Dual-Space Analysis of the Sparse Linear Model

David Wipf and Yi Wu
Visual Computing Group, Microsoft Research Asia
davidwipf@gmail.com, jxwuyi@gmail.com

Abstract

Sparse linear (or generalized linear) models combine a standard likelihood function with a sparse prior on the unknown coefficients. These priors can conveniently be expressed as a maximization over zero-mean Gaussians with different variance hyperparameters. Standard MAP estimation (Type I) involves maximizing over both the hyperparameters and coefficients, while an empirical Bayesian alternative (Type II) first marginalizes the coefficients and then maximizes over the hyperparameters, leading to a tractable posterior approximation. The underlying cost functions can be related via a dual-space framework from [22], which allows both the Type I or Type II objectives to be expressed in either coefficient or hyperparmeter space. This perspective is useful because some analyses or extensions are more conducive to development in one space or the other. Herein we consider the estimation of a trade-off parameter balancing sparsity and data fit. As this parameter is effectively a variance, natural estimators exist by assessing the problem in hyperparameter (variance) space, transitioning natural ideas from Type II to solve what is much less intuitive for Type I. In contrast, for analyses of update rules and sparsity properties of local and global solutions, as well as extensions to more general likelihood models, we can leverage coefficient-space techniques developed for Type I and apply them to Type II. For example, this allows us to prove that Type II-inspired techniques can be successful recovering sparse coefficients when unfavorable restricted isometry properties (RIP) lead to failure of popular ℓ_1 reconstructions. It also facilitates the analysis of Type II when non-Gaussian likelihood models lead to intractable integrations.

1 Introduction

We begin with the likelihood model

$$\boldsymbol{y} = \Phi \boldsymbol{x} + \boldsymbol{\epsilon}, \tag{1}$$

where $\Phi \in \mathbb{R}^{n \times m}$ is a dictionary of unit ℓ_2-norm basis vectors, $\boldsymbol{x} \in \mathbb{R}^m$ is a vector of unknown coefficients we would like to estimate, $\boldsymbol{y} \in \mathbb{R}^n$ is the observed signal, and $\boldsymbol{\epsilon}$ is noise distributed as $\mathcal{N}(\boldsymbol{\epsilon}; 0, \lambda I)$ (later we consider more general likelihood models). In many practical situations where large numbers of features are present relative to the signal dimension, the problem of estimating $\boldsymbol{x}$ given $\boldsymbol{y}$ becomes ill-posed. A Bayesian framework is intuitively appealing for formulating these types of problems because prior assumptions must be incorporated, whether explicitly or implicitly, to regularize the solution space.

Recently, there has been a growing interest in models that employ sparse priors $p(\boldsymbol{x})$ to encourage solutions $\boldsymbol{x}$ with mostly small or zero-valued coefficients and a few large or unrestricted values, i.e., we are assuming the generative $\boldsymbol{x}$ is a sparse vector. Such solutions can be favored by using

$$p(\boldsymbol{x}) \propto \prod_i \exp\left[-\frac{1}{2} g(x_i)\right] = \prod_i \exp\left[-\frac{1}{2} h\left(x_i^2\right)\right], \tag{2}$$

with h concave and non-decreasing on $[0, \infty)$ [15, 16]. Virtually all sparse priors of interest can be expressed in this manner, including the popular Laplacian, Jeffreys, Student's t, and generalized

Gaussian distributions. Roughly speaking, the 'more concave' h, the more sparse we expect $\boldsymbol{x}$ to be. For example, with $h(z) = z$, we recover a Gaussian, which is not sparse at all, while $h(z) = \sqrt{z}$ gives a Laplacian distribution, with characteristic heavy tails and a sharp peak at zero.

All sparse priors of the form (2) can be conveniently framed in terms of a collection of non-negative latent variables or hyperparameters $\boldsymbol{\gamma} \triangleq [\gamma_1, \ldots, \gamma_m]^T$ for purposes of optimization, approximation, and/or inference. The hyperparameters dictate the structure of the prior via

$$p(\boldsymbol{x}) = \prod_i p(x_i), \quad p(x_i) = \max_{\gamma_i \geq 0} \mathcal{N}(x_i; 0, \gamma_i)\varphi(\gamma_i), \tag{3}$$

where $\varphi(\gamma_i)$ is some non-negative function that is sometimes treated as a hyperprior, although it will not generally integrate to one. For the purpose of obtaining sparse point estimates of $\boldsymbol{x}$, which will be our primary focus herein, models with latent variable sparse priors are frequently handled in one of two ways. First, the latent structure afforded by (3) offers a very convenient means of obtaining (possibly local) *maximum a posteriori* (MAP) estimates of $\boldsymbol{x}$ by iteratively solving

$$\boldsymbol{x}_{(I)} = \arg\min_{\boldsymbol{x}} -\log p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x}) = \arg\min_{\boldsymbol{x};\boldsymbol{\gamma}\succeq 0} \|\boldsymbol{y} - \Phi\boldsymbol{x}\|_2^2 + \lambda \sum_i \left[\frac{x_i^2}{\gamma_i} + \log\gamma_i + f(\gamma_i)\right], \tag{4}$$

where $f(\gamma_i) \triangleq -2\log\varphi(\gamma_i)$ and $\boldsymbol{x}_{(I)}$ is commonly referred to as a *Type I* estimator. Examples include minimum ℓ_p-norm approaches [4, 11, 16], Jeffreys prior-based methods sometimes called FOCUSS [7, 6, 9], algorithms for computing the basis pursuit (BP) or Lasso solution [6, 16, 18], and iterative reweighted ℓ_1 methods [3].

