Bayesian Analysis (2015) **10**, Number 2, pp. 247–274

Compound Poisson Processes, Latent Shrinkage Priors and Bayesian Nonconvex Penalization

Zhihua Zhang* and Jin Li†

Abstract. In this paper we discuss Bayesian nonconvex penalization for sparse learning problems. We explore a nonparametric formulation for latent shrinkage parameters using subordinators which are one-dimensional Lévy processes. We particularly study a family of continuous compound Poisson subordinators and a family of discrete compound Poisson subordinators. We exemplify four specific subordinators: Gamma, Poisson, negative binomial and squared Bessel subordinators. The Laplace exponents of the subordinators are Bernstein functions, so they can be used as sparsity-inducing nonconvex penalty functions. We exploit these subordinators in regression problems, yielding a hierarchical model with multiple regularization parameters. We devise ECME (Expectation/Conditional Maximization Either) algorithms to simultaneously estimate regression coefficients and regularization parameters. The empirical evaluation of simulated data shows that our approach is feasible and effective in high-dimensional data analysis.

Keywords: nonconvex penalization, subordinators, latent shrinkage parameters, Bernstein functions, ECME algorithms.

1 Introduction

Variable selection methods based on penalty theory have received great attention in high-dimensional data analysis. A principled approach is due to the lasso of Tibshirani (1996), which uses the ℓ_1-norm penalty. Tibshirani (1996) also pointed out that the lasso estimate can be viewed as the mode of the posterior distribution. Indeed, the ℓ_1 penalty can be transformed into the Laplace prior. Moreover, this prior can be expressed as a Gaussian scale mixture. This has thus led to Bayesian developments of the lasso and its variants (Figueiredo, 2003; Park and Casella, 2008; Hans, 2009; Kyung et al., 2010; Griffin and Brown, 2010; Li and Lin, 2010).

There has also been work on nonconvex penalization under a parametric Bayesian framework. Zou and Li (2008) derived their local linear approximation (LLA) algorithm by combining the expectation maximization (EM) algorithm with an inverse Laplace transform. In particular, they showed that the ℓ_q penalty with $0 < q < 1$ can be obtained by mixing the Laplace distribution with a stable density. Other authors have shown that the prior induced from a penalty, called the nonconvex LOG penalty and defined in equation (2) below, has an interpretation as a scale mixture of Laplace distributions with an inverse Gamma mixing distribution (Cevher, 2009; Garrigues and Olshausen,

*Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China, zhihua@sjtu.edu.cn

†Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China, lijin@sjtu.com

© 2015 International Society for Bayesian Analysis DOI: 10.1214/14-BA892

2010; Lee et al., 2010; Armagan et al., 2013). Recently, Zhang et al. (2012) extended this class of Laplace variance mixtures by using a generalized inverse Gaussian mixing distribution. Related methods include the Bayesian hyper-lasso (Griffin and Brown, 2011), the horseshoe model (Carvalho et al., 2010, 2009) and the Dirichlet Laplace prior (Bhattacharya et al., 2012).

In parallel, nonparametric Bayesian approaches have been applied to variable selection (Ghahramani et al., 2006). For example, in the infinite Gamma Poisson model (Titsias, 2007) negative binomial processes are used to describe non-negative integer valued matrices, yielding a nonparametric Bayesian feature selection approach under an unsupervised learning setting. The beta-Bernoulli process provides a nonparametric Bayesian tool in sparsity modeling (Thibaux and Jordan, 2007; Broderick et al., 2012; Paisley and Carin, 2009; Teh and Görür, 2009). Additionally, Caron and Doucet (2008) proposed a nonparametric approach for normal variance mixtures and showed that the approach is closely related to Lévy processes. Later on, Polson and Scott (2012) constructed sparse priors using increments of subordinators, which embeds finite dimensional normal variance mixtures in infinite ones. Thus, this provides a new framework for the construction of sparsity-inducing priors. Specifically, Polson and Scott (2012) discussed the use of α-stable subordinators and inverted-beta subordinators for modeling joint priors of regression coefficients. Zhang and Tu (2012) established the connection of two nonconvex penalty functions, which are referred to as LOG and EXP and defined in equations (2) and (3) below, with the Laplace transforms of the Gamma and Poisson subordinators. A subordinator is a one-dimensional Lévy process that is almost surely non-decreasing (Sato, 1999).

In this paper we further study the application of subordinators in Bayesian nonconvex penalization problems under supervised learning scenarios. Differing from the previous treatments, we model latent shrinkage parameters using subordinators which are defined as stochastic processes of regularization parameters. In particular, we consider two families of compound Poisson subordinators: continuous compound Poisson subordinators based on a Gamma random variable (Aalen, 1992) and discrete compound Poisson subordinators based on a logarithmic random variable (Sato, 1999). The corresponding Lévy measures are generalized Gamma (Brix, 1999) and Poisson measures, respectively. We show that both the Gamma and Poisson subordinators are limiting cases of these two families of the compound Poisson subordinators.

Since the Laplace exponent of a subordinator is a Bernstein function, we have two families of nonconvex penalty functions, whose limiting cases are the nonconvex LOG and EXP. Additionally, these two families of nonconvex penalty functions can be defined via composition of LOG and EXP, while the continuous and discrete compound Poisson subordinators are mixtures of Gamma and Poisson processes.

Recall that the latent shrinkage parameter is a stochastic process of the regularization parameter. We formulate a hierarchical model with multiple regularization parameters, giving rise to a Bayesian approach for nonconvex penalization. To reduce computational expenses, we devise an ECME (for "Expectation/Conditional Maximization Either") algorithm (Liu and Rubin, 1994) which can adaptively adjust the local regularization parameters in finding the sparse solution simultaneously.

The remainder of the paper is organized as follows. Section 2 reviews the use of Lévy processes in Bayesian sparse learning problems. In Section 3 we study two families of compound Poisson processes. In Section 4 we apply the Lévy processes to Bayesian linear regression and devise an ECME algorithm for finding the sparse solution. We conduct empirical evaluations using simulated data in Section 5, and conclude our work in Section 6.

2 Problem Formulation

Our work is based on the notion of Bernstein and completely monotone functions as well as subordinators.

Definition 1. *Let $g \in C^\infty(0,\infty)$ with $g \geq 0$. The function g is said to be completely monotone if $(-1)^n g^{(n)} \geq 0$ for all $n \in \mathbb{N}$ and Bernstein if $(-1)^n g^{(n)} \leq 0$ for all $n \in \mathbb{N}$.*

Roughly speaking, a *subordinator* is a one-dimensional Lévy process that is non-decreasing almost surely. Our work is mainly motivated by the property of subordinators given in Lemma 1 (Sato, 1999; Applebaum, 2004).

Lemma 1. *If $T = \{T(t) : t \geq 0\}$ is a subordinator, then the Laplace transform of its density takes the form*

$$\mathbb{E}\big(e^{-sT(t)}\big) = \int_0^\infty e^{-s\eta} f_{T(t)}(\eta) d\eta \triangleq e^{-t\Psi(s)} \quad \text{for } s > 0,$$

where $f_{T(t)}$ is the density of $T(t)$ and Ψ, defined on $(0,\infty)$, is referred to as the Laplace exponent *of the subordinator and has the following representation*

$$\Psi(s) = \beta s + \int_0^\infty \big[1 - e^{-su}\big]\nu(du). \tag{1}$$

Here $\beta \geq 0$ and ν is the Lévy measure such that $\int_0^\infty \min(u,1)\nu(du) < \infty$.

Conversely, if Ψ is an arbitrary mapping from $(0,\infty) \to (0,\infty)$ given by expression (1), then $e^{-t\Psi(s)}$ is the Laplace transform of the density of a subordinator.

It is well known that the Laplace exponent Ψ is Bernstein and the corresponding Laplace transform $\exp(-t\Psi(s))$ is completely monotone for any $t \geq 0$ (Applebaum, 2004). Moreover, any function $g : (0,\infty) \to \mathbb{R}$, with $g(0) = 0$, is a Bernstein function if and only if it has the representation as in expression (1). Clearly, Ψ as defined in expression (1) satisfies $\Psi(0) = 0$. As a result, Ψ is nonnegative, nondecreasing and concave on $(0,\infty)$.

2.1 Subordinators for Nonconvex Penalty Functions

We are given a set of training data $\{(\mathbf{x}_i, y_i) : i = 1,\ldots,n\}$, where the $\mathbf{x}_i \in \mathbb{R}^p$ are the input vectors and the y_i are the corresponding outputs. We now discuss the following

linear regression model:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \boldsymbol{\epsilon},$$

where $\mathbf{y} = (y_1, \ldots, y_n)^T$, $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T$, and $\boldsymbol{\epsilon}$ is a Gaussian error vector $N(\mathbf{0}, \sigma\mathbf{I}_n)$. We aim at finding a sparse estimate of the vector of regression coefficients $\mathbf{b} = (b_1, \ldots, b_p)^T$ by using a Bayesian nonconvex approach.

In particular, we consider the following hierarchical model for the regression coefficients b_j's:

$$\begin{aligned} p(b_j|\eta_j, \sigma) &\propto \exp(-\eta_j|b_j|/\sigma), \\ [\eta_j] &\overset{iid}{\sim} p(\eta_j), \\ \sigma &\sim \mathrm{IG}(\alpha_\sigma/2, \beta_\sigma/2), \end{aligned}$$

where the η_j's are referred to as latent shrinkage parameters, and the inverse Gamma prior has the following parametrization:

$$\mathrm{IG}(\alpha_\sigma/2, \beta_\sigma/2) = \frac{(\beta_\sigma/2)^{\alpha_\sigma/2}}{\Gamma(\frac{\alpha_\sigma}{2})}\sigma^{-(\frac{\alpha_\sigma}{2}+1)}\exp\Big(-\frac{\beta_\sigma}{2\sigma}\Big).$$

Furthermore, we regard η_j as $T(t_j)$, that is, $\eta_j = T(t_j)$. Here $\{T(t) : t \geq 0\}$ is defined as a subordinator.

Let $\Psi(s)$, defined on $(0, \infty)$, be the Laplace exponent of the subordinator. Taking $s = |b|$, it can be shown that $\Psi(|b|)$ defines a nonconvex penalty function of b on $(-\infty, \infty)$. Moreover, $\Psi(|b|)$ is nondifferentiable at the origin because $\Psi'(0^+) > 0$ and $\Psi'(0^-) < 0$. Thus, it is able to induce sparsity. In this regard, $\exp(-t\Psi(|b|))$ forms a prior for b. From Lemma 1 it follows that the prior can be defined via the Laplace transform. In summary, we have the following theorem.

Theorem 1. *Let Ψ be a nonzero Bernstein function on $(0, \infty)$. If $\lim_{s\to 0+}\Psi(s) = 0$, then $\Psi(|b|)$ is a nondifferentiable and nonconvex function of b on $(-\infty, \infty)$. Furthermore,*

$$\exp(-t\Psi(|b|)) = \int_0^\infty \exp(-|b|\eta) f_{T(t)}(\eta)d\eta,\ t \geq 0,$$

where $\{T(t) : t \geq 0\}$ is some subordinator.

Recall that $T(t)$ is defined as the latent shrinkage parameter η and in Section 4 we will see that t plays the same role as the regularization parameter (or tuning hyperparameter). Thus, there is an important connection between the latent shrinkage parameter and the corresponding regularization parameter; that is, $\eta = T(t)$. Because $\eta_j = T(t_j)$, each latent shrinkage parameter η_j corresponds to a local regularization parameter t_j. Therefore we have a nonparametric Bayesian formulation for the latent shrinkage parameters η_j's.

