \min_{\gamma \geq 0} \sum_i \frac{x_i^2}{\gamma_i} - \log |\Sigma_y| - 2 \sum_i \log \varphi(\gamma_i) \] (24)

Then

\[ x_{(II)} \triangleq \Gamma_{(II)} \Phi^T (\lambda I + \Phi \Gamma_{(II)} \Phi^T)^{-1} y, \quad \Gamma_{(II)} = \text{diag}(\Gamma_{(II)}) \] (25)

is a global minimum of Eq. (23) iff \( \gamma_{(II)} \) is the global minimum of Eq. (22)
Transforming Type II into Type I-like Problems

Some remarks:

- if $\log \varphi(\exp(\gamma_i))$ is concave with respect to $\gamma_i$ the above extends to local minima as well
- By optimizing in $x$ space we can add positivity constraints which in the latent space are intractable
Properties of Type II in Coefficient Space

- If \( f(\gamma_i) \triangleq -2 \log \varphi(\gamma_i) \) is concave and non-decreasing \( \Rightarrow \mathcal{G}_{(II)} \) is concave non-decreasing as a function of \(|x|\) with 
  \[ |x| \triangleq [|x_1|, \ldots, |x_m|]^T \]

- All local minima satisfy \( \|x\|_0 \leq n \) (basic feasible solutions), regardless of \( \lambda \)

- For \( \lambda \to 0 \) and \( \text{spark}(\Phi) = n + 1 \), global minimum will be the maximally sparse solution obtained by minimizing 
  \( \|x\|_0 \) s.t. \( y = \Phi x \)
$G_{(II)}$ is non separable, i.e., $G_{(II)}(x) \neq \sum_i G_{(II)}(x_i) \Rightarrow$ the prior is \textit{non-factorial}, i.e., elements of $x$ are dependent.

Unlike traditional type I procedures (e.g., Lasso), $G_{(II)}(x)$ explicitly depends on both $\Phi$ and $\lambda$, specifically:

- For $\Phi^T\Phi = I$, $G_{(II)}(x)$ only depends on $\lambda$
- For $\Phi \rightarrow \Phi D$ with $D$ diagonal the solution becomes $x_{(II)} \rightarrow Dx_{(II)}$
Example: Transforming to Type I

We have already seen that for the RVM:

$$\varphi(\gamma_i; a, b) = \exp(-b/\gamma_i)\gamma_i^{-a}$$  \hfill (26)

Accordingly for \(a, b \to 0\),

$$\mathcal{L}_x^{(\text{II})} \triangleq \|y - \Phi x\|^2_2 + \lambda \mathcal{G}_x(\text{II})(x)$$  \hfill (27)

$$\mathcal{G}_x(\text{II})(x) \triangleq \min_{\gamma \geq 0} \sum_i \frac{x_i^2}{\gamma_i} - \log |\Sigma_y| - 2 \sum_i \log \varphi(\gamma_i)$$  \hfill (28)

Whereas, as a Type-I problem with student’s t prior we should have had

$$\mathcal{L}_x^{(\text{I})} \triangleq \|y - \Phi x\|^2_2 + \lambda \sum_i \log |x_i|$$  \hfill (29)
Advantages of the Non-Separable Penalty

- The $\ell_1$ norm is the tightest *convex* relaxation of the $\ell_0$ semi-norm, and therefore it is commonly used (e.g., Lasso)
- However, the $\ell_1$ norm need not be the best relaxation in general
- The authors demonstrate that the non-separable, noise dependent penalty provides a tighter, albeit non-convex, approximation that promotes greater sparsity than $\ell_1$ while conveniently producing *many fewer local minima than when using $\ell_0$ directly*. 
Advantages of the Non-Separable Penalty

An interesting result:

**Theorem 5:** In the limit $\lambda \rightarrow 0$ (noiseless case), no separable penalty $G(x) = \sum_i g(x_i)$ exits such that for all $y$ and $\Phi$ with $\text{spark}(\Phi) = n + 1$, the corresponding Type I optimization problem

$$\min_x \sum_i g(x_i), \quad \text{s.t.} \quad y = \Phi x$$

is: (i) Globally minimized by the maximally sparsest solution, and (ii) Ever has fewer local minima than when solving with the proposed penalty.
Example: How Type II is Smoothing Local Minima

Fig. 1. Plots of the Type II penalty (normalized) across the feasible region as parameterized by $\alpha$. A separable penalty given by $g(x) \propto \sum |x_i|^{0.01} \approx \|x\|_0$ is included for comparison. Both approximations to the $\ell_0$ norm retain the correct global minimum, but only the Type II penalty smooths out local minima. \textit{Left:} $\|x_0\|_0 = 1$ (simple case). \textit{Right:} $\|x_0\|_0 = 9$ (hard case).
Empirical Results

![Empirical Results Diagram](image-url)