Secondly, instead of maximizing over both $\boldsymbol{x}$ and $\boldsymbol{\gamma}$ as in (4), *Type II* methods first integrate out (marginalize) the unknown $\boldsymbol{x}$ and then solve the empirical Bayesian problem [19]

$$\begin{aligned}\boldsymbol{\gamma}_{(II)} &= \arg\max_{\boldsymbol{\gamma}} p(\boldsymbol{\gamma}|\boldsymbol{y}) = \arg\max_{\boldsymbol{\gamma}} \int p(\boldsymbol{y}|\boldsymbol{x}) \prod_i \mathcal{N}(\boldsymbol{x}; 0, \gamma_i)\varphi(\boldsymbol{\gamma}_i)dx_i \\ &= \arg\min_{\boldsymbol{\gamma}} \; \boldsymbol{y}^T\Sigma_y^{-1}\boldsymbol{y} + \log|\Sigma_y| + \sum_{i=1}^m f(\gamma_i), \end{aligned} \tag{5}$$

where $\Sigma_y \triangleq \lambda I + \Phi\Gamma\Phi^T$ and $\Gamma \triangleq \text{diag}[\boldsymbol{\gamma}]$. Once $\boldsymbol{\gamma}_{(II)}$ is obtained, the conditional distribution $p(\boldsymbol{x}|\boldsymbol{y}; \boldsymbol{\gamma}_{(II)})$ is Gaussian, and a point estimate for $\boldsymbol{x}$ naturally emerges as the posterior mean

$$\boldsymbol{x}_{(II)} = \mathrm{E}\left[\boldsymbol{x}|\boldsymbol{y}; \boldsymbol{\gamma}_{(II)}\right] = \Gamma_{(II)}\Phi^T\left(\lambda I + \Phi\Gamma_{(II)}\Phi^T\right)^{-1}\boldsymbol{y}. \tag{6}$$

Pertinent examples include sparse Bayesian learning and the relevance vector machine (RVM) [19], automatic relevance determination (ARD) [14], methods for learning overcomplete dictionaries [8], and large-scale experimental design [17].

While initially these two approaches may seem vastly different, both can be directly compared using a dual-space view [22] of the underlying cost functions. In brief, this involves expressing both the Type I and Type II objective solely in terms of either $\boldsymbol{x}$ or $\boldsymbol{\gamma}$ as reviewed in Section 2. The dual-space view is advantageous for several reasons, such as establishing connections between algorithms, developing efficient update rules, or handling more general (non-Gaussian) likelihood functions. In Section 3, we utilize $\boldsymbol{\gamma}$-space cost functions to develop a principled method for choosing the trade-off parameter λ (which accompanies the Gaussian likelihood model and essentially balances sparsity and data fit) and demonstrate its effectiveness via simulations. Section 4 then derives a new Type II-inspired algorithm in $\boldsymbol{x}$-space that can compute maximally sparse (minimal ℓ_0 norm) solutions even with highly coherent dictionaries, proving a result for clustered dictionaries that previously has only been shown empirically [21]. Finally, Section 5 leverages duality to address Type II methods with generalized likelihood functions that previously were rendered untenable because of intractable integrals. In general, some tasks and analyses are easier to undertake in $\boldsymbol{\gamma}$-space (Section 3), while others are more transparent in $\boldsymbol{x}$-space (Sections 4 and 5). Here we consider both with the goal of advancing the proper understanding and full utilization of the sparse linear model.

2 Dual-Space View of the Sparse Linear Model

Type I is based on a natural cost function in $\boldsymbol{x}$-space, $p(\boldsymbol{x}|\boldsymbol{y})$, while Type II involves an analogous function in $\boldsymbol{\gamma}$-space, $p(\boldsymbol{\gamma}|\boldsymbol{y})$. The dual-space view defines a corresponding $\boldsymbol{\gamma}$-space cost function for Type I and a $\boldsymbol{x}$-space cost function for Type II to complete the symmetry.

Type II in $\boldsymbol{x}$-Space: Using the relationship

$$\boldsymbol{y}\Sigma_y^{-1}\boldsymbol{y} = \min_x \frac{1}{\lambda}\|\boldsymbol{y}-\Phi\boldsymbol{x}\|_2^2 + \boldsymbol{x}^T\Gamma^{-1}\boldsymbol{x} \tag{7}$$

as in [22], it can be shown that the Type II coefficients from (6) satisfy $\boldsymbol{x}_{(II)} = \arg\min_{\boldsymbol{x}} \mathcal{L}_{(II)}(\boldsymbol{x})$, where

$$\mathcal{L}_{(II)}(\boldsymbol{x}) \;\triangleq\; \|\boldsymbol{y}-\Phi\boldsymbol{x}\|_2^2 + \lambda g_{(II)}(\boldsymbol{x}), \tag{8}$$

and

$$g_{(II)}(\boldsymbol{x}) \;\triangleq\; \min_{\boldsymbol{\gamma}\succeq 0} \;\sum_i \frac{x_i^2}{\gamma_i} + \log|\Sigma_y| + \sum_i f(\gamma_i). \tag{9}$$

This reformulation of Type II in $\boldsymbol{x}$-space is revealing for multiple reasons (Sections 4 and 5 will address additional reasons in detail). For many applications of the sparse linear model, the primary goal is simply a point estimate that exhibits some degree of sparsity, meaning many elements of $\hat{\boldsymbol{x}}$ near zero and a few relatively large coefficients. This requires a penalty function $g(\boldsymbol{x})$ that is concave and non-decreasing in $\boldsymbol{x}^2 \triangleq [x_1^2,\ldots,x_m^2]^T$. In the context of Type I, any prior $p(\boldsymbol{x})$ expressible via (2) will satisfy this condition by definition; such priors are said to be *strongly super-Gaussian* and will always have positive kurtosis [15]. Regarding Type II, because the associated $\boldsymbol{x}$-space penalty (9) is represented as a minimum of upper-bounding hyperplanes with respect to $\boldsymbol{x}^2$ (and the slopes are all non-negative given $\boldsymbol{\gamma}\succeq 0$), it must therefore be concave and non-decreasing in $\boldsymbol{x}^2$ [1].