It is also worth pointing out that

$$\exp(-t\Psi(|b|)) = 2\int_0^\infty L(b|0, (2\eta)^{-1})\eta^{-1} f_{T(t)}(\eta)d\eta,$$

where $L(b|u,\eta)$ denotes a Laplace distribution with density given by

$$p(b|u,\eta) = \frac{1}{4\eta}\exp\Big(-\frac{1}{2\eta}|b-u|\Big).$$

Thus, if $0 < \int_0^\infty \eta^{-1} f_{T(t)}(\eta)d\eta = M < \infty$, then $f_{T^*(t)} \triangleq \eta^{-1} f_{T(t)}(\eta)/M$ defines the proper density of some random variable (denoted $T^*(t)$). Subsequently, we obtain a proper prior $\exp(-t\Psi(|b|))/M$ for b. Moreover, this prior can be regarded as a Laplace scale mixture, i.e., the mixture of $L(b|0,(2\eta)^{-1})$ with mixing distribution $f_{T^*(t)}(\eta)$. If $\int_0^\infty \eta^{-1} f_{T(t)}(\eta)d\eta = \infty$, then $f_{T^*(t)}$ is not a proper density. Thus, $\exp(-t\Psi(|b|))$ is also improper as a prior of b. However, we still treat $\exp(-t\Psi(|b|))$ as the mixture of $L(b|0,(2\eta)^{-1})$ with mixing distribution $f_{T^*(t)}(\eta)$. In this case, we employ the terminology of pseudo-priors for the density, which is also used by Polson and Scott (2011).

2.2 The Gamma and Poisson Subordinators

Obviously, $\Psi(s) = s$ is Bernstein. It is an extreme case, because we have that $\beta = 1$, $\nu(du) = \delta_0(u)du$ and that $f_{T(t)}(\cdot) = \delta_t(\cdot)$, where $\delta_t(\cdot)$ denotes the Dirac Delta measure at t, which corresponds to the deterministic process $T(t) = t$. We can exclude this case by assuming $\beta = 0$ in expression (1) to obtain a strictly concave Bernstein function. In fact, we can impose the condition $\lim_{s\to\infty} \frac{\Psi(s)}{s} = 0$. This in turn leads to $\beta = 0$ due to $\lim_{s\to\infty} \frac{\Psi(s)}{s} = \beta$. In this paper we exploit Laplace exponents in nonconvex penalization problems. For this purpose, we will only consider a subordinator without drift, i.e., $\beta = 0$. Equivalently, we always assume that $\lim_{s\to\infty} \frac{\Psi(s)}{s} = 0$.

We here take the nonconvex LOG and EXP penalties as two concrete examples (also see Zhang and Tu, 2012). The LOG penalty is defined by

$$\Psi(s) = \frac{1}{\xi}\log\big(\gamma s+1\big), \quad \gamma,\ \xi > 0, \tag{2}$$

while the EXP penalty is given by

$$\Psi(s) = \frac{1}{\xi}(1-\exp(-\gamma s)), \quad \gamma,\ \xi > 0. \tag{3}$$

Clearly, these two functions are Bernstein on $(0,\infty)$. Moreover, they satisfy $\Psi(0) = 0$ and $\lim_{s\to\infty} \frac{\Psi(s)}{s} = \lim_{s\to\infty} \Psi'(s) = 0$. It is also directly verified that

$$\frac{1}{\xi}\log\big(\gamma s+1\big) = \int_0^\infty \big[1-\exp(-su)\big]\nu(du),$$

where the Lévy measure ν is given by

$$\nu(du) = \frac{1}{\xi u}\exp(-u/\gamma)du.$$

The corresponding subordinator $\{T(t) : t \geq 0\}$ is a Gamma subordinator, because each $T(t)$ follows a Gamma distribution with parameters $(t/\xi, \gamma)$, with density given by

$$f_{T(t)}(\eta) = \frac{\gamma^{-\frac{t}{\xi}}}{\Gamma(t/\xi)} \eta^{\frac{t}{\xi}-1} \exp(-\gamma^{-1}\eta) \text{ (also denoted } \mathrm{Ga}(t/\xi, \gamma)).$$

We also note that the corresponding pseudo-prior is given by

$$\exp(-t\Psi(|b|)) = \big(\gamma|b|+1\big)^{-t/\xi} \propto \int_0^\infty L(b|0, \eta^{-1})\eta^{-1} f_{T(t)}(\eta) d\eta.$$

Furthermore, if $t > \xi$, the pseudo-prior is a proper distribution, which is the mixture of $L(b|0, \eta^{-1})$ with mixing distribution $\mathrm{Ga}(\eta|\xi^{-1}t-1, \gamma)$.

As for the EXP penalty, the Lévy measure is $\nu(du) = \xi^{-1}\delta_\gamma(u)du$. Since

$$\int_{\mathbb{R}} \big[1-\exp(-\gamma|b|)\big] db = \infty,$$

then $\xi^{-1}[1 - \exp(-\gamma|b|)]$ is an improper prior of b. Additionally, $\{T(t) : t \geq 0\}$ is a Poisson subordinator. Specifically, $T(t)$ is a Poisson distribution with intensity $1/\xi$ taking values on the set $\{k\gamma : k \in \mathbb{N} \cup \{0\}\}$. That is,

$$\Pr(T(t) = k\gamma) = \frac{(t/\xi)^k}{k!} e^{-t/\xi}, \text{ for } k = 0, 1, 2, \ldots \tag{4}$$

which we denote by $T(t) \sim \mathrm{Po}(1/\xi)$.

3 Compound Poisson Subordinators

In this section we explore the application of compound Poisson subordinators in constructing nonconvex penalty functions. Let $\{Z(k) : k \in \mathbb{N}\}$ be a sequence of independent and identically distributed (i.i.d.) real valued random variables with common law μ_Z, and let $K \in \mathbb{N} \cup \{0\}$ be a Poisson process with intensity λ that is independent of all the $Z(k)$. Then $T(t) \triangleq Z(K(1)) + \cdots + Z(K(t))$, for $t \geq 0$, follows a compound Poisson distribution with density $f_{T(t)}(\eta)$ (denoted $\mathrm{CP}(\lambda t, \mu_Z)$), and hence $\{T(t) : t \geq 0\}$ is called a compound Poisson process. A compound Poisson process is a subordinator if and only if the $Z(k)$ are nonnegative random variables (Sato, 1999). It is worth pointing out that if $\{T(t) : t \geq 0\}$ is the Poisson subordinator given in expression (4), it is equivalent to saying that $T(t)$ follows $\mathrm{CP}(t/\xi, \delta_\gamma)$.

We particularly study two families of nonnegative random variables $Z(i)$: nonnegative continuous random variables and nonnegative discrete random variables. Accordingly, we have continuous and discrete compound Poisson subordinators $\{T(t) : t \geq 0\}$. We will show that both the Gamma and Poisson subordinators are limiting cases of the compound Poisson subordinators.

3.1 Compound Poisson Gamma Subordinators

In the first family $Z(i)$ is a Gamma random variable. In particular, let $\lambda = \frac{\rho+1}{\rho\xi}$ and the $Z(i)$ be i.i.d. from the $\mathrm{Ga}\big(\rho, \frac{\rho+1}{\gamma}\big)$ distribution, where $\rho > 0$, $\xi > 0$ and $\gamma > 0$. The compound Poisson subordinator can be written as follows

$$T(t) = \begin{cases} Z(K(1)) + \cdots + Z(K(t)) & \text{if } K(t) > 0, \\ 0 & \text{if } K(t) = 0. \end{cases}$$

The density of the subordinator is then given by

$$f_{T(t)}(\eta) = \exp\Big(-\frac{(\rho+1)t}{\rho\xi}\Big)\Big\{\delta_0(\eta) + \exp\Big(-\frac{(\rho+1)\eta}{\gamma}\Big)\sum_{k=1}^{\infty}\frac{(\rho+1)^{k(\rho+1)}(\frac{t}{\xi})^k(\frac{\eta}{\gamma})^{k\rho}}{k!\rho^k\Gamma(k\rho)\eta}\Big\}. \quad (5)$$

We denote it by $\mathrm{PG}(t/\xi, \gamma, \rho)$. The mean and variance are

$$\mathbb{E}(T(t)) = \frac{\gamma t}{\xi} \quad \text{and} \quad \mathbb{V}ar(T(t)) = \frac{\gamma^2 t}{\xi},$$

respectively. The Laplace transform is given by

$$\mathbb{E}(\exp(-sT(t))) = \exp(-t\Psi_\rho(s)),$$

where Ψ_ρ is a Bernstein function of the form

$$\Psi_\rho(s) = \frac{\rho+1}{\rho\xi}\Big[1 - \big(1 + \frac{\gamma}{\rho+1}s\big)^{-\rho}\Big]. \quad (6)$$

The corresponding Lévy measure is given by

$$\nu(du) = \frac{\gamma}{\xi}\frac{((\rho+1)/\gamma)^{\rho+1}}{\Gamma(\rho+1)}u^{\rho-1}\exp\Big(-\frac{\rho+1}{\gamma}u\Big)du. \quad (7)$$

Notice that $\frac{\xi}{\gamma}u\nu(du)$ is a Gamma measure for the random variable u. Thus, the Lévy measure $\nu(du)$ is referred to as a generalized Gamma measure (Brix, 1999).

The Bernstein function $\Psi_\rho(s)$ was studied by Aalen (1992) for survival analysis. However, we consider its application in sparsity modeling. It is clear that $\Psi_\rho(s)$ for $\rho > 0$ and $\gamma > 0$ satisfies the conditions $\Psi_\rho(0) = 0$ and $\lim_{s\to\infty}\frac{\Psi_\rho(s)}{s} = 0$. Also, $\Psi_\rho(|b|)$ is a nonnegative and nonconvex function of b on $(-\infty, \infty)$, and it is an increasing function of $|b|$ on $[0, \infty)$. Moreover, $\Psi_\rho(|b|)$ is continuous w.r.t. b but nondifferentiable at the origin. This implies that $\Psi_\rho(|b|)$ can be treated as a sparsity-inducing penalty.

We are interested in the limiting cases that $\rho = 0$ and $\rho = +\infty$.

Proposition 1. *Let* $\mathrm{PG}(t/\xi, \gamma, \rho)$, $\Psi_\rho(s)$ *and* $\nu(du)$ *be defined by expressions (5), (6) and (7), respectively. Then*

(1) $\lim_{\rho\to 0+}\Psi_\rho(s) = \frac{1}{\xi}\log\big(\gamma s+1\big)$ *and* $\lim_{\rho\to\infty}\Psi_\rho(s) = \frac{1}{\xi}(1 - \exp(-\gamma s))$;

(2) $\lim_{\rho\to 0+} \mathrm{PG}(t/\xi,\gamma,\rho) = \mathrm{Ga}(t/\xi,\gamma)$ *and* $\lim_{\rho\to\infty} \mathrm{PG}(t/\xi,\gamma,\rho) = \mathrm{CP}(t/\xi,\delta_\gamma)$;

(3) $\lim_{\rho\to 0+} \nu(du) = \frac{1}{\xi u}\exp(-\frac{u}{\gamma})du$ *and* $\lim_{\rho\to\infty} \nu(du) = \frac{1}{\xi}\delta_\gamma(u)du$.

This proposition can be obtained by using direct algebraic computations. Proposition 1 tells us that the limiting cases yield the nonconvex LOG and EXP functions. Moreover, we see that $T(t)$ converges in distribution to a Gamma random variable with shape t/ξ and scale γ, as $\rho \to 0+$, and to a Poisson random variable with mean t/ξ, as $\rho \to \infty$.

It is well known that Ψ_0 degenerates to the LOG function (Aalen, 1992; Brix, 1999). Here we have shown that Ψ_ρ approaches to EXP as $\rho \to \infty$. We list another special example in Table 1 when $\rho = 1$. We refer to the corresponding penalty as a *linear-fractional* (LFR) function. For notational simplicity, we respectively replace $\gamma/2$ and $\xi/2$ by γ and ξ in the LFR function. The density of the subordinator for the LFR function is given by

$$f_{T(t)}(\eta) = e^{-\frac{t}{\xi}}\Big\{\delta_0(\eta) + e^{-\frac{\eta}{\gamma}}\frac{\sqrt{t/\xi}I_1\big(2\sqrt{t\eta/(\xi\gamma)}\big)}{\gamma\sqrt{\eta/\gamma}}\Big\}.$$

We thus say each $T(t)$ follows a squared Bessel process without drift (Yuan and Kalbfleisch, 2000), which is a mixture of a Dirac delta measure and a randomized Gamma distribution (Feller, 1971). We denote the density of $T(t)$ by $\mathrm{SB}(t/\xi,\gamma)$.

Table 1: Bernstein functions LOG, EXP, LFR, and CEL, defined on $[0,\infty)$, and the corresponding Lévy measures and subordinators ($\xi > 0$ and $\gamma > 0$).

