Analytic solution and stationary phase approximation for the Bayesian lasso and elastic net

Tom Michoel
tom.michoel@uib.no | lab.michoel.info

Abstract

- The lasso and elastic net linear regression models impose a **double-exponential prior distribution** on the model parameters to achieve regression shrinkage and variable selection.
- There has been limited success in deriving estimates for the posterior distribution of regression coefficients in these models, due to a need to evaluate **analytically intractable partition function integrals**, not amenable to a convential Laplace approximation.
- We used the **Fourier transform** to express these integrals as complex-valued oscillatory integrals over "regression frequencies", which are amenable to a Gaussian approximation.
- An **analytic expansion** and **stationary phase approximation** in Fourier space is derived for the partition functions of the Bayesian lasso and elastic net.
- This approximation results in **highly accurate numerical estimates** at much **reduced computational cost** compared to Gibbs sampling.

Bayesian elastic net model

Response data $y \in \mathbb{R}^n$, predictor data $A \in \mathbb{R}^{n\times p}$.
Regression coefficients $x \in \mathbb{R}^p$.
Hierarchical model with double-exponential (ℓ_1) prior:

$$p(y \mid A, x) = \mathcal{N}(Ax, \sigma^2\mathbb{1}) \propto e^{-\frac{1}{2\sigma^2}\|y-Ax\|^2}$$

$$p(x) \propto e^{-\frac{n}{\sigma^2}\left(\lambda\|x\|^2+2\mu\|x\|_1\right)}$$

Posterior distribution of x:

$$p(x \mid y, A) \propto p(y \mid x, A)\; p(x) \propto e^{-\frac{n}{\sigma^2}\mathcal{L}(x|y,A)},$$

where

$$\mathcal{L}(x \mid y, A) = x^T\left(\frac{A^TA}{2n} + \lambda\mathbb{1}\right)x - 2\left(\frac{A^Ty}{2n}\right)^Tx + 2\mu\|x\|_1 + \frac{1}{2n}\|y\|^2$$

is minus the posterior log-likelihood function.

Problem formulation

The Bayesian elastic net belongs to a more general class of models with cost functions

$$H(x \mid C, w, \mu) = x^TCx - 2w^Tx + 2\mu\|x\|_1,$$

where $C \in \mathbb{R}^{p\times p}$ is positive definite, $w \in \mathbb{R}^p$ and $\mu > 0$.

These cost/energy functions define Gibbs distributions

$$p(x \mid C, w, \mu, \tau) = \frac{e^{-\tau H(x|C,w,\mu)}}{Z(C, w, \mu, \tau)}.$$

We seek to compute the **partition function**

$$Z(C, w, \mu, \tau) = \int_{\mathbb{R}^p} e^{-\tau H(x|C,w,\mu)}dx$$

when the inverse temperature τ is large, but **cannot apply a Laplace approximation** because H is not twice differentiable.

Key idea

Maximum-likelihood	$\Rightarrow$	Bayesian inference
Legendre transform	$\Rightarrow$	Fourier transform
Fenchel's duality theorem	$\Rightarrow$	Parseval's identity

Hence, write $f(x) = \frac{1}{2}x^TCx - w^Tx$, $g(x) = \mu\|x\|_1 = \mu\sum_{j=1}^p |x_j|$ and use Parseval's identity to write the partition function as a p-dimensional **complex contour integral** in Fourier space:

$$Z = \int_{\mathbb{R}^p} e^{-2\tau f(x)}e^{-2\tau g(x)}dx = \int_{\mathbb{R}^p} \overline{\mathcal{F}(e^{-\tau f})}(k)\mathcal{F}(e^{-\tau g})(k)dk$$
$$= \frac{(-i\mu)^p}{(\pi\tau)^{\frac{p}{2}}\sqrt{\det(C)}}\int_{i\mathbb{R}^p} e^{\tau(z-w)^TC^{-1}(z-w)}\prod_{j=1}^p\frac{1}{\mu^2-z_j^2}dz \quad (1)$$