For compression, interpretability, or other practical reasons, it is sometimes desirable to have *exactly sparse* point estimates, with many (or most) elements of $\boldsymbol{x}$ equal to exactly zero. This then necessitates a penalty function $g(\boldsymbol{x})$ that is concave and non-decreasing in $|\boldsymbol{x}| \triangleq [|x_1|,\ldots,|x_m|]^T$, a much stronger condition. In the case of Type I, if $\log\gamma + f(\gamma)$ is concave and non-decreasing in γ, then $g(\boldsymbol{x}) = \sum_i g(x_i)$ satisfies this condition. The Type II analog, which emerges by further inspection of (9) stipulates that if

$$\log|\Sigma_y| + \sum_i f(\gamma_i) = \log\left|\lambda^{-1}\Phi^T\Phi + \Gamma^{-1}\right| + \log|\Gamma| + \sum_i f(\gamma_i) \tag{10}$$

is a concave and non-decreasing function of $\boldsymbol{\gamma}$, then $g_{(II)}(\boldsymbol{x})$ will be a concave, non-decreasing function of $|\boldsymbol{x}|$. For this purpose it is sufficient, but not necessary, that f be a concave and non-decreasing function. Note that this is a somewhat stronger criteria than Type I since the first term on the righthand side of (10) (which is absent from Type I) is actually convex in $\boldsymbol{\gamma}$. Regardless, it is now very transparent how Type II may promote sparsity akin to Type I.

The dual-space view also leads to efficient, convergent algorithms such as iterative reweighted ℓ_1 minimization and its variants as discussed in [22]. However, building on these ideas, we can demonstrate here that it also elucidates the original, widely applied update procedures developed for implementing the relevance vector machine (RVM), a popular Type II method for regression and classification that assumes $f(\gamma) = 0$ [19]. In fact these updates, which were inspired by a fixed-point heuristic from [12], have been widely used for a number of Bayesian inference tasks without any formal analyses or justification.[1] The dual-space formulation can be leveraged to show that these updates are in fact executing a coordinate-wise, iterative min-max procedure in search of a saddle point. Specifically we have the following result (all proofs are in the supplementary material):

Theorem 1. The original RVM update rule from [19, Equation (16)] is equivalent to a closed-form, coordinate-wise optimization of

$$\min_{\boldsymbol{x};\boldsymbol{\gamma}\succeq 0}\;\max_{\boldsymbol{z}\succeq 0}\left[\|\boldsymbol{y}-\Phi\boldsymbol{x}\|_2^2 + \sum_i\left(\frac{x_i^2}{\gamma_i} + z_i\log\gamma_i\right) - \vartheta(\boldsymbol{z})\right] \tag{11}$$

over $\boldsymbol{x}$, $\boldsymbol{\gamma}$, and $\boldsymbol{z}$, where $\vartheta(\boldsymbol{z})$ is the convex conjugate function [1] of $\log\left|\lambda I + \Phi\mathrm{diag}[\exp(\boldsymbol{u})]\Phi^T\right|$ with respect to $\boldsymbol{u}$.

[1] Although a more recent, step-wise variant of the RVM has been shown to be substantially faster [20], the original version is still germane since it can easily be extended to handle more general structured sparsity problems. The step-wise method cannot without introducing additional approximations [10].

Type I in γ-Space: Similar methodology and the expansion of $\boldsymbol{y}^T\Sigma_y^{-1}\boldsymbol{y}$ can be used to express the Type I optimization problem in $\boldsymbol{\gamma}$-space, which serves several useful purposes. Let $\boldsymbol{\gamma}_{(I)} \triangleq \arg\min_{\boldsymbol{\gamma}\geq 0}\mathcal{L}_{(I)}(\boldsymbol{\gamma})$, with

$$\mathcal{L}_{(I)}(\boldsymbol{\gamma}) \triangleq \boldsymbol{y}^T\Sigma_y^{-1}\boldsymbol{y} + \log|\Gamma| + \sum_{i=1}^{m} f(\gamma_i). \tag{12}$$

Then the Type I coefficients obtained from (4) satisfy

$$\boldsymbol{x}_{(I)} = \Gamma_{(I)}\Phi^T\left(\lambda I + \Phi\Gamma_{(I)}\Phi^T\right)^{-1}\boldsymbol{y}. \tag{13}$$

Section 3 will use $\boldsymbol{\gamma}$-space cost functions to derive well-motivated approaches for learning the trade-off parameter λ.

3 Choosing the Trade-off Parameter λ

The trade-off parameter is crucial for obtaining good estimates of $\boldsymbol{x}$. In general, if λ is too large, $\hat{\boldsymbol{x}} \rightarrow 0$; too small and $\hat{\boldsymbol{x}}$ is overfitted to the noise. In practice, either expensive cross-validation or some heuristic procedure is often required. However, because λ can be interpreted as a variance, it is useful to address its estimation in $\boldsymbol{\gamma}$-space, in which existing unknowns (i.e., $\boldsymbol{\gamma}$) are also variances.

Learning λ with Type I: Consider the Type I cost function $\mathcal{L}_{(I)}(\boldsymbol{\gamma})$. The data-dependent term can be shown to be a convex, non-increasing function of $\boldsymbol{\gamma}$, which encourages each element to be large. The second term is a penalty factor that regulates the size of $\boldsymbol{\gamma}$. It is here that a convenient regularizer for λ can be incorporated.