Bernstein Functions	Lévy Measures $\nu(du)$	Subordinators $T(t)$	Priors
LOG $\Psi_0(s) = \Phi_0(s) = \frac{1}{\xi}\log(\gamma s+1)$	$\frac{1}{\xi u}\exp(-\frac{u}{\gamma})du$	$\mathrm{Ga}(t/\xi,\gamma)$	Proper[a]
EXP $\Psi_\infty(s) = \Phi_\infty(s) = \frac{1}{\xi}[1-\exp(-\gamma s)]$	$\frac{1}{\xi}\delta_\gamma(u)du$	$\mathrm{CP}(t/\xi,\delta_{k\gamma})$	Improper
LFR $\Psi_1(s) = \frac{1}{\xi}\frac{\gamma s}{\gamma s+1}$	$\frac{1}{\xi\gamma}\exp(-\frac{u}{\gamma})du$	$\mathrm{SB}(t/\xi,\gamma)$	Improper
CEL $\Phi_1(s) = \frac{1}{\xi}\log[2-\exp(-\gamma s)]$	$\frac{1}{\xi}\sum_{k=1}^{\infty}\frac{1}{k2^k}\delta_{k\gamma}(u)du$	$\mathrm{NB}(t/\xi,1/2,\delta_{k\gamma})$	Improper

[a]It is proper only when $t > \xi$.

3.2 Negative Binomial Subordinators

In the second case, we consider a family of discrete compound Poisson subordinators. Particularly, $Z(i)$ is discrete and takes values on $\{k\alpha : k \in \mathbb{N}\cup\{0\}\}$. And it is defined as logarithmic distribution $\log(1-q)$, where $\alpha \neq 0$ and $q \in (0,1)$, with probability mass function given by

$$\Pr(Z(i) = k\alpha) = -\frac{(1-q)^k}{k\log(q)}.$$

Moreover, we let $K(t)$ have a Poisson distribution with intensity $-(\rho{+}1)\log(q)/\xi$, where $\rho > 0$. Then $T(t)$ is distributed according to a negative binomial (NB) distribution (Sato, 1999). The probability mass function of $T(t)$ is given by

$$\Pr(T(t) = k\alpha) = \frac{\Gamma(k{+}(\rho+1)t/\xi)}{k!\Gamma((\rho+1)t/\xi)} q^{\frac{(\rho+1)t}{\xi}} (1-q)^k, \tag{8}$$

which is denoted as $\mathrm{NB}((\rho{+}1)t/\xi, q, \delta_{k\alpha})$. We thus say that $T(t)$ follows an NB subordinator. Let $q = \frac{\rho}{\rho+1}$ and $\alpha = \frac{\rho}{\rho+1}\gamma$. It can be verified that $\mathrm{NB}\big((\rho{+}1)t/\xi, \frac{\rho}{\rho+1}, \delta_{\frac{k\gamma\rho}{\rho+1}}\big)$ has the same mean and variance as the $\mathrm{PG}(t/\xi, \gamma, \rho)$ distribution. The corresponding Laplace transform then gives rise to a new family of Bernstein functions, which is given by

$$\Phi_\rho(s) \triangleq \frac{\rho{+}1}{\xi} \log\Big[\frac{1{+}\rho}{\rho} - \frac{1}{\rho}\exp(-\frac{\rho}{\rho{+}1}\gamma s)\Big]. \tag{9}$$

We refer to this family of Bernstein functions as *compound EXP-LOG* (CEL) functions. The first-order derivative of $\Phi_\rho(s)$ w.r.t. s is given by

$$\Phi'_\rho(s) = \frac{\gamma}{\xi} \frac{\rho\exp(-\frac{\rho}{\rho+1}\gamma s)}{1{+}\rho - \exp(-\frac{\rho}{\rho+1}\gamma s)}.$$

The Lévy measure for $\Phi_\rho(s)$ is given by

$$\nu(du) = \frac{\rho+1}{\xi} \sum_{k=1}^{\infty} \frac{1}{k(1{+}\rho)^k} \delta_{\frac{k\gamma\rho}{\rho+1}}(u)du. \tag{10}$$

The proof is given in Appendix 1. We call this Lévy measure a *generalized Poisson measure* relative to the generalized Gamma measure.

Like $\Psi_\rho(s)$, $\Phi_\rho(s)$ can define a family of sparsity-inducing nonconvex penalties. Also, $\Phi_\rho(s)$ for $\rho > 0$, $\xi > 0$ and $\gamma > 0$ satisfies the conditions $\Phi_\rho(0) = 0$, $\lim_{s\to\infty} \frac{\Phi_\rho(s)}{s} = 0$ and $\lim_{s\to 0} \Phi'_\rho(s) = \frac{\gamma}{\xi}$. We present a special CEL function Φ_1 as well as the corresponding $T(t)$ and $\nu(du)$ in Table 1, where we replace $\xi/2$ and $\gamma/2$ by ξ and γ for notational simplicity. We now consider the limiting cases.

Proposition 2. *Assume $\nu(du)$ is defined by expression (10) for fixed $\xi > 0$ and $\gamma > 0$. Then we have that*

(a) $\lim_{\rho\to\infty} \Phi_\rho(s) = \frac{1}{\xi}(1 - \exp(-\gamma s))$ *and* $\lim_{\rho\to 0+} \Phi_\rho(s) = \frac{1}{\xi}\log(1+\gamma s)$.

(b) $\lim_{\rho\to\infty} \Phi'_\rho(s) = \frac{\gamma}{\xi}\exp(-\gamma s)$ *and* $\lim_{\rho\to 0+} \Phi'_\rho(s) = \frac{\gamma}{\xi}\frac{1}{1+\gamma s}$.

(c) $\lim_{\rho\to\infty} \nu(du) = \frac{1}{\xi}\delta_\gamma(u)du$ *and* $\lim_{\rho\to 0+} \nu(du) = \frac{1}{\xi u}\exp(-\frac{u}{\gamma})du$.

(d) $\lim_{\rho\to\infty} \mathrm{NB}\big((\rho{+}1)t/\xi, \rho/(\rho{+}1), \delta_{k\rho\gamma/(\rho+1)}\big) = \mathrm{CP}(t/\xi, \delta_\gamma)$ *and*

$$\lim_{\rho\to 0+} \Pr(T(t) \le \eta) = \int_0^\eta \frac{\gamma^{-t/\xi}}{\Gamma(t/\xi)} u^{\frac{t}{\xi}-1} \exp(-\frac{u}{\gamma}) du.$$

Notice that

$$\lim_{\rho\to 0+} \int_0^\infty \exp(-us) u\nu(du) = \lim_{\rho\to 0+} \Phi'_\rho(s) = \frac{\gamma}{\xi}\frac{1}{1{+}\gamma s} = \frac{1}{\xi}\int_0^\infty \exp\Big(-us - \frac{u}{\gamma}\Big) du.$$

This shows that $\nu(du)$ converges to $\frac{1}{\xi}u^{-1}\exp(-\frac{u}{\gamma})$, as $\rho \to 0$. Analogously, we obtain the second part of Proposition 2-(d), which implies that as $\rho \to 0$, $T(t)$ converges in distribution to a Gamma random variable with shape parameter t/ξ and scale parameter γ. An alternative proof is given in Appendix 2.

Proposition 2 shows that $\Phi_\rho(s)$ degenerates to EXP as $\rho \to \infty$, while to LOG as $\rho \to 0$. This shows an interesting connection between $\Psi_\rho(s)$ in expression (6) and $\Phi_\rho(s)$ in expression (9); that is, they have the same limiting behaviors.

3.3 Gamma/Poisson Mixture Processes

We note that for $\rho > 0$,

$$\Psi_\rho(s) = \frac{\rho{+}1}{\rho\xi}\Big[1 - \exp\Big(-\rho\log(\frac{\gamma s}{\rho{+}1} + 1)\Big)\Big]$$

which is a composition of the LOG and EXP functions, and that

$$\Phi_\rho(s) = \frac{\rho{+}1}{\xi}\log\Big[1{+}\frac{1}{\rho}(1{-}\exp(-\frac{\rho}{\rho{+}1}\gamma s))\Big]$$

which is a composition of the EXP and LOG functions. In fact, the composition of any two Bernstein functions is still Bernstein. Thus, the composition is also the Laplace exponent of some subordinator, which is then a mixture of the subordinators corresponding to the original two Bernstein functions (Sato, 1999). This leads us to an alternative derivation for the subordinators corresponding to Ψ_ρ and Φ_ρ. That is, we have the following theorem whose proof is given in Appendix 3.

Theorem 2. *The subordinator $T(t)$ associated with $\Psi_\rho(s)$ is distributed according to the mixture of* $\mathrm{Ga}(k\rho, \gamma/(\rho{+}1))$ *distributions with* $\mathrm{Po}(k|(\rho{+}1)t/(\rho\xi))$ *mixing, while $T(t)$ associated with $\Phi_\rho(s)$ is distributed according to the mixture of* $\mathrm{CP}(\lambda, \delta_{k\rho\gamma/(\rho+1)})$ *distributions with* $\mathrm{Ga}(\lambda|(\rho{+}1)t/\xi, 1/\rho)$ *mixing.*

Additionally, the following theorem illustrates a limiting property of the subordinators as γ approaches 0.

Theorem 3. *Let ρ be a fixed constant on $[0,\infty]$.*

(a) *If $T(t) \sim \mathrm{PG}(t/\xi,\gamma,\rho)$ where $\xi = \frac{\rho+1}{\rho}\Big[1-(1+\frac{\gamma}{\rho+1})^{-\rho}\Big]$ or $\xi=\gamma$, then $T(t)$ converges in probability to t, as $\gamma\to 0$.*

(b) *If $T(t)\sim \mathrm{NB}((\rho+1)t/\xi,\rho/(\rho+1),\delta_{k\rho\gamma/(\rho+1)})$ where*

$$\xi=(\rho+1)\log\Big[1+\frac{1}{\rho}(1-\exp(-\frac{\rho}{\rho+1}\gamma))\Big]$$

or $\xi=\gamma$, then $T(t)$ converges in probability to t, as $\gamma\to 0$.

The proof is given in Appendix 4. Since "$T(t)$ converges in probability to t" implies "$T(t)$ converges in distribution to t," we have that

$$\lim_{\gamma\to 0}\mathrm{PG}(t/\xi,\gamma,\rho)\overset{d}{=}\delta_t \text{ and } \lim_{\gamma\to 0}\mathrm{NB}((\rho+1)t/\xi,\rho/(\rho+1),\delta_{k\rho\gamma/(\rho+1)})\overset{d}{=}\delta_t.$$

Finally, consider the four nonconvex penalty function given in Table 1. We present the following property. That is, when $\xi=\gamma$ and for any fixed $\gamma>0$, we have

$$\frac{1}{\gamma}\log[2-\exp(-\gamma s)]\le\frac{s}{\gamma s+1}\le\frac{1}{\gamma}[1-\exp(-\gamma s)]\le\frac{1}{\gamma}\log\big(\gamma s+1\big)\le s, \tag{11}$$

with equality only when $s=0$. The proof is given in Appendix 5. This property is also illustrated in Figure 1.