The exponential factor in eq. (1) has a saddle point at $z = w$, but poles at $z_j = \pm\mu$ prevent application of the standard stationary phase approximation (Fig. 1). We therefore write Z as

$$Z = \frac{(-i\mu)^p}{(\pi\tau)^{\frac{p}{2}}\sqrt{\det(C)}}\int_{i\mathbb{R}^p} e^{\tau H^*_\tau(z)}dz \quad (2)$$

and expand around the saddle point of

$$H^*_\tau(z) = (z-w)^TC^{-1}(z-w) - \frac{1}{\tau}\sum_{j=1}^p \ln(\mu^2 - z_j^2) \quad (3)$$

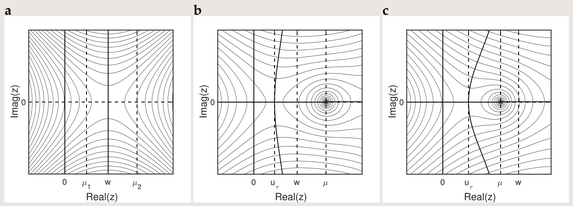

Fig. 1: Illustration in 1d: The quadratic exponent has a saddle point at $z = w$ but the integration contour cannot be deformed to pass through it if $|w| \geq \mu$ (a). By adding the logarithmic barrier function to the exponent, there is a unique saddle point $\hat{u}_\tau$ with $|\hat{u}_\tau| \leq \mu$ regardless of $|w| < \mu$ (b) or not (c).

The saddle point equations

H^*_τ has a unique saddle point $\hat{u}_\tau$ in the domain $\mathcal{D} = \{z \in \mathbb{C}^p : |\Re z_j| < \mu,\ j = 1, \ldots, p\}$. $\hat{u}_\tau$ is real, $\hat{u}_\tau \in \mathcal{D} \cap \mathbb{R}^p$, and solves the set of third order equations

$$(\mu^2 - u_j^2)[C^{-1}(w-u)]_j - \frac{u_j}{\tau} = 0,\ j \in \{1, \ldots, p\}. \quad (4)$$

As $\tau \to \infty$, $\hat{u} = \lim_{\tau\to\infty}\hat{u}_\tau$ is a solution to the set of equations

$$(u_j - \mu)(u_j + \mu)[C^{-1}(w-u)]_j = 0 \;\text{ subject to }\; |u_j| \leq \mu$$

These are the optimality conditions for the convex dual problem of maximizing the elastic net log-likelihood. The maximum-likelihood coefficients $\hat{x}$ satisfy $\hat{x} = C^{-1}(w-\hat{u}) = \lim_{\tau\to\infty}\hat{x}_\tau$ where $\hat{x}_\tau = C^{-1}(w-\hat{u}_\tau)$, and $\hat{x}_j \neq 0 \Leftrightarrow \hat{u}_j = \pm\mu$.

The stationary phase approximation

By shifting the integration contour in eq. (2) to a **steepest descent contour** passing through the saddle point (Fig. 1), a **Gaussian approximation** to the partition function is obtained:

$$Z \sim \left(\frac{\mu}{\sqrt{\tau}}\right)^p e^{\tau(w-\hat{u}_\tau)^TC^{-1}(w-\hat{u}_\tau)}\prod_{j=1}^p\frac{1}{\sqrt{\mu^2+\hat{u}_{\tau,j}^2}}\frac{1}{\sqrt{\det(C+D_\tau)}}, \quad (5)$$

where D_τ is a diagonal matrix with diagonal elements $\frac{\tau(\mu^2-\hat{u}_{\tau,j}^2)^2}{\mu^2+\hat{u}_{\tau,j}^2}$.

Although H^*_τ and the saddle point depend on τ, standard stationary phase approximation estimates still hold and allow to prove that this is a bona fide **analytic approximation** (i.e. higher-order terms vanish for large τ relative to the quadratic term).