This can be accomplished as follows. First we expand Σ_y via $\Sigma_y = \sum_{j=1}^{m}\gamma_i\boldsymbol{\phi}_{\cdot i}\boldsymbol{\phi}_{\cdot i}^T + \sum_{j=1}^{n}\lambda\boldsymbol{e}_j\boldsymbol{e}_j^T$, where $\boldsymbol{\phi}_{\cdot i}$ denotes the i-th column of Φ and $\boldsymbol{e}_j$ is a column vector of zeros with a '1' in the j-th location. Thus we observe that λ is embedded in the data-dependent term in the exact same fashion as each γ_i. This motivates a penalty on λ with similar correspondence, leading to the objective

$$\begin{aligned}\mathcal{L}_{(I)}(\boldsymbol{\gamma},\lambda) &\triangleq \boldsymbol{y}^T\Sigma_y^{-1}\boldsymbol{y} + \sum_{i=1}^{m}\left[\log\gamma_i + f(\gamma_i)\right] + \sum_{j=1}^{n}\left[\log\lambda + f(\lambda)\right] \\ &= \boldsymbol{y}^T\Sigma_y^{-1}\boldsymbol{y} + \sum_{i=1}^{m}\left[\log\gamma_i + f(\gamma_i)\right] + n\log\lambda + nf(\lambda).\end{aligned} \tag{14}$$

While admittedly simple, this construction is appealing because, regardless of how each γ_i is penalized, λ is penalized in a proportional manner, so both $\boldsymbol{\gamma}$ and λ have a properly balanced chance of explaining the observed data. This is important because the optimal λ will be highly dependent on both the true noise level, *and crucially*, the particular sparse prior assumed $p(\boldsymbol{x})$ (as reflected by f).

For analysis or implementational purposes, we may convert $\mathcal{L}_{(I)}(\boldsymbol{\gamma},\lambda)$ back to $\boldsymbol{x}$-space, with λ-dependency now removed. It can then be shown that solving (4), with λ fixed to the value that minimizes (14), is equivalent to solving

$$\min_{\boldsymbol{x},\boldsymbol{u}}\sum_i g(x_i) + ng\left(\frac{1}{\sqrt{n}}\|\boldsymbol{u}\|_2\right), \qquad \text{s.t.} \quad \boldsymbol{y} = \Phi\boldsymbol{x} + \boldsymbol{u}. \tag{15}$$

If $\boldsymbol{x}_*$ and $\boldsymbol{u}_*$ minimize (15), then we can demonstrate using [15] that the corresponding λ estimate, which also minimizes (14), is given by $\lambda_* = \partial h(z)/\partial z$ evaluated at $z = 1/n\|\boldsymbol{u}_*\|_2^2$. Note that if we were just performing maximum likelihood estimation of λ given $\boldsymbol{x}_*$, the optimal value would reduce to simply $\lambda_* = 1/n\|\boldsymbol{u}_*\|_2^2$, with no influence from the prior on $\boldsymbol{x}$. This is a fundamental weakness.

Solving (15), or equivalently (14), can be accomplished using simple iterative reweighted least squares, or if g is concave in $|x_i|$, an iterative reweighted second-order-cone (SOC) minimization.

Learning λ with Type II: The same procedure can be adopted for Type II yielding the cost function

$$\mathcal{L}_{(II)}(\boldsymbol{\gamma},\lambda) = \boldsymbol{y}^T\Sigma_y^{-1}\boldsymbol{y} + \log|\Sigma_y| + \sum_i f(\gamma_i) + nf(\lambda), \tag{16}$$

where we note that, unlike in the Type I case above, the $\log$-based term is already naturally balanced between λ and $\boldsymbol{\gamma}$ by virtue of the symmetric embedding in Σ_y. It is important to stress that this Type II prescription for learning λ is not the same as originally proposed in the literature for Type II models of this genre. In this context, $\varphi(\gamma_i)$ is interpreted a hyperprior on γ_i, and an equivalent distribution is assumed on the noise variance λ. Importantly, these assumptions leave out the factor of n in (16), and so an asymmetry is created.

Simulation Examples: Empirical tests help to illustrate the efficacy of this procedure. As in many applications of sparse reconstruction, here we are only concerned with accurately estimating $\boldsymbol{x}$, whose nonzero entries may have physical significance (e.g., source localization [16], compressive sensing [2], etc.), as opposed to predicting new values of $\boldsymbol{y}$. Therefore, automatically learning the value of λ is particularly relevant, since cross-validation is often not possible.[2] Simulations are helpful for evaluation purposes since we then have access to the true sparse generating vector.

Figure 1 compares the estimation performance obtained by minimizing (15) with two different selections for g: $g(\boldsymbol{x}) = \|\boldsymbol{x}\|_p^p = \sum_i |x_i|^p$, with $p = 0.01$ and $p = 1.0$. Data generation proceeds as follows: We create a random 100×50 dictionary Φ, with ℓ_2-normalized, iid Gaussian columns. $\boldsymbol{x}$ is randomly generated with 10 unit Gaussian nonzero elements. We then compute $\boldsymbol{y} = \Phi\boldsymbol{x} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is iid Gaussian noise producing an SNR of 0dB. To determine what λ values lead to optimal performance we solve (4) with the appropriate g over a range of fixed λ values (10^{-4} to 10^{1}) and then compute the error between $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$. The minimum of this curve reflects the best performance we can hope to achieve when learning λ blindly. In Figure 1 (*Top*) we plot these curves for both Type I methods averaged over 1000 independent trials.

Next we solve (15), which produces an estimate of both $\boldsymbol{x}$ and λ. We mark with an '+' the learned λ versus the corresponding error of $\hat{\boldsymbol{x}}$. In both cases the learned λ's (averaged across trials) perform just as well as if we knew the optimal value a priori. Results using other noise levels, problem dimensions n and m, sparsity levels $\|\boldsymbol{x}\|_0$, and sparsity penalties g are similar. See the supplementary material for more examples.