4 Bayesian Linear Regression with Latent Subordinators

We apply the compound Poisson subordinators to the Bayesian sparse learning problem given in Section 2. Defining $T(t)=\eta$, we rewrite the hierarchical representation for the joint prior of the b_j under the regression framework. That is, we assume that

$$\begin{aligned}[b_j|\eta_j,\sigma] &\overset{ind}{\sim} L(b_j|0,\sigma(2\eta_j)^{-1}),\\ f_{T^*(t_j)}(\eta_j) &\propto \eta_j^{-1}f_{T(t_j)}(\eta_j),\end{aligned}$$

which implies that

$$p(b_j,\eta_j|t_j,\sigma)\propto\sigma^{-1}\exp\Big(-\frac{\eta_j}{\sigma}|b_j|\Big)f_{T(t_j)}(\eta_j).$$

The joint marginal pseudo-prior of the b_j's is given by

$$\begin{aligned}p^*(\mathbf{b}|\mathbf{t},\sigma)&=\prod_{j=1}^p p^*(b_j|t_j,\sigma)=\prod_{j=1}^p\sigma^{-1}\int_0^\infty\exp\Big(-\frac{\eta_j}{\sigma}|b_j|\Big)f_{T(t_j)}(\eta_j)d\eta_j\\&=\prod_{j=1}^p\sigma^{-1}\exp\Big(-t_j\Psi\Big(\frac{|b_j|}{\sigma}\Big)\Big).\end{aligned}$$

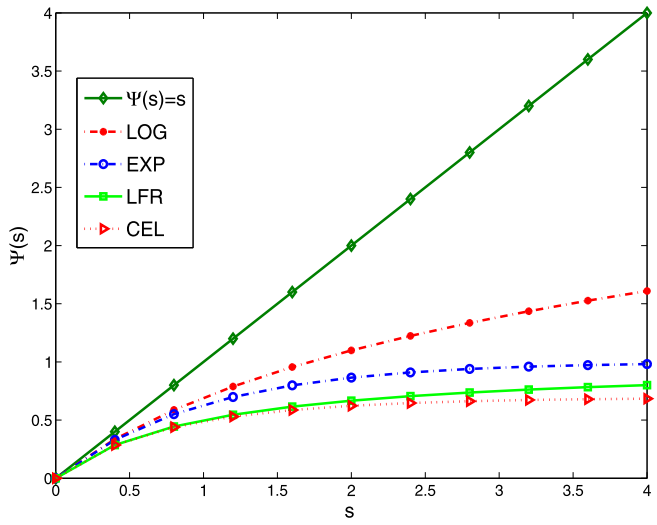


Figure 1: The four nonconvex penalties $\Psi(s)$ in Table 1 with $\xi = \gamma = 1$ and $\Psi(s) = s$.

We will see in Theorem 4 that the full conditional distribution $p(\mathbf{b}|\sigma, \mathbf{t}, \mathbf{y})$ is proper. Thus, the maximum *a posteriori* (MAP) estimate of $\mathbf{b}$ is based on the following optimization problem:

$$\min_{\mathbf{b}} \Big\{ \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2^2 + \sigma \sum_{j=1}^{p} t_j \Psi(|b_j|/\sigma) \Big\}.$$

Clearly, the t_j's are local regularization parameters and the η_j's are latent shrinkage parameters. Moreover, it is interesting that $\{T(t) : t \geq 0\}$ (or η) is defined as a subordinator w.r.t. t.

The full conditional distribution $p(\sigma|\mathbf{b}, \boldsymbol{\eta}, \mathbf{y})$ is conjugate w.r.t. the prior, which is $\sigma \sim \mathrm{IG}(\frac{a_\sigma}{2}, \frac{b_\sigma}{2})$. Specifically, it is an inverse Gamma distribution of the form

$$p(\sigma|\mathbf{b}, \boldsymbol{\eta}, \mathbf{y}) \propto \frac{1}{\sigma^{\frac{n+2p+a_\sigma}{2}+1}} \exp\Big[-\frac{b_\sigma + \|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2^2 + 2\sum_{j=1}^{p} \eta_j |b_j|}{2\sigma} \Big].$$

In the following experiment, we use an improper prior of the form $p(\sigma) \propto \frac{1}{\sigma}$ (i.e., $a_\sigma = b_\sigma = 0$). Clearly, $p(\sigma|\mathbf{b}, \boldsymbol{\eta}, \mathbf{y})$ is still an inverse Gamma distribution in this setting. Additionally, based on

$$p(\mathbf{b}|\boldsymbol{\eta}, \sigma, \mathbf{y}) \propto \exp\Big[-\frac{1}{2\sigma}\|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2^2 \Big] \prod_{j=1}^{p} \exp(-\frac{\eta_j}{\sigma}|b_j|)$$

and the proof of Theorem 4 (see Appendix 6), we have that the conditional distribution $p(\mathbf{b}|\boldsymbol{\eta}, \sigma, \mathbf{y})$ is proper. However, the absolute terms $|b_j|$ make the form of $p(\mathbf{b}|\boldsymbol{\eta}, \sigma, \mathbf{y})$ unfamiliar. Thus, a Gibbs sampling algorithm is not readily available and we resort to an EM algorithm to estimate the model.

4.1 The ECME Estimation Algorithm

Notice that if $p^*(b_j|t_j, \sigma)$ is proper, the corresponding normalizing constant is given by

$$2\int_0^\infty \sigma^{-1} \exp\Big[-t_j\Psi\Big(\frac{|b_j|}{\sigma}\Big)\Big]d|b_j| = 2\int_0^\infty \exp\Big[-t_j\Psi\Big(\frac{|b_j|}{\sigma}\Big)\Big]d(|b_j|/\sigma),$$

which is independent of σ. Also, the conditional distribution $p(\eta_j|b_j, t_j, \sigma)$ is independent of the normalizing term. Specifically, we always have that

$$p(\eta_j|b_j, t_j, \sigma) = \frac{\exp\big(-\frac{\eta_j}{\sigma}|b_j|\big) f_{T(t_j)}(\eta_j)}{\exp(-t_j\Psi(|b_j|/\sigma))},$$

which is proper.

As shown in Table 1, except for LOG with $t > \xi$ which can be transformed into a proper prior, the remaining Bernstein functions cannot be transformed into proper priors. In any case, our posterior computation is directly based on the marginal pseudo-prior $p^*(\mathbf{b}|\mathbf{t}, \sigma)$. We ignore the involved normalizing term, because it is infinite if $p^*(\mathbf{b}|\mathbf{t}, \sigma)$ is improper and it is independent of σ if $p^*(\mathbf{b}|\mathbf{t}, \sigma)$ is proper.

Given the kth estimates $(\mathbf{b}^{(k)}, \sigma^{(k)})$ of $(\mathbf{b}, \sigma)$ in the E-step of the EM algorithm, we compute

$$\begin{aligned} Q(\mathbf{b}, \sigma|\mathbf{b}^{(k)}, \sigma^{(k)}) &\triangleq \log p(\mathbf{y}|\mathbf{b}, \sigma) + \sum_{j=1}^{p}\int \log p[b_j|\eta_j, \sigma] p(\eta_j|b_j^{(k)}, \sigma^{(k)}, t_j) d\eta_j + \log p(\sigma) \\ &\propto -\frac{n+\alpha_\sigma}{2}\log\sigma - \frac{\|\mathbf{y}-\mathbf{X}\mathbf{b}\|_2^2 + \beta_\sigma}{2\sigma} - (p+1)\log\sigma \\ &\quad - \frac{1}{\sigma}\sum_{j=1}^{p}|b_j|\int \eta_j p(\eta_j|b_j^{(k)}, \sigma^{(k)}, t_j) d\eta_j. \end{aligned}$$

Here we omit some terms that are independent of parameters σ and $\mathbf{b}$. In fact, we only need to compute $\mathbb{E}(\eta_j|b_j^{(k)}, \sigma^{(k)})$ in the E-step. Considering that

$$\int_0^\infty \exp\big(-\frac{\eta_j}{\sigma}|b_j|\big) f_{T(t_j)}(\eta_j) d\eta_j = \exp(-t_j\Psi(|b_j|/\sigma)),$$

and taking the derivative w.r.t. $|b_j|$ on both sides of the above equation, we have that

$$w_j^{(k+1)} \triangleq \mathbb{E}(\eta_j|b_j^{(k)}, \sigma^{(k)}, t_j) = t_j\Psi'(|b_j^{(k)}|/\sigma^{(k)}).$$

The M-step maximizes $Q(\mathbf{b}, \sigma|\mathbf{b}^{(k)}, \sigma^{(k)})$ w.r.t. $(\mathbf{b}, \sigma)$. In particular, it is obtained that:

$$\mathbf{b}^{(k+1)} = \underset{\mathbf{b}}{\operatorname{argmin}} \ \frac{1}{2}\|\mathbf{y}-\mathbf{X}\mathbf{b}\|_2^2 + \sum_{j=1}^{p} w_j^{(k+1)}|b_j|,$$

$$\sigma^{(k+1)} = \frac{1}{n+\alpha_\sigma+2p+2}\Big\{\|\mathbf{y}-\mathbf{X}\mathbf{b}^{(k+1)}\|_2^2 + \beta_\sigma + 2\sum_{j=1}^{p} w_j^{(k+1)}|b_j^{(k+1)}|\Big\}.$$

The above EM algorithm is related to the linear local approximation (LLA) procedure (Zou and Li, 2008). Moreover, it shares the same convergence property given in Zou and Li (2008) and Zhang et al. (2012).

Subordinators help us to establish a direct connection between the local regularization parameters t_j's and the latent shrinkage parameters η_j's (or $T(t_j)$). However, when we implement the MAP estimation, it is challenging how to select these local regularization parameters. We employ an ECME (for "Expectation/Conditional Maximization Either") algorithm (Liu and Rubin, 1994; Polson and Scott, 2010) for learning about the b_j's and t_j's simultaneously. For this purpose, we suggest assigning t_j Gamma prior $\mathrm{Ga}(\alpha_t, 1/\beta_t)$, namely,

$$p(t_j)=\frac{\beta_t^{\alpha_t}}{\Gamma(\alpha_t)} t_j^{\alpha_t-1}\exp(-\beta_t t_j),$$

because the full conditional distribution is also Gamma and given by

$$[t_j|b_j,\sigma] \sim \mathrm{Ga}\big(\alpha_t, 1/[\Psi(|b_j|/\sigma)+\beta_t]\big).$$

Recall that we here compute the full conditional distribution directly using the marginal pseudo-prior $p^*(b_j|t_j,\sigma)$, because our used Bernstein functions in Table 1 cannot induce proper priors. However, if $p^*(b_j|t_j,\sigma)$ is proper, the corresponding normalizing term would rely on t_j. As a result, the full conditional distribution of t_j is possibly no longer Gamma or even not analytically available.

Figure 2-(a) depicts the hierarchical model for the Bayesian penalized linear regression, and Table 2 gives the ECME procedure where the E-step and CM-step are respectively identical to the E-step and the M-step of the EM algorithm, with $t_j = t_j^{(k)}$. The CME-step updates the t_j's with

$$t_j^{(k+1)} = \frac{\alpha_t - 1}{\Psi(|b_j^{(k)}|/\sigma^{(k)}) + \beta_t},\ j = 1,\ldots,p.$$

In order to make sure that $t_j > 0$, it is necessary to assume that $\alpha_t > 1$. In the following experiments, we set $\alpha_t = 10$.

We conduct experiments with the prior $p(\mathbf{b}_j) \propto t_j\sigma^{-1/2}\exp(-t_j(|b_j|/\sigma)^{1/2})$ for comparison. This prior is induced from the $\ell_{1/2}$-norm penalty, so it is a proper specification. Moreover, the full conditional distribution of t_j w.r.t. its Gamma prior $\mathrm{Ga}(\alpha_t, 1/\beta_t)$ is

still Gamma; that is,

$$[t_j|b_j,\sigma] \sim \text{Ga}\Big(\alpha_t{+}2,\ 1/(\beta_t{+}\sqrt{|b_j|/\sigma})\Big).$$

Thus, the CME-step for updating the t_j's is given by

$$t_j^{(k+1)} = \frac{\alpha_t + 1}{\sqrt{|b_j^{(k)}|/\sigma^{(k)}} + \beta_t},\ j = 1,\ldots,p. \tag{12}$$

The convergence analysis of the ECME algorithm was presented by Liu and Rubin (1994), who proved that the ECME algorithm retains the monotonicity property from the standard EM. Moreover, the ECME algorithm based on pseudo-priors was also used by Polson and Scott (2011).

Table 2: The Basic Procedure of the ECME Algorithm

E-step	Identical to the E-step of the EM with $t_j = t_j^{(k)}$.
CM-step	Identical to the M-step of the EM with $t_j = t_j^{(k)}$.
CME-step	Set $t_j^{(k+1)} = \frac{\alpha_t - 1}{\Psi(\lvert b_j^{(k)}\rvert/\sigma^{(k)}) + \beta_t}$.