The stationary phase approximation is used to compute:

- **Posterior expectation values**:

$$\mathbb{E}_\tau(x) \sim \hat{x}_\tau = C^{-1}(w - \hat{u}_\tau)$$

- **Marginal posterior distributions**:

$$p(x_j) = e^{-\tau(C_{jj}x_j^2 - 2w_jx_j + 2\mu|x_j|)}\frac{Z(C_{\setminus j}, w_{\setminus j} - x_jC_{\setminus j,j}, \mu)}{Z}$$

- **Posterior predictive distributions**:

$$p(y) = \left(\frac{\tau}{2\pi n}\right)^{\frac{1}{2}}e^{-\frac{\tau}{2n}y^2}\frac{Z(C + \frac{1}{2n}aa^T, w + \frac{y}{2n}a, \mu)}{Z}$$

Numerical results

Marginal posterior distributions using eq. (5) are indistinguishable from those obtained by Gibbs sampling, but not if we approximate $\hat{u}_\tau$ by its maximum-likelihood limit $\hat{u}$ (Fig. 2).

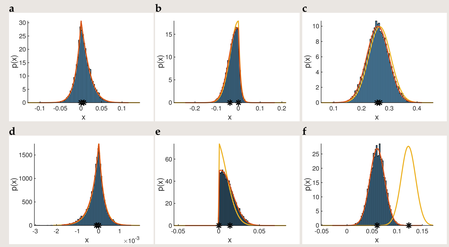

Fig. 2: Marginal posterior distributions for the diabetes (top) and leukemia (bottom) data for a zero, transition and non-zero maximum-likelihood predictor (left to right). Blue, Gibbs sampling; red, stationary phase approximation; yellow, maximum-likelihood-based approximation.

Predictive accuracy of the stationary phase approximation is comparable to state-of-the-art Gibbs sampling implementations for the Bayesian lasso and horseshoe (BayReg package) (Fig. 3).

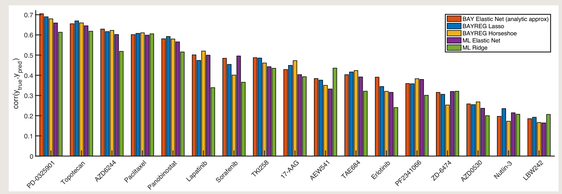

Fig 3: Median accuracy for predicting dug sensitivity from gene expression in the Cancer Cell Line Encyclopedia using 10× cross-validation, using the analytic approximation for the Bayesian elastic net (red), BayReg's Bayesian lasso (blue) and horseshoe (yellow), and maximum-likelihood elastic net (purple) and ridge regression (green).

FAQ

- **Why Bayesian lasso/elastic net? Expectation values for the coefficients are not sparse!**
  Bayesian lasso/elastic net is suitable to model **heavy-tail** effect sizes. This is more realistic for many natural systems (e.g. genetic association studies) than normally distributed or strictly sparse effect sizes.
- **Can results be extended to logistic regression or other generalized linear models?**
  The non-differentiable double-exponential prior transforms to a well-behaved exponential of a log-barrier function in **all** ℓ_1-penalized GLMs. In eqs. (4), C will then be replaced by the Hessian of the unpenalized model evaluated at the saddle point itself, making the numerics more complicated.
- **How to determine the inverse temperature and why can it be assumed large?**
  The inverse temperature $\tau = n/\sigma^2$ is large because we assume large sample size n and small residual variance σ^2. In our experiments, we determined $\hat{\tau}$ by its MAP value or by cross-validation.

Conclusions

- Expressing intractable partition function integrals as complex-valued oscillatory integrals through the **Fourier transform** is a powerful approach for performing Bayesian inference in ℓ_1-penalized models.
- Use of the **stationary phase approximation** to these integrals results in highly accurate estimates for the posterior expectation values, marginal posterior distributions, and posterior predictive distributions at a much reduced computational cost compared to Gibbs sampling.
- The analytical methods are generic and show that **powerful duality principles exist to study Bayesian inference problems**. These are generalizations to finite inverse temperature of the convex duality principles that have been essential to characterize analytical properties of the maximum-likelihood solutions of ℓ_1-penalized and other regression models.

Acknowledgments

This work was supported by grants from the BBSRC (B/J004235/1, BB/M020053/1) while the author was affiliated with the University of Edinburgh (2012–2018).

Paper

Available at **arxiv.org/abs/1709.08535**

Code

Available at **github.com/tmichoel/bayonet**

AI/ML positions available

UiB is expanding research in AI/ML and recruiting at all levels. See **uib.no/en/ii**.

COMPUTATIONAL BIOLOGY UNIT
DEPARTMENT OF INFORMATICS
UNIVERSITY OF BERGEN