Figure 1 (*Bottom*) shows the average sparsity of estimates $\hat{\boldsymbol{x}}$, as quantified by the ℓ_0 norm $\|\hat{\boldsymbol{x}}\|_0$, across λ values ($\|\boldsymbol{x}\|_0$ returns a count of the number of nonzero elements in $\boldsymbol{x}$). The '+' indicates the average sparsity of each $\hat{\boldsymbol{x}}$ for the learned λ as before. In general, the $\ell_{(0.01)}$ penalty produces a much sparser estimate, very near the true value of $\|\boldsymbol{x}\|_0 = 10$ at the optimal λ. The ℓ_1 penalty, which is substantially less concave/sparsity-inducing, still sets some elements to exactly zero, but also substantially shrinks nonzero coefficients in achieving a similar overall reconstruction error. This highlights the importance of learning a λ via a penalty that is properly matched to the prior on $\boldsymbol{x}$: if we instead tried to force a particular sparsity value (in this case 10), then the ℓ_1 solution would be very suboptimal. Finally we note that maximum likelihood (ML) estimation of λ performs very poorly (not shown), except in the special case where the ML estimate is equivalent to solving (14) as occurs when $f(\gamma) = 0$ (see [6]). The proposed method can be viewed as adding a principled hyperprior on λ, properly matched to $p(\boldsymbol{x})$, that compensates for this shortcoming of standard ML.

Type II λ estimation has been explored elsewhere for the special case where $f(\gamma) = 0$ [19], which renders the factor of n in (16) irrelevant; however, for other selections we have found this factor to improve performance (not shown). For space considerations we have focused our attention here on Type I, which has frequently been noted for not lending itself well to λ estimation (or related parameters) [6, 13]. In fact, the symmetry afforded by the dual-space perspective reveals that Type I is just as natural a candidate for this task as Type II, and may be preferred in high-dimensional settings where computational resources are at a premium.

4 Maximally Sparse Estimation

With the advent of compressive sensing and other related applications, there has been growing interest in finding *maximally sparse* signal representations from redundant dictionaries ($m \gg n$) [3, 5]. The canonical form of this problem involves solving

$$\boldsymbol{x}_0 \triangleq \arg\min_{\boldsymbol{x}} \|\boldsymbol{x}\|_0, \qquad \text{s.t. } \boldsymbol{y} = \Phi\boldsymbol{x}. \tag{17}$$

[2] For example, in non-stationary environments, the value of both $\boldsymbol{x}$ and λ may be completely different for any new $\boldsymbol{y}$, which then necessitates that we estimate both jointly.

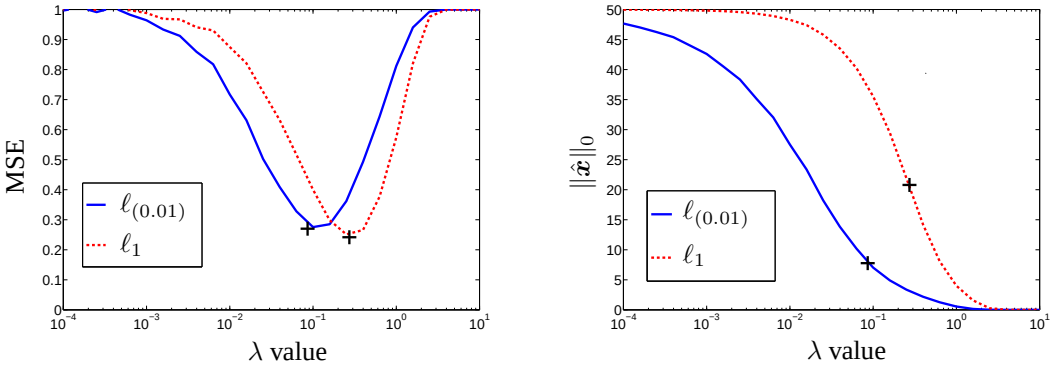

Figure 1: *Left*: Normalized mean-squared error (MSE) given by $\left\langle \|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2^2 / \|\boldsymbol{x}\|_2 \right\rangle$ (where the average is across 1000 trials) plotted versus λ for two different Type I approaches. Each black '+' represents the estimated value of λ (averaged across trials) and the associated MSE produced with this estimate. In both cases the estimated value achieves the lowest possible MSE (it can actually be slightly *lower* than the curve because its value is allowed to fluctuate from trial to trial). *Right*: Solution sparsity $\|\hat{\boldsymbol{x}}\|_0$ versus λ. Even though they both lead to similar MSE, the $\ell_{(0.01)}$ penalty produces a much sparser estimate at the optimal λ value.

While (17) is NP-hard, whenever the dictionary Φ satisfies a *restricted isometry property* (RIP) [2] or a related structural assumption, meaning that each $\|\boldsymbol{x}_0\|_0$ columns of Φ are sufficiently close to orthonormal (i.e., mutually uncorrelated), then replacing ℓ_0 with ℓ_1 in (17) leads to a convex problem with an equivalent global solution. Unfortunately however, in many situations (e.g., feature selection, source localization) these RIP equivalence conditions are grossly violated, implying that the ℓ_1 solution may deviate substantially from $\boldsymbol{x}_0$.