As we have seen, $p(\mathbf{t}|\mathbf{y},\mathbf{b},\sigma) = p(\mathbf{t}|\mathbf{b},\sigma) = \prod_{j=1}^p p(t_j|b_j,\sigma)$ and $p(\sigma|\mathbf{b},\boldsymbol{\eta},\mathbf{t},\mathbf{y})$ are proper. In the following theorem, we show that $p(\mathbf{b}|\sigma,\mathbf{t},\mathbf{y})$ and $p(\mathbf{b},\sigma,\mathbf{t}|\mathbf{y})$ are also proper. Moreover, when the improper prior $p(\sigma) \propto \frac{1}{\sigma}$ (i.e., $\alpha_\sigma = \beta_\sigma = 0$ in the inverse Gamma prior) is used, Theorem 4 shows that $p(\mathbf{b},\sigma,\mathbf{t}|\mathbf{y})$ is proper under certain conditions.

Theorem 4. *With the previous prior specifications for* $\mathbf{b}$, σ *and* $\mathbf{t}$, *we have that* $p(\mathbf{b}|\sigma,\mathbf{t},\mathbf{y})$, $p(\mathbf{b},\sigma|\mathbf{t},\mathbf{y})$ *and* $p(\mathbf{b},\sigma,\mathbf{t}|\mathbf{y})$ *are proper. Suppose we use the improper prior* $p(\sigma) \propto \frac{1}{\sigma}$ *for* σ. *If* $\mathbf{y} \notin \text{range}(\mathbf{X})$ (*the subspace spanned by the columns of* $\mathbf{X}$), $p(\mathbf{b},\sigma|\mathbf{t},\mathbf{y})$ *and* $p(\mathbf{b},\sigma,\mathbf{t}|\mathbf{y})$ *are proper.*

The proof of Theorem 4 is given in Appendix 6. Notice that the proof only requires that $\Psi(s) \geq 0$, and does not involve the other properties of the Bernstein function. In other words, Theorem 4 is still held for any nonnegative but not necessarily Bernstein function Ψ. Theorem 4 shows that our ECME algorithm is to find the MAP estimates of the parameters $\mathbf{b}$ and σ as well as the MAP estimates of the local regularization parameters t_j's.

In the EM algorithm of Polson and Scott (2012), the authors set $t_1 = \cdots = t_p \triangleq \nu$ as a global regularization parameter and assumed it to be prespecified (see Section 5.3 of Polson and Scott, 2012). This in fact leads to a parametric setting for the latent shrinkage parameters η (Zou and Li, 2008; Cevher, 2009; Garrigues and Olshausen, 2010; Lee et al., 2010; Armagan et al., 2013). However, Polson and Scott (2012) aimed to construct sparse priors using increments of subordinators. It is worth noting that

Caron and Doucet (2008) regarded their model as a nonparametric model w.r.t. the regression coefficients b; that is, they treated b as a stochastic process of T. Thus, the treatment of Caron and Doucet (2008) is also different from ours.

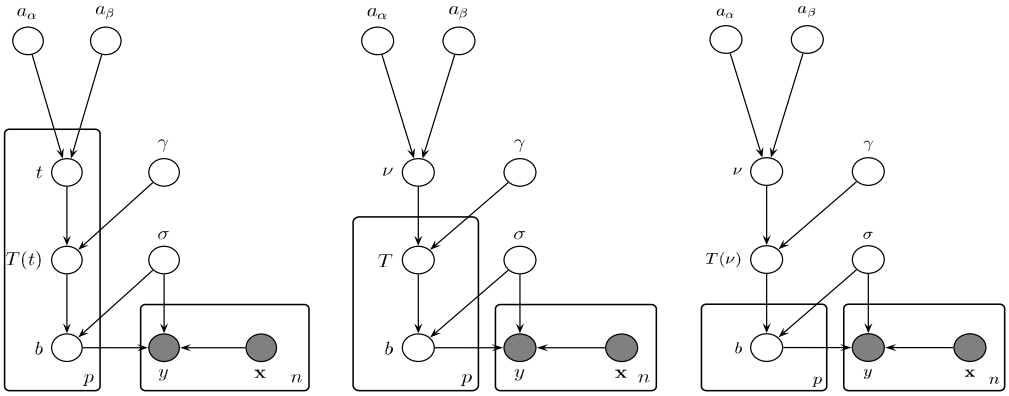


Figure 2: Graphical representations for hierarchical regression models. (a) Nonparametric setting for T, i.e., different T have different t; (b) Parametric setting for T, i.e., different T share a common ν; (c) nonseparable setting, i.e., one T.

5 Experimental Analysis

We now conduct empirical analysis with our ECME procedure described in Algorithm 1 based on Figure 2-(a). We also implement the setting in Polson and Scott (2012), i.e., $t_1 = \cdots = t_p \triangleq \nu$ and the T_j's are independent given ν. Polson and Scott (2012) suggested that ν is prespecified as the global regularization parameter. In fact, we can also estimate ν under the ECME algorithm. This setting is described in Figure 2-(b) and the corresponding ECME algorithm is given in Algorithm 2.

Notice that in the setting $t_1 = \cdots = t_p \triangleq \nu$, if the latent shrinkage $T(t)$ is treated as a stochastic process of t, then the b_j's share a common $T(\nu)$. In this case, the marginal pseudo-prior for $\mathbf{b}$ is nonseparable; that is, $p(\mathbf{b}) \propto \exp(-\frac{\nu}{\xi}\Psi(\|\mathbf{b}\|_1/\sigma))$. Figure 2-(c) illustrates the resulting model and the corresponding ECME algorithm (see Algorithm 3) is also performed.

We refer to the algorithms based on Figures 2-(a), (b) and (c) as "Alg 1," "Alg 2" and "Alg 3," respectively. We consider the nonconvex $\ell_{1/2}$, LOG, EXP, LFR and CEL penalties to respectively implement these three algorithms. The CME-step with the $\ell_{1/2}$-norm is based on expression (12). According to Theorem 3, we can set, for instance, $\xi = \frac{\gamma}{1+\gamma}$ in LFR. However, Theorem 3 also shows that when $\xi = \gamma$, the two settings have the same asymptotic properties as $\gamma \to 0$. That is, the resulting model approaches the lasso. We thus set $\xi = \gamma$ in "Alg 1," and $\xi = p\gamma$ in both "Alg 2" and "Alg 3." The settings are empirically validated to be effective. As we have mentioned,

Algorithm 1: ECME for Bayesian Regression with Penalty $\Psi_\rho(|b|)$ or $\Phi_\rho(|b|)$

E-step Given the current estimates $\mathbf{b}^{(k)}$ and $t_j = t_j^{(k)}$, compute
$w_j^{(k)} = t_j^{(k)}\Psi'_\rho(|b_j^{(k)}|/\sigma^{(k)})$ or $w_j^{(k)} = t_j^{(k)}\Phi'_\rho(|b_j^{(k)}|/\sigma^{(k)})$, $j = 1, \ldots, p$

CM-step Solve the following problem:
$\mathbf{b}^{(k+1)} = \text{argmin}_{\mathbf{b}} \ \frac{1}{2}\|\mathbf{y}-\mathbf{X}\mathbf{b}\|_2^2 + \sum_{j=1}^p w_j^{(k+1)}|b_j|$,
$\sigma^{(k+1)} = \frac{1}{\alpha_\sigma+n+2p+2}\Big\{\beta_\sigma + \|\mathbf{y}-\mathbf{X}\mathbf{b}^{(k+1)}\|_2^2 + 2\sum_{j=1}^p w_j^{(k+1)}|b_j^{(k+1)}|\Big\}$.

CME-step Compute
$t_j^{(k+1)} = \frac{\alpha_t-1}{\beta_t+\Psi_\rho(|b_j^{(k)}|/\sigma^{(k)})}$ or $t_j^{(k+1)} = \frac{\alpha_t-1}{\beta_t+\Phi_\rho(|b_j^{(k)}|/\sigma^{(k)})}$.

Algorithm 2: ECME for Bayesian Regression with Penalty $\Psi_\rho(|b|)$ or $\Phi_\rho(|b|)$

E-step Given the current estimates $\mathbf{b}^{(k)}$ and $\nu = \nu^{(k)}$, compute
$w_j^{(k)} = \nu^{(k)}\Psi'_\rho(|b_j^{(k)}|/\sigma^{(k)})$ or $w_j^{(k)} = \nu^{(k)}\Phi'_\rho(|b_j^{(k)}|/\sigma^{(k)})$, $j = 1, \ldots, p$

CM-step Solve the following problem:
$\mathbf{b}^{(k+1)} = \text{argmin}_{\mathbf{b}} \ \frac{1}{2}\|\mathbf{y}-\mathbf{X}\mathbf{b}\|_2^2 + \sum_{j=1}^p w_j^{(k+1)}|b_j|$,
$\sigma^{(k+1)} = \frac{1}{\alpha_\sigma+n+2p+2}\Big\{\beta_\sigma + \|\mathbf{y}-\mathbf{X}\mathbf{b}^{(k+1)}\|_2^2 + 2\sum_{j=1}^p w_j^{(k+1)}|b_j^{(k+1)}|\Big\}$.

CME-step Compute
$\nu^{(k+1)} = \frac{\alpha_t-1}{\beta_t+\sum_{j=1}^p \Psi_\rho(|b_j^{(k)}|/\sigma^{(k)})}$ or $\nu^{(k+1)} = \frac{\alpha_t-1}{\beta_t+\sum_{j=1}^p \Phi_\rho(|b_j^{(k)}|/\sigma^{(k)})}$.

Algorithm 3: ECME for Bayesian Regression with Penalty $\Psi_\rho(\|\mathbf{b}\|_1)$ or $\Phi_\rho(\|\mathbf{b}\|_1)$

E-step Given the current estimates $\mathbf{b}^{(k)}$ and $\nu = \nu^{(k)}$, compute
$w^{(k)} = \nu^{(k)}\Psi'(\|\mathbf{b}^{(k)}\|_1/\sigma^{(k)})$ or $w^{(k)} = \nu^{(k)}\Phi'(\|\mathbf{b}^{(k)}\|_1/\sigma^{(k)})$

CM-step Solve the following problem:
$\mathbf{b}^{(k+1)} = \text{argmin}_{\mathbf{b}} \ \frac{1}{2}\|\mathbf{y}-\mathbf{X}\mathbf{b}\|_2^2 + w^{(k+1)}\|\mathbf{b}\|_1$,
$\sigma^{(k+1)} = \frac{1}{\alpha_\sigma+n+2p+2}\Big\{\beta_\sigma + \|\mathbf{y}-\mathbf{X}\mathbf{b}^{(k+1)}\|_2^2 + 2w^{(k+1)}\|\mathbf{b}^{(k+1)}\|_1\Big\}$.

CME-step Compute
$\nu^{(k+1)} = \frac{\alpha_t-1}{\beta_t+\Psi(\|\mathbf{b}^{(k)}\|_1/\sigma^{(k)})}$ or $\nu^{(k+1)} = \frac{\alpha_t-1}{\beta_t+\Psi(\|\mathbf{b}^{(k)}\|_1/\sigma^{(k)})}$.

γ is a global shrinkage parameter, so we call it the global tuning parameter. In the experiments, γ and β_t are selected via cross validation. As hyperparameters α_σ, β_σ, and α_t, we simply set $\alpha_\sigma = \beta_\sigma = 0$, $\alpha_t = 10$.

Our analysis is based on a set of simulated data, which are generated according to Mazumder et al. (2011). In particular, we consider the following three data models — "small," "medium" and "large."

Data S: $n = 35$, $p = 30$, $\mathbf{b}^S = (0.03, 0.07, 0.1, 0.9, 0.93, 0.97, \mathbf{0})^T$, and $\mathbf{\Sigma}^S$ is a $p \times p$ matrix with 1 on the diagonal and 0.4 on the off-diagonal.

Table 3: Results of the three algorithms with $\ell_{1/2}$, LOG, EXP, LFR and CEL on the simulated data sets. Here a standardized prediction error (SPE) is used to evaluate the model prediction ability, and the minimal achievable value for SPE is 1. And "✓" denotes the proportion of correctly predicted zero entries in $\mathbf{b}$, that is, $\frac{\#\{i|b_i=0 \text{ and } \hat{b}_i=0\}}{\#\{i|b_i=0\}}$; if all the nonzero entries are correctly predicted, this score should be 100%.

