Efficient Derivation of Sparse Representations in Radial Basis Function Regression

Ernest Fokoué
Rochester Institute of Technology, Center for Quality and Applied Statistics,
98 Lomb Memorial Drive, Rochester, NY 14623, USA.
E-mail: ernest.fokoue@gmail.com

Introduction

We are given a training set $\mathcal{D} = \{(\mathbf{x}_i, \mathrm{y}_i), i = 1, \cdots, n : \ \mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^p, \mathrm{y}_i \in \mathbb{R}\}$ where y_i are realizations of $Y_i = f^*(\mathbf{x}_i) + \epsilon_i$, and ϵ_i is the noise term. For simplicity and without loss of generality, we shall assume throughout this paper that the data are standardized. We assume that the true function f^* can be approximated by

$$f_n(\mathbf{x}) = \sum_{j=1}^{n} \mathrm{w}_j \phi(\|\mathbf{x} - \mathbf{x}_j\|), \tag{1}$$

where $\phi : \mathbb{R}^+ \to \mathbb{R}$ is the basis function, and the norm $\| \cdot \|$ is the ordinary Euclidean norm on $\mathbb{R}^p$. We further assume that the basis function ϕ is a fixed radially symmetric function with respect to the norm, so that it has all the symmetries of the unit ball in $\mathbb{R}^p$, $\forall \boldsymbol{u} \in \mathcal{X}$, $\Phi(\boldsymbol{u}) = \phi(\|\boldsymbol{u}\|)$. With all that, the function f_n as defined in Eq. (1) is called a radial basis function (RBF) with weights $\mathrm{w}_1, \mathrm{w}_2, \cdots, \mathrm{w}_n$ and centers $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$. In this paper, we use the most popular radial basis function namely the so called Gaussian radial basis function kernel corresponding to $\phi(\mathrm{u}) = \exp(-r\mathrm{u}^2)$, where r represents a function of the bandwidth. This approach to regression really became popular in the machine learning and statistics communities after the publication by [5] of the Relevance Vector Machine (RVM). One of the most important aspects of radial basis function regression - besides the crucial issue of the choice of the kernel - is the search for a sparse representation. Indeed, sparsity was the professed motivation of [5], and later of [1]. In fact, for most situations and indeed most kernels, the statistical estimation of the weights w_j's by traditional error minimization (least squares) or density maximization (MLE) methods turns out to be an illposed problem, for which there is no hope of a decent solution without some form of regularization or constraints to help stabilize the solution. Regularization in and of itself does not necessarily yield a sparse solution. Indeed, the form of the regularizer and/or appropriate subsequent refinements performed on the regularized solution are the keys to obtaining the desired level of sparsity. We later argue in this paper, in the spirit of [3] and [1] that sparsity in the end is analogous to traditional model selection and can therefore be arrived at using search techniques. In fact, we argue with an even greater emphasis that our technique goes a step further by making the search straightforward, computationally efficient, interpretable and predictively optimal. [6] adds credence to our claim with his theoretical study of the consistency of Silverman's g-Prior in kernel regression. The rest of this paper is organized as follows: Section 2 presents our proposed algorithm and provides details of its properties. Section 3 shows some computational results along with important comparisons. Section 4 concludes and gives a few pointers for extensions and improvements on the present work.

Naturally Efficient Sparsity Tuner

We make the usual assumption that the noise terms are independent zero-mean Gaussian random variables with the same variance σ^2, i.e. $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0,\sigma^2)$. As a result, the likelihood is

$$p(\mathbf{y}\,|\,\mathbf{w},\sigma^2) = \mathcal{N}_n(\mathbf{y}\,|\,\mathbf{K}\mathbf{w},\sigma^2\mathbf{I}_n), \tag{2}$$

where $\mathbf{y} = (\mathrm{y}_1, \mathrm{y}_2, \cdots, \mathrm{y}_n)^\mathsf{T}$, $\mathbf{w} = (\mathrm{w}_1, \mathrm{w}_2, \cdots, \mathrm{w}_n)^\mathsf{T}$, $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \cdots, \epsilon_n)^\mathsf{T}$ and $\mathbf{K} = (\mathbf{K}_{ij})$ where $\mathbf{K}_{ij} = \phi(\|\mathbf{x}_i - \mathbf{x}_j\|)$, $i,j = 1,\cdots,n$. We specify a Gaussian prior for the weights, namely $\mathbf{w} \sim \mathcal{N}_n(0,\mathbf{Q})$, where the prior variance matrix $\mathbf{Q}$ is allowed to take a variety of forms. Finally, we adopt the maximum a posteriori (MAP) approach as our initial estimation technique. The resulting posterior is therefore conveniently Gaussian, namely, $p(\mathbf{w}|\mathbf{y},\mathbf{Q},\mathbf{K},\sigma^2) = \mathcal{N}_n(\mathbf{w}|\mu_\mathbf{w},\Sigma_\mathbf{w})$, with the mean and the variance matrix given respectively by $\mu_\mathbf{w} = \sigma^{-2}\left[\sigma^{-2}\mathbf{K}^\mathsf{T}\mathbf{K} + \mathbf{Q}^{-1}\right]^{-1}\mathbf{K}^\mathsf{T}\mathbf{y}$ and $\Sigma_\mathbf{w} = \left[\sigma^{-2}\mathbf{K}^\mathsf{T}\mathbf{K} + \mathbf{Q}^{-1}\right]^{-1}$. Crucially, the posterior information matrix