An alternative is to instead replace (17) with minimization of (8) and then take the limit as $\lambda \to 0$. (Note that the extension to the noisy case with $\lambda > 0$ is straightforward, but analysis is more difficult.) In this regime the optimization problem reduces to

$$\boldsymbol{x}_{(II)} = \lim_{\lambda \to 0} \arg\min_{\boldsymbol{x}} \; g_{(II)}(\boldsymbol{x}), \qquad \text{s.t. } \boldsymbol{y} = \Phi \boldsymbol{x}. \tag{18}$$

If $\log|\Sigma_y| + \sum_i f(\gamma_i)$ is concave, then (18) can be minimized using reweighted ℓ_1 minimization. With initial weight vector $\boldsymbol{w}^{(0)} = \mathbf{1}$, the $(k+1)$-th iteration involves computing

$$\boldsymbol{x}^{(k+1)} \leftarrow \arg\min_{\boldsymbol{x}:\, \boldsymbol{y} = \Phi \boldsymbol{x}} \sum_i w_i^{(k)} |x_i|, \qquad \boldsymbol{w}^{(k+1)} \leftarrow \left. \frac{\partial g_{(II)}(\boldsymbol{x})}{\partial |x_i|} \right|_{\boldsymbol{x} = \boldsymbol{x}^{(k+1)}}. \tag{19}$$

With $f(\gamma) = 0$, iterating (19) will provably lead to an estimate of $\boldsymbol{x}_0$ that is as good or better than the ℓ_1 solution [21], in particular when Φ has highly correlated columns. Additionally, the assumption $f(\gamma) = 0$ leads to a closed-form expression for the weights $\boldsymbol{w}^{(k+1)}$. Let

$$\eta_i(\boldsymbol{x}; \alpha, q) \triangleq \left[\boldsymbol{\phi}_{\cdot i}^T \left(\alpha I + \Phi |X^{(k+1)}|^2 \Phi^T \right)^{-1} \boldsymbol{\phi}_{\cdot i} \right]^q, \tag{20}$$

where $|X^{(k+1}|$ denotes a diagonal matrix with i-th diagonal entry given by $|x_i^{(k+1)}|$. Then $\boldsymbol{w}^{(k+1)}$ can be computed via $w_i^{(k+1)} = \eta_i(\boldsymbol{x}; 0, 1/2), \;\; \forall i$. It remains unclear however in what circumstances this type of update can lead to guaranteed improvement nor if the functions $\eta_i(\boldsymbol{x}; 0, 1/2)$ are even the optimal choice. We will now demonstrate that for certain selections of α and q, we can guarantee that reweighted ℓ_1 using $\eta_i(\boldsymbol{x}; \alpha, q)$ is guaranteed to recover $\boldsymbol{x}_0$ exactly if Φ is drawn from what we call a *clustered dictionary model*.

Definition 1. *Clustered Dictionary Model:* Let $\Phi_{uncorr}^{(d)}$ denote any dictionary such that ℓ_1 minimization succeeds in solving (17) for all $\|\boldsymbol{x}_0\|_0 \leq d$. Let $\Phi_{corr}^{(d,\epsilon)}$ denote any dictionary obtained by replacing each column of $\Phi_{uncorr}^{(d)}$ with a "cluster" of m_i basis vectors such that the angle between any two vectors within a cluster is less than some $\epsilon > 0$. We also define the cluster support

$\Omega_0 \subset \{1, 2, \ldots, m\}$ as the set of cluster indices whereby $\boldsymbol{x}_0$ has at least one nonzero element. Finally, we assume that the resulting $\Phi_{corr}^{(d,\epsilon)}$ is such that every $n \times n$ submatrix is full rank.

Theorem 2. For any sparse vector $\boldsymbol{x}_0$ and any dictionary $\Phi_{corr}^{(d,\epsilon)}$ obtained from the clustered dictionary model with ϵ sufficiently small, reweighted ℓ_1 minimization using weights $\eta_i(\boldsymbol{x}; \lambda, q)$ with some $q \geq 1$ and α sufficiently small will recover $\boldsymbol{x}_0$ exactly provided that $|\Omega_0| \leq d$, $\sum_{i \in \Omega_0} m_i \leq n$, and within each cluster $k \in \Omega_0$ the coefficients do not sum to zero.

Theorem 2 implies that even though ℓ_1 may fail to find the maximally sparse $\boldsymbol{x}_0$ because of severe RIP violations (high correlations between groups of dictionary columns as dictated by ϵ lead directly to a poor RIP), a Type II-inspired method can still be successful. Moreover, because whenever ℓ_1 does succeed, Type II will always succeed as well (assuming a reweighted ℓ_1 implementation), the converse (RIP violation leading to Type II failure but not ℓ_1 failure) can never happen. Recent work from [21] has argued that Type II may be useful for addressing the sparse recovery problem with correlated dictionaries, and empirical evidence is provided showing vastly superior performance on clustered dictionaries. However, we stress that no results proving global convergence to the correct, maximally sparse solution have been shown before in the case of structured dictionaries (except in special cases with strong, unverifiable constraints on coefficient magnitudes [21]). Moreover, the proposed weighting strategy $\eta_i(\boldsymbol{x}; \lambda, q)$ accomplishes this without any particular tuning to the clustered dictionary model under consideration and thus likely holds in many other cases as well.

5 Generalized Likelihood functions

Type I methods naturally accommodate alternative likelihood functions. We simply must replace the quadratic data fit term from (4) with some preferred function and then coordinate-wise optimization may proceed provided we have an efficient means of computing a weighted ℓ_2-norm penalized solution. In contrast, generalizing Type II is substantially more complicated because it is no longer possible to compute the marginalization (5) or the posterior distribution $p(\boldsymbol{x}|\boldsymbol{y}; \boldsymbol{\gamma}_{(II)})$. Therefore, to obtain a tractable estimate $\boldsymbol{x}_{(II)}$ additional heuristics are required. For example, the RVM classifier from [19] employs a Laplace approximation for this purpose; however, it is not clear what cost function is being minimized nor rigorous properties of the estimated solutions.