	SPE(±STD)	✓(%)	SPE(±STD)	✓(%)	SPE(±STD)	✓(%)
	Data S		Data M		Data L	
Alg 1+LOG	1.0914(±0.1703)	98.24	1.1526(±0.1025)	97.42	1.4637(±0.1735)	90.04
Alg 2+LOG	1.1508(±0.1576)	85.25	1.3035(±0.1821)	87.35	1.5084(±0.1676)	88.67
Alg 3+LOG	1.1268(±0.1754)	86.33	1.5524(±0.1437)	91.21	1.5273(±0.1567)	85.25
Alg 1+EXP	1.1106(±0.1287)	98.67	1.1587(±0.1527)	97.98	1.4608(±0.1557)	87.55
Alg 2+EXP	1.1654(±0.1845)	87.36	1.3134(±0.1152)	88.45	1.5586(±0.1802)	85.34
Alg 3+EXP	1.1552(±0.1495)	80.33	1.5047(±0.1376)	93.67	1.5145(±0.1594)	84.56
Alg 1+LFR	1.0985(±0.1824)	98.67	1.1603(±0.1158)	98.34	1.4536(±0.1697)	89.23
Alg 2+LFR	1.1326(±0.1276)	86.35	1.3089(±0.1367)	87.28	1.5183(±0.1507)	85.67
Alg 3+LFR	1.1723(±0.1534)	84.28	1.3972(±0.2356)	88.33	1.5962(±0.1467)	86.53
Alg 1+CEL	1.1238(±0.1145)	96.12	1.1642(±0.1236)	98.26	1.4633(±0.1346)	89.58
Alg 2+CEL	1.1784(±0.1093)	84.67	1.4059(±0.1736)	89.67	1.5903(±0.1785)	85.23
Alg 3+CEL	1.1325(±0.1282)	85.23	1.3762(±0.1475)	90.32	1.5751(±0.1538)	82.65
Alg 1+$\ell_{1/2}$	1.2436(±0.1458)	89.55	1.2937(±0.2033)	94.83	1.5032(±0.1633)	85.86
Alg 2+$\ell_{1/2}$	1.2591(±0.1961)	79.88	1.5902(±0.2207)	83.50	1.6859(±0.1824)	83.58
Alg 3+$\ell_{1/2}$	1.2395(±0.2045)	75.34	1.5630(±0.1642)	80.83	1.6732(±0.1711)	80.67
Lasso	1.3454(±0.3098)	60.17	1.6708(±0.2149)	66.08	1.6839(±0.1825)	71.33

Data M: $n = 100$, $p = 200$, $\mathbf{b}^M$ has 10 non-zeros such that $b^M_{20i+1} = 1$ and $i = 0, 1, \cdots, 9$, and $\mathbf{\Sigma}^M = \{0.7^{|i-j|}\}_{1\le i,j\le p}$.

Data L: $n = 500$, $p = 1000$, $\mathbf{b}^L = (\mathbf{b}^M, \cdots, \mathbf{b}^M)$, and $\mathbf{\Sigma}^L = \text{diag}(\mathbf{\Sigma}^M, \cdots, \mathbf{\Sigma}^M)$ (five blocks).

For each data model, we generate $n\times p$ data matrices $\mathbf{X}$ such that each row of $\mathbf{X}$ is generated from a multivariate Gaussian distribution with mean $\mathbf{0}_p$ and covariance matrix $\mathbf{\Sigma}^S$, $\mathbf{\Sigma}^M$, or $\mathbf{\Sigma}^L$.

We assume a linear model $\mathbf{y} = \mathbf{Xb} + \boldsymbol{\epsilon}$ with multivariate Gaussian predictors $\mathbf{X}$ and Gaussian errors. We choose σ such that the Signal-to-Noise Ratio (SNR) is a specified value. Following the setting in Mazumder et al. (2011), we use SNR = 3.0 in all the experiments. We employ a standardized prediction error (SPE) to evaluate the model prediction ability. The minimal achievable value for SPE is 1. Variable selection accuracy is measured by the correctly predicted zeros and incorrectly predicted zeros in $\hat{\mathbf{b}}$. The

SNR and SPE are defined as

$$\text{SNR} = \frac{\sqrt{\mathbf{b}^T \boldsymbol{\Sigma} \mathbf{b}}}{\sigma} \quad \text{and} \quad \text{SPE} = \frac{\mathbb{E}(\mathbf{y} - \mathbf{x}\hat{\mathbf{b}})^2}{\sigma^2}.$$

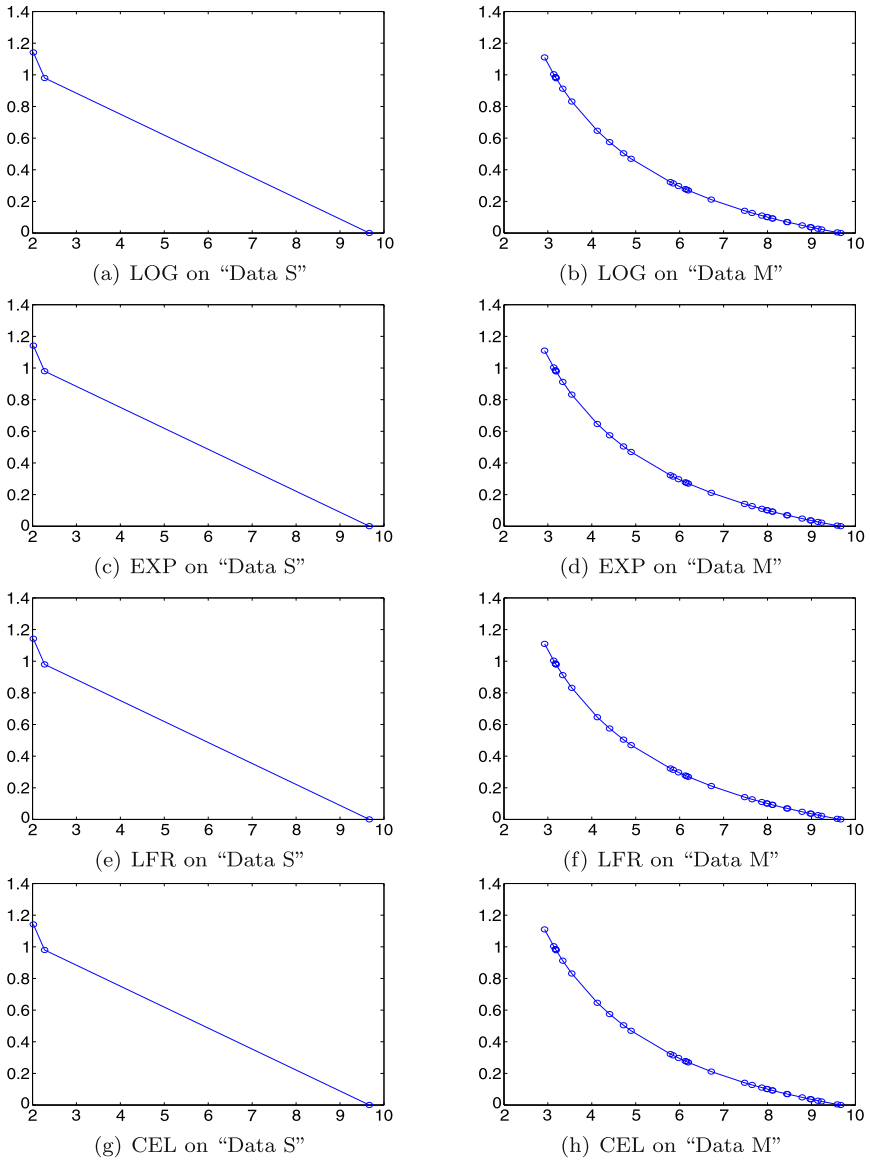
For each data model, we generate training data of size n, very large validation data and test data, each of size 10000. For each algorithm, the optimal global tuning parameters are chosen by cross validation based on minimizing the average prediction errors. With the model $\hat{\mathbf{b}}$ computed on the training data, we compute SPE on the test data. This procedure is repeated 100 times, and we report the average and standard deviation of SPE and the average of zero-nonzero error. We use "✓" to denote the proportion of correctly predicted zero entries in $\mathbf{b}$, that is, $\frac{\#\{i|b_i=0 \text{ and } \hat{b}_i=0\}}{\#\{i|b_i=0\}}$; if all the nonzero entries are correctly predicted, this score should be 100%.

We report the results in Table 3. It is seen that our setting in Figure 2-(a) is better than the other two settings in Figures 2-(b) and (c) in both model prediction accuracy and variable selection ability. Especially, when the size of the dataset takes large values, the prediction performance of the second setting becomes worse. The several nonconvex penalties are competitive, but they outperform the lasso. Moreover, we see that LOG, EXP, LFR and CEL slightly outperform $\ell_{1/2}$. The $\ell_{1/2}$ penalty indeed suffers from the problem of numerical instability during the EM computations. As we know, the priors induced from LFR, CEL and EXP as well as LOG with $t \leq \xi$ are improper, but the prior induced from $\ell_{1/2}$ is proper. The experimental results show that these improper priors work well, even better than the proper case.

Recall that in our approach each regression variable b_j corresponds to a distinct local tuning parameter t_j. Thus, it is interesting to empirically investigate the inherent relationship between b_j and t_j. Let $\hat{t}_j$ be the estimate of t_j obtained from our ECME algorithm ("Alg 1"), and $(\pi_1, \ldots, \pi_p)$ be the permutation of $(1, \ldots, p)$ such that $\hat{t}_{\pi_1} \leq \cdots \leq \hat{t}_{\pi_p}$. Figure 3 depicts the change of $|\hat{b}_{\pi_j}|$ vs. $\hat{t}_{\pi_j}$ with LOG, EXP, LFR and CEL on "Data S" and "Data M." We see that $|\hat{b}_{\pi_j}|$ is decreasing w.r.t. $\hat{t}_{\pi_j}$. Moreover, $|\hat{b}_{\pi_j}|$ becomes 0 when $\hat{t}_{\pi_j}$ takes some large value. A similar phenomenon is also observed for "Data L." This thus shows that the subordinator is a powerful Bayesian approach for variable selection.

6 Conclusion

In this paper we have introduced subordinators into the definition of nonconvex penalty functions. This leads us to a Bayesian approach for constructing sparsity-inducing pseudo-priors. In particular, we have illustrated the use of two compound Poisson subordinators: the compound Poisson Gamma subordinator and the negative binomial subordinator. In addition, we have established the relationship between the two families of compound Poisson subordinators. That is, we have proved that the two families of compound Poisson subordinators share the same limiting behaviors. Moreover, their densities at each time have the same mean and variance.



(a) LOG on "Data S" (b) LOG on "Data M"

(c) EXP on "Data S" (d) EXP on "Data M"

(e) LFR on "Data S" (f) LFR on "Data M"

(g) CEL on "Data S" (h) CEL on "Data M"

Figure 3: The change of $|\hat{b}_{\pi_j}|$ vs. $\hat{t}_{\pi_j}$ on "Data S" and "Data M" where $(\pi_1, \ldots, \pi_p)$ is the permutation of $(1, \ldots, p)$ such that $\hat{t}_{\pi_1} \leq \cdots \leq \hat{t}_{\pi_p}$.

We have developed the ECME algorithms for solving sparse learning problems based on the nonconvex LOG, EXP, LFR and CEL penalties. We have conducted the experimental comparison with the state-of-the-art approach. The results have shown that our nonconvex penalization approach is potentially useful in high-dimensional Bayesian modeling. Our approach can be cast into a point estimation framework. It is also interesting to fit a fully Bayesian framework based on the MCMC estimation. We would like to address this issue in future work.

Appendix 1: The Lévy Measure of the CEL Function

Consider that

$$\begin{aligned}\log\Big[\frac{1+\rho}{\rho}-\frac{1}{\rho}\exp(-\frac{\rho}{1+\rho}\gamma s)\Big] &= \log\Big[1-\frac{1}{1+\rho}\exp(-\frac{\rho}{1+\rho}\gamma s)\Big]-\log\Big[1-\frac{1}{1+\rho}\Big]\\ &= \sum_{k=1}^{\infty}\frac{1}{k(1+\rho)^k}\Big[1-\exp\Big(-\frac{\rho}{1+\rho}k\gamma s\Big)\Big]\\ &= \sum_{k=1}^{\infty}\frac{1}{k(1+\rho)^k}\int_0^{\infty}(1-\exp(-us))\delta_{\frac{\rho k\gamma}{1+\rho}}(u)du.\end{aligned}$$

We thus have that $\nu(du)=\frac{1+\rho}{\xi}\sum_{k=1}^{\infty}\frac{1}{k(1+\rho)^k}\delta_{\frac{\rho k\gamma}{1+\rho}}(u)du$.

