MAP Estimation via Simulated Annealing for Sparse Bayesian Regression using MCMC sampling

Sudhir Raman
University of Basel, Switzerland
sudhir.raman@unibas.ch

Volker Roth
University of Basel, Switzerland
volker.roth@unibas.ch

1 Background

One of the frequently encountered modeling scenarios involves high-dimensional data with small number of measurements. In such situations, finding meaningful explanations to data through simpler models is often desired. Such simple explanations are highly desired especially in the context of biological data. Consider the standard linear regression model which explains real-valued observations $\boldsymbol{y} = (y_1, \ldots, y_n)^t$ as products of input vectors $\boldsymbol{x}_i \in \mathbb{R}^p$ and regression coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^t$, with additional additive noise:

$$y_i = \boldsymbol{x}_i^t \boldsymbol{\beta} + \epsilon_i \Leftrightarrow \boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where X is the $n \times p$ "design" matrix containing the vectors of input variables as rows. It is usually assumed that the noise terms ϵ_i are uncorrelated and follow a normal distribution. In many practical applications of regression, we are not only interested in finding regression models (i.e. coefficients $\boldsymbol{\beta}$) which are good for predicting the target variable $\boldsymbol{y}$, but also in identifying important explanatory factors through sparse models. These explanatory factors might correspond to individual input variables and higher-order interactions, and sparsity in this context interprets to a classical variable selection problem. Traditionally, such sparsity, in terms of variables, has been imposed in the form of a Lasso (ℓ_1-norm) constraint on $\boldsymbol{\beta}$ as introduced in (Tibshirani, 1996). In order to provide detailed information about the posterior distribution of the regression coefficients, a Bayesian formulation for the same has been introduced by (Park & Casella, 2008) and (Kyung et al., 2010). This model has been further generalized to handling sparsity on grouped domains in (Raman et al., 2009) and (van Gerven et al., 2009). A key feature of some of these models is the introduction of auxiliary variables which makes inference feasible through the use of MCMC sampling and these variables are integrated out stochastically.

Although we can summarize the posterior distribution in traditional ways like estimating the first moment, for practical applications, additionally a sparse point estimate can provide a simple and meaningful way of communicating the results of data analysis effectively to the applications side. In this paper, first we show that it is trivial to extend an existing Bayesian framework (the lasso case of the model defined in (Raman & Roth, 2009)) to generate a MAP estimate. This extension is based on simulated annealing (as defined in (Kirkpatrick et al., 1983) and (Černý, 1985)) and is achieved by introducing a computational temperature parameter to the existing framework. The inference is again carried out by MCMC sampling since all conditional posterior distributions are of standard form. It is however, non-trivial to assume that such an estimate will also be sparse. This is due to the fact that the MAP estimation problem changes due to the introduction of auxiliary variables and annealing proceeds to find a MAP of regression coefficients and auxiliary variables jointly (instead of MAP of regression coefficients with auxiliary variables integrated out). A key contribution of this work is to show that inspite of this change, the MAP estimate found by annealing is still sparse with respect to the regression coefficients and that it only results in a re-adjustment of the level of sparsity in the estimation.

Another problem associated with the classical Lasso is the inclusion of too many features (Meinshausen, 2007). This problem can be overcome by applying an ℓ_p norm constraint where $p < 1$. It

however makes the problem non-convex which can no longer be suitably addressed by methods like gradient descent. We show that the same Bayesian framework can be used to address this problem as well by using a flexible class of priors on the $\boldsymbol{\beta}$ coefficients as introduced in (Caron & Doucet, 2008) and generalized to grouped domains in (Raman & Roth, 2009).

2 Method

Model Description. We begin by describing the problem setting with respect to a standard regression problem given by:

$$y_i = \boldsymbol{x_i}^t \boldsymbol{\beta} + \epsilon, \qquad \text{where } \epsilon \sim N(0, \sigma^2). \tag{2}$$

The Bayesian view of the same can be expressed as:

$$p(\boldsymbol{\beta}|\boldsymbol{x}_i, y_i, \sigma^2) \propto N(y_i|\boldsymbol{x_i}^t\boldsymbol{\beta}, \sigma^2)\, p(\boldsymbol{\beta}), \tag{3}$$

where we need to translate the Lasso constraint in the form of a suitable prior over the regression coefficients $\boldsymbol{\beta}$ to obtain the posterior distribution. We use the Lasso case of the flexible prior defined in (Raman & Roth, 2009), which is expressed as a two-level hierarchical model, by introducing auxiliary variables Λ, in order to make posterior analysis feasible. The complete hierarchical model including all the hyperpriors (see (Raman et al., 2009) for details) is given in Figure 1.