Fortunately, the dual $\boldsymbol{x}$-space view provides a natural mechanism for generalizing the basic Type II methodology to address alternative likelihood functions in a more principled manner. In the case of classification problems, we might want to replace the Gaussian likelihood $p(\boldsymbol{y}|\boldsymbol{x})$ implied by (1) with a multivariate Bernoulli distribution $p(\boldsymbol{y}|\boldsymbol{x}) \propto \log[-\psi(\boldsymbol{y}, \boldsymbol{x})]$ where $\psi(\boldsymbol{y}, \boldsymbol{x})$ is the function

$$\psi(\boldsymbol{y}, \boldsymbol{x}) \triangleq \sum_j \left(y_j \log\left[\sigma_j(\boldsymbol{x})\right] + (1 - y_j) \log\left[1 - \sigma_j(\boldsymbol{x})\right]\right). \tag{21}$$

Here $y_j \in \{0, 1\}$ and $\sigma_j(\boldsymbol{x}) \triangleq 1/[1+\exp(\boldsymbol{\phi}_{j\cdot}^T \boldsymbol{x})]$, with $\boldsymbol{\phi}_{j\cdot}$ denoting the j-th row of Φ. This function may be naturally substituted into the $\boldsymbol{x}$-space Type II cost function (8) giving us the candidate penalized logistic regression function

$$\min_{\boldsymbol{x}} \; \psi(\boldsymbol{y}, \boldsymbol{x}) + \lambda g_{(II)}(\boldsymbol{x}). \tag{22}$$

Importantly, recasting Type II classification using $\boldsymbol{x}$-space in this way, with its attendant well-specified cost function, facilitates more concrete analyses (see below) regarding properties of global and local minima that were previously rendered inaccessible because of intractable integrals and compensatory approximations. Moreover, we retain a tight connection with the original Type II marginalization process as follows.

Consider the strict upper bound on the function $\psi(\boldsymbol{y}, \boldsymbol{x})$ (obtained by a Taylor series approximation and a Hessian bound) given by

$$\psi(\boldsymbol{y}, \boldsymbol{x}) \leq \pi(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{v}) \triangleq \psi(\boldsymbol{y}, \boldsymbol{v}) + (\boldsymbol{v} - \boldsymbol{x})^T \Phi^T \boldsymbol{t} + 1/8 \, (\boldsymbol{v} - \boldsymbol{x})^T \Phi^T \Phi \, (\boldsymbol{v} - \boldsymbol{x}), \tag{23}$$

where $\boldsymbol{t} = [t_1, \ldots, t_n]^T$ with $t_j \triangleq y_j - \sigma_j(\boldsymbol{v})$. This bound holds for all $\boldsymbol{v}$ with equality when $\boldsymbol{v} = \boldsymbol{x}$. Using this result we obtain the lower bound on the marginal likelihood given by $\int \log[-\psi(\boldsymbol{y}, \boldsymbol{x})] p(\boldsymbol{x}) d\boldsymbol{x} \geq \int \log[-\pi(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{v})] p(\boldsymbol{x}) d\boldsymbol{x}$. The dual-space framework can then be used to derive the following result:

Theorem 3. Minimization of (22) with $\lambda = 4$ is equivalent to solving

$$\max_{\boldsymbol{v};\boldsymbol{\gamma}\succeq 0} \int \exp\left[-\pi(\boldsymbol{y},\boldsymbol{x},\boldsymbol{v})\right] \prod_i \mathcal{N}(\boldsymbol{x};0,\gamma_i)\varphi(\boldsymbol{\gamma}_i)dx_i \tag{24}$$

and then computing $\boldsymbol{x}_{(II)}$ by plugging the resulting $\boldsymbol{\gamma}$ into (6).

Thus we may conclude that (22) provides a principled approximation to (5) when a Bernoulli likelihood function is used for classification purposes. In empirical tests on benchmark data sets (see supplementary material) using $f(\gamma) = 0$, it performs nearly identically to the original RVM (which also implicitly assumes $f(\gamma) = 0$), but nonetheless provides a more solid theoretical justification for Type II classifiers because of the underlying similarities and identical generative model. But while the RVM and its attendant approximations are difficult to analyze, (22) is relatively transparent. Additionally, for other sparse priors, or equivalently other selections for f, we can still perform optimization and analyze cost functions without any conjugacy requirements on the implicit $p(\boldsymbol{x})$.

Theorem 4. If $\log|\Sigma_y| + \sum_i f(\gamma_i)$ is a concave, non-decreasing function of $\boldsymbol{\gamma}$ (as will be the case if f is concave and non-decreasing), then every local optimum of (24) is achieved at a solution with at most n nonzero elements in $\boldsymbol{\gamma}$ and therefore $\boldsymbol{x}_{(II)}$. In contrast, if $-\log p(\boldsymbol{x})$ is convex, then (24) can be globally solved via a convex program.

Despite the practical success of the RVM and related Bayesian techniques, and empirical evidence of sparse solutions, there is currently no proof that the standard variants of these classification methods will always produce exactly sparse estimates. Thus Theorem 4 provides some analytical validation of these types of classifiers.

Finally, if we take (22) as our starting point, we may naturally consider modifications tailored to specific sparse classification tasks (that may or may not retain an explicit connection with the original Type II probabilistic model). For example, suppose we would like to obtain a maximally sparse classifier, where regularization is provided by a $\|\boldsymbol{x}\|_0$ penalty. Direct optimization is combinatorial because of what we call the *global zero attraction property*: Whenever any individual coefficient x_i goes to zero, we are necessarily at a local minimum with respect to this coefficient because of the infinite slope (discontinuity) of the ℓ_0 norm at zero. However, (22) can be modified to approximate the ℓ_0 without this property as follows.