Appendix 2: The Proof of Proposition 2

We here give an alternative proof of Proposition 2-(d), which is immediately obtained from the following lemma.

Lemma 2. *Let X take discrete value on $\mathbb{N}\cup\{0\}$ and follow negative binomial distribution $\mathrm{Nb}(r,p)$. If r converges to a positive constant as $p\to 0$, pX converges in distribution to a Gamma random variable with shape r and scale 1.*

Proof. Since

$$F_{pX}(x)=\Pr(pX\le x)=\sum_{\substack{kp\le x\\ k=0}}^{\infty}\frac{\Gamma(k+r)}{\Gamma(r)\Gamma(k+1)}p^r(1-p)^k,$$

we have that

$$\lim_{p\to 0+}F_{pX}(x)=\lim_{p\to 0+}\sum_{\substack{kp\le x\\ k=0}}^{\infty}\frac{\Gamma(k+r)}{\Gamma(r)\Gamma(k+1)}p^r(1-p)^k$$

$$= \lim_{p\to 0+} \sum_{\substack{kp\le x \\ k=1}}^{\infty} \frac{\Gamma(\frac{kp}{p}+r)}{\Gamma(r)\Gamma(\frac{kp}{p}+1)} p p^{r-1}(1-p)^k$$

$$= \frac{1}{\Gamma(r)} \int_0^x \lim_{p\to 0} \frac{\Gamma(\frac{u}{p}+r)}{\Gamma(\frac{u}{p}+1)} p^{r-1}(1-p)^{u/p} du.$$

Notice that $\lim_{p\to 0}(1-p)^{u/p} = \exp(-u)$ and

$$\lim_{p\to 0} \frac{\Gamma(\frac{u}{p}+r)}{\Gamma(\frac{u}{p}+1)} p^{r-1} = \lim_{p\to 0} \frac{\left(\frac{u}{p}+r\right)^{\frac{u}{p}+r-\frac{1}{2}} \exp(-\frac{u}{p}-r)}{\left(\frac{u}{p}+1\right)^{\frac{u}{p}+1-\frac{1}{2}} \exp(-\frac{u}{p}-1)} p^{r-1}$$

$$= \exp(1-r) \lim_{p\to 0} \left(\frac{\frac{u}{p}+1+r-1}{\frac{u}{p}+1}\right)^{\frac{u}{p}+1} \left(\frac{\frac{u}{p}+r}{\frac{u}{p}+1}\right)^{-\frac{1}{2}} \left(\frac{u}{p}+r\right)^{r-1} p^{r-1}$$

$$= \exp(1-r)\exp(r-1)u^{r-1} = u^{r-1}.$$

This leads us to

$$\lim_{p\to 0+} F_{pX}(x) = \int_0^x \frac{u^{r-1}}{\Gamma(r)} \exp(-u) du. \qquad \square$$

Similarly, we have that

$$\lim_{\rho\to 0} \nu(du) = \frac{\rho+1}{\xi} \sum_{k=1}^{\infty} \frac{\rho}{k\rho(1+\rho)^{k\rho/\rho}} \delta_{k\rho\gamma/(1+\rho)} du$$

$$= \frac{1}{\xi} \int_0^{\infty} z^{-1} \exp(-z) \delta_{z\gamma}(u) dz$$

$$= \frac{1}{\xi} u^{-1} \exp(-u/\gamma).$$

Appendix 3: The Proof of Theorem 2

Proof. Consider a mixture of $\mathrm{Ga}(\eta|k\nu,\beta)$ with $\mathrm{Po}(k|\lambda)$ mixing. That is,

$$p(\eta) = \sum_{k=0}^{\infty} \mathrm{Ga}(\eta|k\nu,\beta)\mathrm{Po}(k|\lambda)$$

$$= \sum_{k=0}^{\infty} \frac{\beta^{-k\nu}}{\Gamma(k\nu)} \eta^{k\nu-1} \exp(-\frac{\eta}{\beta}) \frac{\lambda^k}{k!} \exp(-\lambda)$$

$$= \lim_{k\to 0} \frac{\eta^{k\nu-1}\lambda^k}{\beta^{k\nu}\Gamma(k\nu)k!} \exp(-\frac{\eta}{\beta}) \exp(-\lambda) + \sum_{k=1}^{\infty} \frac{\lambda^k(\eta/\beta)^{k\nu}}{\eta\Gamma(k\nu)k!} \exp(-(\frac{\eta}{\beta}+\lambda))$$

$$= \exp(-\lambda)\Big\{\delta_0(\eta) + \exp(-\frac{\eta}{\beta})\sum_{k=1}^{\infty}\frac{\lambda^k(\eta/\beta)^{k\nu}}{\eta\Gamma(k\nu)k!}\Big\}.$$

Letting $\lambda = \frac{\rho t}{\xi(\rho-1)}$, $\nu = \rho - 1$ and $\beta = \frac{\gamma}{\rho}$, we have that

$$p(\eta) = \exp\left(-\frac{\rho t}{\xi(\rho-1)}\right)\left\{\delta_0(\eta) + \exp\left(-\frac{\rho\eta}{\gamma}\right)\eta^{-1}\sum_{k=1}^{\infty}\frac{(\rho t/\xi)^k(\rho\eta/\gamma)^{k(\rho-1)}}{k!(\rho-1)^k\Gamma(k(\rho-1))}\right\}.$$

We now consider a mixture of $\mathrm{Po}(k|\phi\lambda)$ with $\mathrm{Ga}(\lambda|\psi, 1/\beta)$ mixing. That is,

$$\begin{aligned}\Pr(T(t) = k\alpha) &= \int_0^{\infty}\mathrm{Po}(k|\lambda\phi)\mathrm{Ga}(\lambda|\psi, 1/\beta)d\lambda \\ &= \int_0^{\infty}\frac{(\lambda\phi)^k}{k!}\exp(-\lambda\phi)\frac{\beta^{\psi}}{\Gamma(\psi)}\lambda^{\psi-1}\exp(-\beta\lambda)d\lambda \\ &= \frac{\beta^{\psi}}{\Gamma(\psi)}\frac{\Gamma(\psi+k)}{k!}\frac{\phi^k}{(\phi+\beta)^{k+\psi}},\end{aligned}$$

which is $\mathrm{Nb}(T(t)|\psi, \beta/(\beta+\phi))$. Let $\psi = (\rho+1)t/\xi$, $\phi = 1$, $\beta = \rho$ and $q = \frac{\beta}{\phi+\beta}$. Thus,

$$\Pr(T(t) = k\alpha) = \frac{\Gamma(k+(\rho+1)t/\xi)}{k!\Gamma((\rho+1)t/\xi)}q^{(\rho+1)t/\xi}(1-q)^k. \qquad \square$$

Appendix 4: The Proof of Theorem 3

Proof. Since $\lim_{\gamma\to 0}\frac{\rho+1}{\rho\gamma}\left[1 - (1+\frac{\gamma}{\rho+1})^{-\rho}\right] = 1$, we only need to consider the case that $\xi = \gamma$. Recall that $\mathrm{PG}(t/\xi, \gamma, \rho)$, whose mean and variance are

$$\mathbb{E}(T(t)) = \frac{\gamma t}{\xi} = t \quad \text{and} \quad \mathbb{V}ar(T(t)) = \frac{\gamma^2 t}{\xi} = \gamma t$$

whenever $\xi = \gamma$. By Chebyshev's inequality, we have that

$$\Pr\{|T(t) - t| \geq \epsilon\} \leq \frac{\gamma t}{\epsilon^2}.$$

Hence, we have that

$$\lim_{\gamma\to 0}\Pr\{|T(t) - t| \geq \epsilon\} = 0.$$

Similarly, we have Part (b). $\square$

Appendix 5: The Proof of Proposition in Expression (11)

Proof. We first note that

$$2\exp(\gamma s) = 2 + 2\gamma s + (\gamma s)^2 + \frac{2}{3}(\gamma s)^3 + \cdots,$$

which implies that $2\exp(\gamma s) - 1 - (\gamma s + 1)^2 > 0$ for $s > 0$. Subsequently, we have that $\frac{d}{ds}\big[\log(2-\exp(-\gamma s)) - \frac{\gamma s}{1+\gamma s}\big] \leq 0$. As a result, $\log(2-\exp(-\gamma s)) - \frac{\gamma s}{1+\gamma s} < 0$ for $s > 0$. As for $\frac{\gamma s}{\gamma s+1} \leq 1-\exp(-\gamma s)$, it is directly obtained from that

$$\frac{\gamma s}{\gamma s+1} = 1 - \frac{1}{1+\gamma s} = 1 - \exp(-\log(1+\gamma s)) \leq 1 - \exp(-\gamma s).$$

Since $\frac{d}{ds}\big[1-\exp(-\gamma s) - \log\big(\gamma s+1\big)\big] = \frac{\gamma}{\exp(\gamma s)} - \frac{\gamma}{1+\gamma s} < 0$ for $s > 0$, we have that $1-\exp(-\gamma s) - \log\big(\gamma s+1\big) < 0$ for $s > 0$. □

Appendix 6: The Proof of Theorem 4

Proof. First consider that

$$p(\mathbf{b}|\sigma, \mathbf{t}, \mathbf{y}) \propto \frac{1}{(2\pi\sigma)^{\frac{n}{2}}} \exp\big[-\frac{1}{2\sigma}\|\mathbf{y}-\mathbf{X}\mathbf{b}\|_2^2\big] \prod_{j=1}^{p} \sigma^{-1} \exp\Big(-t_j \Psi\Big(\frac{|b_j|}{\sigma}\Big)\Big).$$

To prove that $p(\mathbf{b}|\sigma, \mathbf{t}, \mathbf{y})$ is proper, it suffices to obtain that

$$\frac{1}{(2\pi\sigma)^{\frac{n}{2}}} \int \exp\big[-\frac{1}{2\sigma}\|\mathbf{y}-\mathbf{X}\mathbf{b}\|_2^2\big] \prod_{j=1}^{p} \sigma^{-1} \exp\Big(-t_j \Psi\Big(\frac{|b_j|}{\sigma}\Big)\Big) d\mathbf{b} < \infty.$$

It is directly computed that

$$\begin{aligned} &\exp\big[-\frac{1}{2\sigma}\|\mathbf{y}-\mathbf{X}\mathbf{b}\|_2^2\big] \\ &= \exp\big[-\frac{1}{2\sigma}(\mathbf{b}-\mathbf{z})^T\mathbf{X}^T\mathbf{X}(\mathbf{b}-\mathbf{z})\big] \times \exp\big[-\frac{1}{2\sigma}\mathbf{y}^T(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^+\mathbf{X}^T)\mathbf{y}\big], \end{aligned} \tag{13}$$

where $\mathbf{z} = (\mathbf{X}^T\mathbf{X})^+\mathbf{X}^T\mathbf{y}$ and $(\mathbf{X}^T\mathbf{X})^+$ is the Moore-Penrose pseudo inverse of matrix $\mathbf{X}^T\mathbf{X}$ (Magnus and Neudecker, 1999). Here we use the well-established properties that $\mathbf{X}(\mathbf{X}^T\mathbf{X})^+(\mathbf{X}^T\mathbf{X}) = \mathbf{X}$ and $(\mathbf{X}^T\mathbf{X})^+(\mathbf{X}^T\mathbf{X})(\mathbf{X}^T\mathbf{X})^+ = (\mathbf{X}^T\mathbf{X})^+$. Notice that if $\mathbf{X}^T\mathbf{X}$ is nonsingular, then $(\mathbf{X}^T\mathbf{X})^+ = (\mathbf{X}^T\mathbf{X})^{-1}$. In this case, we consider a conventional multivariate normal distribution $N(\mathbf{b}|\mathbf{z}, \sigma(\mathbf{X}^T\mathbf{X})^{-1})$. Otherwise, we consider a singular multivariate normal distribution $N(\mathbf{b}|\mathbf{z}, \sigma(\mathbf{X}^T\mathbf{X})^+)$ (Mardia et al., 1979), the density of which is given by

$$\frac{\prod_{j=1}^{q}\sqrt{\lambda_j(\mathbf{X}^T\mathbf{X})}}{(2\pi\sigma)^{q/2}} \exp\big[-\frac{1}{2\sigma}(\mathbf{b}-\mathbf{z})^T\mathbf{X}^T\mathbf{X}(\mathbf{b}-\mathbf{z})\big].$$