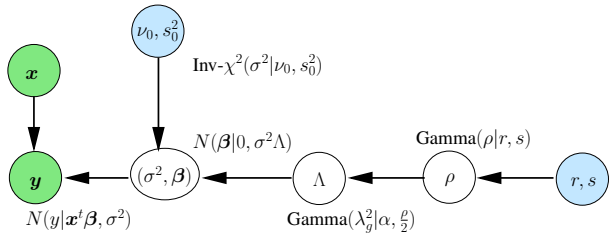

Figure 1: Shows the full hierarchical model including all the variables and their respective distributions.

Since the conditional posteriors for all the variables are of standard form, the inference is carried out using Gibbs sampling. The two variables σ^2 and ρ together decide the constraint value in the Lasso optimization problem. Hence, it would not make sense to optimize over these variables as well since the solution would favor having a lesser influence of the constraint. Therefore, based on the resulting posterior samples, we find the expected values for σ^2 and ρ and fix it for the next step and justify it as a model selection procedure. We now extend this framework to generate MAP estimates, by annealing the posterior distribution over the remaining two variables ($\boldsymbol{\beta}$ and Λ). A standard annealing procedure occurs over a discrete domain, where samples are generated from the posterior distribution over discrete variables parameterized by a computational temperature parameter T, based on a cooling schedule. Asymptotically, this results in the sampling process converging to the set of global maxima. Following the same procedure for our model, we introduce a computational temperature parameter T as follows:

$$p(\boldsymbol{\beta}, \Lambda|\boldsymbol{X}, \boldsymbol{y}, \sigma^2, \alpha, \rho, T) \propto \overbrace{\left(\prod_{i=1}^{n} N(y_i|\boldsymbol{x}_i^t\boldsymbol{\beta}, \sigma^2)\right)^{\frac{1}{T}}}^{Likelihood} \cdot \overbrace{q(\Lambda)\left(\prod_{g=1}^{p} N(\beta_g|0, \lambda_g^2\sigma^2 I)\right)^{\frac{1}{T}}\left(\prod_{g=1}^{p} \text{Gamma}(\lambda_g^2|\alpha, \frac{\rho}{2})\right)^{\frac{1}{T}}}^{JointPrior}, \tag{4}$$

where $q(\Lambda)$ is a function in Λ introduced for normalization purposes. To perform posterior inference, we again use a Gibbs sampling strategy, since for this case, all the conditional posteriors retain the same standard forms as in the original problem with only a change in the parameters. We now discuss the sparse nature of this MAP estimate.

Sparse nature of the joint MAP estimate. The above defined annealing is done on a joint distribution of $\boldsymbol{\beta}$ and Λ, hence resulting in solving the optimization problem MAP_{joint} : $max_{\beta,\Lambda}[p(\boldsymbol{\beta}, \Lambda)]$ instead of $MAP_{original}$: $max_{\boldsymbol{\beta}}[p(\boldsymbol{\beta})]$ with Λ integrated out, which was originally the desired optimization. We now justify that inspite of this change, MAP_{joint} will still result in a sparse estimate with respect to the regression coefficients. Assuming the convergence of the overall annealing procedure, we analyze the asymptotic behavior of the conditional posterior distribution of each λ_g^2 and then use that to infer the asymptotic behavior of the conditional posterior distribution of $\boldsymbol{\beta}$. We start the analysis by fixing the value of α (the sparsity parameter) based on the level of sparsity required. Based on the annealed conditional posterior for λ_g^2 (a generalized inverse gaussian distribution), we show that as $T \to 0$, the variance of this distribution goes to zero and hence λ_g^2 asymptotically converges to its mode. For the special case of $\alpha = 1.5$, it can be shown analytically using standard asymptotic results related to modified Bessel function of the second kind. With this result, we now turn back to our original joint prior of (β_g, λ_g^2) in eq. (4) and replace λ_g^2 with the mode value. We then look at the properties of this prior defined based on a fixed asymptotic conditional value of λ_g^2. For the special case of $\alpha = 1.5$, this prior (after simplification) results in a Laplace distribution in β_g, Laplace$(\beta_g|\rho, \sigma)$ (see Left panel of Figure 2 for the plot of this distribution). As mentioned earlier, we observe that this prior is parameterized by (ρ, σ^2) which plays the role of the constraint value in the Lasso based optimization problem.