Theorem 5. Consider the Type II-inspired minimization problem

$$\hat{\boldsymbol{x}}, \hat{\boldsymbol{\gamma}} = \arg\min_{\boldsymbol{x};\boldsymbol{\gamma}\succeq 0} \psi\left(\boldsymbol{y},\boldsymbol{x}\right) + \alpha_1 \sum_i \frac{x_i^2}{\gamma_i} + \log\left|\alpha_2 I + \Phi\Gamma\Phi^T\right| \tag{25}$$

which is equivalent to (22) with $f(\gamma) = 0$ when $\alpha_1 = \alpha_2 = \lambda$. For some α_1 and α_2 sufficiently small (but not necessarily equal), the support[3] of $\hat{\boldsymbol{x}}$ will match the support of $\arg\min_{\boldsymbol{x}} \psi\left(\boldsymbol{y},\boldsymbol{x}\right) + \lambda\|\boldsymbol{x}\|_0$. Moreover, (25) does *not* satisfy the global zero attraction property.

Thus Type II affords the possibility of mimicking the ℓ_0 norm in the presence of generalized likelihoods but with the advantageous potential for drastically fewer local minima. This is a direction for future research. Additionally, while here we have focused our attention on classification via logistic regression, these ideas can presumably be extended to other likelihood functions provided certain conditions are met. To the best of our knowledge, while already demonstrably successful in an empirical setting, Type II classifiers and other related Bayesian generalized likelihood models have never been analyzed in the context of sparse estimation as we have done in this section.

6 Conclusion

The dual-space view of sparse linear or generalized linear models naturally allows us to transition $\boldsymbol{x}$-space ideas originally developed for Type I and apply them to Type II, and conversely, apply $\boldsymbol{\gamma}$-space techniques from Type II to Type I. The resulting symmetry promotes a mutual understanding of both methodologies and helps ensure that they are not underutilized.

[3] *Support* refers to the index set of the nonzero elements.

References

[1] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[2] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Information Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[3] E. Candès, M. Wakin, and S. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *J. Fourier Anal. Appl.*, vol. 14, no. 5, pp. 877–905, 2008.

[4] R. Chartrand and W. Yin, "Iteratively reweighted algorithms for compressive sensing," *Proc. Int. Conf. Accoustics, Speech, and Signal Proc.*, 2008.

[5] D.L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization," *Proc. National Academy of Sciences*, vol. 100, no. 5, pp. 2197–2202, March 2003.

[6] M.A.T. Figueiredo, "Adaptive sparseness using Jeffreys prior," *Advances in Neural Information Processing Systems 14*, pp. 697–704, 2002.

[7] C. Févotte and S.J. Godsill, "Blind separation of sparse sources using Jeffreys inverse prior and the EM algorithm," *Proc. 6th Int. Conf. Independent Component Analysis and Blind Source Separation*, Mar. 2006.

[8] M. Girolami, "A variational method for learning sparse and overcomplete representations," *Neural Computation*, vol. 13, no. 11, pp. 2517–2532, 2001.

[9] I.F. Gorodnitsky and B.D. Rao, "Sparse signal reconstruction from limited data using FOCUSS: A re-weighted minimum norm algorithm," *IEEE Transactions on Signal Processing*, vol. 45, no. 3, pp. 600–616, March 1997.

[10] S. Ji, D. Dunson, and L. Carin, "Multi-task compressive sensing," *IEEE Trans. Signal Processing*, vol. 57, no. 1, pp. 92–106, Jan 2009.

[11] K. Kreutz-Delgado, J. F. Murray, B.D. Rao, K. Engan, T.-W. Lee, and T.J. Sejnowski, "Dictionary learning algorithms for sparse representation," *Neural Computation*, vol. 15, no. 2, pp. 349–396, February 2003.

[12] D.J.C. MacKay, "Bayesian interpolation," *Neural Computation*, vol. 4, no. 3, pp. 415–447, 1992.

[13] J. Mattout, C. Phillips, W.D. Penny, M.D. Rugg, and K.J. Friston, "MEG source localization under multiple constraints: An extended Bayesian framework," *NeuroImage*, vol. 30, pp. 753–767, 2006.

[14] R.M. Neal, *Bayesian Learning for Neural Networks*, Springer-Verlag, New York, 1996.

[15] J.A. Palmer, D.P. Wipf, K. Kreutz-Delgado, and B.D. Rao, "Variational EM algorithms for non-Gaussian latent variable models," *Advances in Neural Information Processing Systems 18*, pp. 1059–1066, 2006.

[16] B.D. Rao, K. Engan, S. F. Cotter, J. Palmer, and K. Kreutz-Delgado, "Subset selection in noise based on diversity measure minimization," *IEEE Trans. Signal Processing*, vol. 51, no. 3, pp. 760–770, March 2003.

[17] M. Seeger and H. Nickisch, "Large scale Bayesian inference and experimental design for sparse linear models," *SIAM J. Imaging Sciences*, vol. 4, no. 1, pp. 166–199, 2011.

[18] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society*, vol. 58, no. 1, pp. 267–288, 1996.

[19] M.E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, vol. 1, pp. 211–244, 2001.

[20] M.E. Tipping and A.C. Faul, "Fast marginal likelihood maximisation for sparse Bayesian models," *Ninth Int. Workshop. Artificial Intelligence and Statistics*, Jan. 2003.

[21] D.P. Wipf "Sparse estimation with structured dictionaries," *Advances in Nerual Information Processing 24*, 2011.

[22] D.P. Wipf, B.D. Rao, and S. Nagarajan, "Latent variable Bayesian models for promoting sparsity," *IEEE Trans. Information Theory*, vol. 57, no. 9, Sept. 2011.