Here $q = \text{rank}(\mathbf{X})$, and $\lambda_j(\mathbf{X}^T\mathbf{X})$, $j = 1, \ldots, q$, are the positive eigenvalues of $\mathbf{X}^T\mathbf{X}$. In any case, we always write $N(\mathbf{b}|\mathbf{z}, \sigma(\mathbf{X}^T\mathbf{X})^+)$. Thus, $\int \exp\big[-\frac{1}{2\sigma}\|\mathbf{y}-\mathbf{X}\mathbf{b}\|_2^2\big] d\mathbf{b} < \infty$. It then follows the propriety of $p(\mathbf{b}|\sigma, \mathbf{t}, \mathbf{y})$ because

$$\exp\big[-\frac{1}{2\sigma}\|\mathbf{y}-\mathbf{X}\mathbf{b}\|_2^2\big] \prod_{j=1}^{p} \exp\Big(-t_j \Psi\Big(\frac{|b_j|}{\sigma}\Big)\Big) \leq \exp\big[-\frac{1}{2\sigma}\|\mathbf{y}-\mathbf{X}\mathbf{b}\|_2^2\big].$$

We now consider that

$$p(\mathbf{b},\sigma|\mathbf{t},\mathbf{y}) \propto \sigma^{-(\frac{n+\alpha_\sigma+2p}{2}+1)} \exp\Big[-\frac{\|\mathbf{y}-\mathbf{X}\mathbf{b}\|_2^2+\beta_\sigma}{2\sigma}\Big] \prod_{j=1}^{p} \exp\Big(-t_j\Psi\Big(\frac{|b_j|}{\sigma}\Big)\Big).$$

Let $\nu = \mathbf{y}^T[\mathbf{I}_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^+\mathbf{X}^T]\mathbf{y}$. Since the matrix $\mathbf{I}_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^+\mathbf{X}^T$ is positive semidefinite, we obtain $\nu \geq 0$. Based on expression (13), we can write

$$\sigma^{-(\frac{n+\alpha_\sigma+2p}{2}+1)} \exp\Big[-\frac{\|\mathbf{y}-\mathbf{X}\mathbf{b}\|_2^2+\beta_\sigma}{2\sigma}\Big] \propto N(\mathbf{b}|\mathbf{z},\sigma(\mathbf{X}^T\mathbf{X})^+)\mathrm{IG}(\sigma|\frac{\alpha_\sigma+n+2p-q}{2},\nu+\beta_\sigma).$$

Subsequently, we have that

$$\int \sigma^{-(\frac{n+\alpha_\sigma+2p}{2}+1)} \exp\Big[-\frac{\|\mathbf{y}-\mathbf{X}\mathbf{b}\|_2^2+\beta_\sigma}{2\sigma}\Big] d\mathbf{b}d\sigma < \infty,$$

and hence,

$$\int \sigma^{-(\frac{n+\alpha_\sigma+2p}{2}+1)} \exp\Big[-\frac{\|\mathbf{y}-\mathbf{X}\mathbf{b}\|_2^2+\beta_\sigma}{2\sigma}\Big] \prod_{j=1}^{p} \exp\Big(-t_j\Psi\Big(\frac{|b_j|}{\sigma}\Big)\Big) d\mathbf{b}d\sigma < \infty.$$

Therefore $p(\mathbf{b},\sigma|\mathbf{t},\mathbf{y})$ is proper.

Thirdly, we take

$$\begin{aligned} p(\mathbf{b},\sigma,\mathbf{t}|\mathbf{y}) &\propto \frac{\exp\Big[-\frac{\|\mathbf{y}-\mathbf{X}\mathbf{b}\|_2^2+\beta_\sigma}{2\sigma}\Big]}{\sigma^{\frac{n+\alpha_\sigma+2p}{2}+1}} \prod_{j=1}^{p} \Big\{\exp\Big(-t_j\Psi\Big(\frac{|b_j|}{\sigma}\Big)\Big)\frac{t_j^{\alpha_t-1}\exp(-\beta_t t_j)}{\Gamma(\alpha_t)}\Big\} \\ &\triangleq F(\mathbf{b},\sigma,\mathbf{t}). \end{aligned}$$

In this case, we compute

$$\int F(\mathbf{b},\sigma,\mathbf{t})d\mathbf{b}d\sigma d\mathbf{t} = \int \frac{\exp\Big[-\frac{\|\mathbf{y}-\mathbf{X}\mathbf{b}\|_2^2+\beta_\sigma}{2\sigma}\Big]}{\sigma^{\frac{n+\alpha_\sigma+2p}{2}+1}} \prod_{j=1}^{p} \frac{1}{\Big(\beta_t+\Psi\Big(\frac{|b_j|}{\sigma}\Big)\Big)^{\alpha_t}} d\mathbf{b}d\sigma.$$

Similar to the previous proof, we also have that

$$\int F(\mathbf{b},\sigma,\mathbf{t})d\mathbf{b}d\sigma d\mathbf{t} < \infty$$

because $\Big(\beta_t+\Psi\Big(\frac{|b_j|}{\sigma}\Big)\Big)^{-\alpha_t} \leq {\beta_t}^{-\alpha_t}$. As a result, $p(\mathbf{b},\sigma,\mathbf{t}|\mathbf{y})$ is proper.

Finally, consider the setting that $p(\sigma) \propto \frac{1}{\sigma}$. That is, $\alpha_\sigma = 0$ and $\beta_\sigma = 0$. In this case, if $\mathbf{y} \notin \mathrm{range}(\mathbf{X})$, we obtain $\nu > 0$ and $q < n$. As a result, we use the inverse Gamma distribution $\mathrm{IG}(\sigma|\frac{n+2p-q}{2},\nu)$. Thus, the results still hold. □

References

Aalen, O. O. (1992). "Modelling heterogeneity in survival analysis by the compound Poisson distribution." *The Annals of Applied Probability*, 2(4): 951–972. 248, 253, 254

Applebaum, D. (2004). *Lévy Processes and Stochastic Calculus*. Cambridge, UK: Cambridge University Press. 249

Armagan, A., Dunson, D., and Lee, J. (2013). "Generalized double Pareto shrinkage." *Statistica Sinica*, 23: 119–143. 247, 261

Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2012). "Bayesian Shrinkage." Technical report. arXiv:1212.6088. 248

Brix, A. (1999). "Generalized Gamma measures and shot-noise Cox processes." *Advances in Applied Probability*, 31(4): 929–953. 248, 253, 254

Broderick, T., Jordan, M. I., and Pitman, J. (2012). "Beta Processes, Stick-Breaking and Power Laws." *Bayesian Analysis*, 7(2): 439–476. 248

Caron, F. and Doucet, A. (2008). "Sparse Bayesian nonparametric regression." In *Proceedings of the 25th International Conference on Machine Learning (ICML)*, 88. 248, 262

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). "Handling sparsity via the horsehoe." In *The Twelfth International Conference on Artificial Intelligence and Statistics*, 73–80. 248

— (2010). "The horseshoe estimator for sparse signals." *Biometrika*, 97: 465–480. 248

Cevher, V. (2009). "Learning with compressible priors." In *Advances in Neural Information Processing Systems (NIPS) 22*, 261–269. 247, 261

Feller, W. (1971). *An Introduction to Probability Theory and Its Applications*, volume II. New York: John Wiley and Sons, second edition. 254

Figueiredo, M. A. T. (2003). "Adaptive Sparseness for Supervised Learning." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9): 1150–1159. 247

Garrigues, P. J. and Olshausen, B. A. (2010). "Group Sparse Coding with a Laplacian Scale Mixture Prior." In *Advances in Neural Information Processing Systems (NIPS) 22*. 247, 261

Ghahramani, Z., Griffiths, T., and Sollich, P. (2006). "Bayesian Nonparametric latent feature models." In *World Meeting on Bayesian Statistics*. 248

Griffin, J. E. and Brown, P. J. (2010). "Inference with normal-gamma prior distributions in regression problems." *Bayesian Analysis*, 5(1): 171–183. 247

— (2011). "Bayesian Hyper-lassos with Non-convex Penalization." *Australian & New Zealand Journal of Statistics*, 53(4): 423–442. 248

Hans, C. (2009). "Bayesian lasso regression." *Biometrika*, 96: 835–845. 247

Kyung, M., Gill, J., Ghosh, M., and Casella, G. (2010). "Penalized regression, standard errors, and Bayesian lassos." *Bayesian Analysis*, 5(2): 369–382. 247

Lee, A., Caron, F., Doucet, A., and Holmes, C. (2010). "A Hierarchical Bayesian Framework for Constructing Sparsity-inducing Priors." Technical report, University of Oxford, UK. 247, 261

Li, Q. and Lin, N. (2010). "The Bayesian Elastic Net." *Bayesian Analysis*, 5(1): 151–170. 247

Liu, C. and Rubin, D. B. (1994). "The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence." *Biometrika*, 84(4): 633–648. 248, 260, 261

Magnus, J. R. and Neudecker, H. (1999). *Matrix Calculus with Applications in Statistics and Econometrics*. New York: John Wiley & Sons, revised edition. 270

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate Analysis*. New York: Academic Press. 270

Mazumder, R., Friedman, J., and Hastie, T. (2011). "SparseNet: Coordinate Descent with Nonconvex Penalties." *Journal of the American Statistical Association*, 106(495): 1125–1138. 263, 264

Paisley, J. and Carin, L. (2009). "Nonparametric factor analysis with beta process priors." In *The 26th International Conference on Machine Learning (ICML)*. 248

Park, T. and Casella, G. (2008). "The Bayesian Lasso." *Journal of the American Statistical Association*, 103(482): 681–686. 247

Polson, N. G. and Scott, J. G. (2010). "Shrink Globally, Act Locally: Sparse Bayesian Regularization and Prediction." In Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., and West, M. (eds.), *Bayesian Statistics 9*. Oxford University Press. 260

— (2011). "Data Augmentation for Support Vector Machines." *Bayesian Analysis*, 6(1): 1–24. 251, 261

— (2012). "Local shrinkage rules, Lévy processes and regularized regression." *Journal of the Royal Statistical Society, Series B*, 74(2): 287–311. 248, 261, 262

Sato, S.-I. P. (1999). *Lévy Processes and Infinitely Divisible Distributions*. Cambridge, UK: Cambridge University Press. 248, 249, 252, 255, 256

Teh, Y. W. and Görür, D. (2009). "Indian buffet processes with power-law behavior." In *Advances in Neural Information Processing Systems (NIPS)*. 248

Thibaux, R. and Jordan, M. I. (2007). "Hierachical Beta Processes and the Indian Buffet Processes." In *The International Conference on AI and Statistics*. 248

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society, Series B*, 58: 267–288. 247

Titsias, M. K. (2007). "The Infinite Gamma-Poisson Feature Models." In *Advances in Neural Information Processing Systems (NIPS) 20*. 248

Yuan, L. and Kalbfleisch, J. D. (2000). "On the Bessel Distribution and Related Problems." *Annals of the Institute of Statistical Mathematics*, 52(3): 438–447. 254

Zhang, Z. and Tu, B. (2012). "Nonconvex Penalization Using Laplace Exponents and Concave Conjugates." In *Advances in Neural Information Processing Systems (NIPS) 26*. 248, 251

Zhang, Z., Wang, S., Liu, D., and Jordan, M. I. (2012). "EP-GIG priors and applications in Bayesian sparse learning." *Journal of Machine Learning Research*, 13: 2031–2061. 248, 260

Zou, H. and Li, R. (2008). "One-step sparse estimates in nonconcave penalized likelihood models." *The Annals of Statistics*, 36(4): 1509–1533. 247, 260, 261

Acknowledgments

The authors would like to thank the Editors and two anonymous referees for their constructive comments and suggestions on the original version of this paper. The authors would especially like to thank the Associate Editor for giving extremely detailed comments on earlier drafts. This work has been supported in part by the Natural Science Foundation of China (No. 61070239).