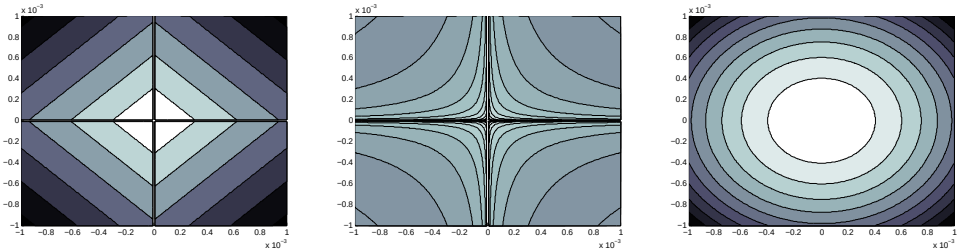

Figure 2: **Left:** Plot of the prior for $\boldsymbol{\beta}$ for $\alpha = 1.5$ which resembles the Lasso constraint. **Center:** Plot of the prior for $\boldsymbol{\beta}$ for $\alpha = 1.0$ which is also observed to be a sparsity inducing prior. **Right:** Plot of the prior for $\boldsymbol{\beta}$ for $\alpha = 2.0$ which resembles a gaussian distribution.

For ($\alpha < 1.5$), we obtain a more complicated asymptotic prior for $\boldsymbol{\beta}$ (see Center panel of Figure 2). Hence we observe that for $\alpha \leq 1.5$, the distribution over $\boldsymbol{\beta}$ represents a sparsity inducing prior. For the case of $\alpha < 1.5$, this prior is a Bayesian counterpart of a constraint in the optimization-based formulation which is non-convex. Further, for $\alpha > 1.5$, the distribution over $\boldsymbol{\beta}$ becomes a non-sparsity inducing prior and resembles the Gaussian distribution (see Right panel of Figure 2).

3 Conclusion

In conclusion, we have defined a unified Bayesian framework for sparse regression, which can be used for either a full posterior analysis of the regression coefficients, and additionally to generate a sparse MAP estimate. We observed the change in the MAP estimation problem due to the introduction of auxiliary variables. A key part of our analysis is the justification that the MAP estimate will be sparse. Apart from this dual purpose viewpoint, this model is also capable of solving a non-convex optimization problem to deal with the issue of too many non-zero coefficients in the classical Lasso case. As future work, we hope to extend the above arguments to the Group-Lasso problem where the regression coefficients are grouped (with the groups known in advance). Also, we would extend the model beyond standard regression to generalized linear models like the Poisson, Binomial models etc.

References

Caron, F., & Doucet, A. (2008). Sparse bayesian nonparametric regression. *ICML '08* (pp. 88–95). Helsinki, Finland: ACM.

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, *220, 4598*, 671–680.

Kyung, M., Gill, J., Ghosh, M., & Casella, G. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis*, *5*, 369–412.

Meinshausen, N. (2007). Relaxed lasso. *Computational Statistics & Data Analysis*, *52*, 374–393.

Park, T., & Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association*, *103*, 681–686.

Raman, S., Fuchs, T., Wild, P., Dahl, E., & Roth, V. (2009). The Bayesian Group-Lasso for analyzing contingency tables. *Proceedings of the 26th International Conference on Machine Learning* (pp. 881–888). Montreal: Omnipress.

Raman, S., & Roth, V. (2009). Sparse Bayesian regression for grouped variables in generalized linear models. *Proceedings of the 31st DAGM Symposium on Pattern Recognition* (pp. 242–251). Berlin, Heidelberg: Springer-Verlag.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Stat. Soc. B*, *58*, 267–288.

van Gerven, M., Cseke, B., Oostenveld, R., & Heskes, T. (2009). Bayesian source localization with the multivariate laplace prior. *Advances in Neural Information Processing Systems 22* (pp. 1901–1909).

Černý, V. (1985). Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. *Journal of Optimization Theory and Applications*, *45*, 41–51.