$$\mathbf{M} = \sigma^{-2}\mathbf{K}^\mathsf{T}\mathbf{K} + \mathbf{Q}^{-1} \tag{3}$$

is central to our work, as we ultimately seek to select those atoms that maximize the size of $\mathbf{M}$. Thanks to Gaussianity, our MAP estimator of $\mathbf{w}$ is therefore the above $\mu_\mathbf{w}$, or more explicitly

$$\hat{\mathrm{w}}_{\mathsf{MAP}} = \sigma^{-2}\left[\sigma^{-2}\mathbf{K}^\mathsf{T}\mathbf{K} + \mathbf{Q}^{-1}\right]^{-1}\mathbf{K}^\mathsf{T}\mathbf{y} = \sigma^{-2}\mathbf{M}^{-1}\mathbf{K}^\mathsf{T}\mathbf{y}. \tag{4}$$

The Naturally Efficient Sparsity Tuner (NEST) technique consists in selecting the most relevant basis functions for equation (1) by first forming the MAP estimator of $\mathbf{w}$, then sorting the w_j^2 in decreasing

order. Now, starting from an empty set, expand the set with the next index that does not decrease the size of the posterior information matrix, skipping those that cause negligible changes.

Let $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \cdots, \gamma_n)^\top \in \{0,1\}^n$ denote the vector of indicator variables such that

$$\gamma_j = \begin{cases} 1 & \text{basis function } j \text{ is part of the best model} \\ 0 & \text{otherwise} \end{cases}$$

Let $d_{\boldsymbol{\gamma}} = |\boldsymbol{\gamma}| = \sum_{j=1}^{n} \gamma_j$ denote the number of active atoms in the current model. We will use $\mathbf{K}_{\boldsymbol{\gamma}} \in I\!R^{n \times d_{\boldsymbol{\gamma}}}$ to denote the $n \times d_{\boldsymbol{\gamma}}$ matrix made up of the $d_{\boldsymbol{\gamma}}$ currently active atoms acting on the whole training set. Also, we will use $\mathbf{K}_{\boldsymbol{\gamma}\boldsymbol{\gamma}} \in I\!R^{d_{\boldsymbol{\gamma}} \times d_{\boldsymbol{\gamma}}}$ to denote the $d_{\boldsymbol{\gamma}} \times d_{\boldsymbol{\gamma}}$ design matrix restricted to the $d_{\boldsymbol{\gamma}}$ currently active atoms. Finally, we use $\mathbf{M}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}$ to denote $d_{\boldsymbol{\gamma}} \times d_{\boldsymbol{\gamma}}$ submatrix of $\mathbf{M}$ associated with the $d_{\boldsymbol{\gamma}}$ currently active atoms.

NEST: Naturally Efficient Sparsity Tuner

```
1. Compute $\hat{\mathbf{w}} = \sigma^{-2}\mathbf{M}^{-1}\mathbf{K}^{\mathsf{T}}\mathbf{y}$, where $\mathbf{M} = \sigma^{-2}\mathbf{K}^{\mathsf{T}}\mathbf{K} + \mathbf{Q}^{-1}$.
2. Rank $\hat{\mathrm{w}}^2_{(1)} \leq \hat{\mathrm{w}}^2_{(2)} \leq \cdots \leq \hat{\mathrm{w}}^2_{(n-1)} \leq \hat{\mathrm{w}}^2_{(n)}$
3. Repeat
    (a) Let $i := i + 1$ and set $j = \mathrm{index}(\hat{\mathrm{w}}^2_{(i)})$
    (b) Compute $\mathcal{E}_i = \log\det\left(\mathbf{M}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\right)$
    (c) if $|\mathcal{E}_i - \mathcal{E}_{i-1}| > \epsilon$ then activate $\gamma_j = 1$ and update $\boldsymbol{\gamma}$
    (d) stop $= (\mathcal{E}_i < \mathcal{E}_{i-1})$
4. Until stop $=$ TRUE
5. Find optimal $k$
        $k^* = \arg\max_{i \in 1,\cdots,n} \{\mathcal{E}_1, \cdots, \mathcal{E}_n\}.$
6. Extract dominant points
        $\boldsymbol{\gamma}^* = \{\boldsymbol{\gamma} : \ \gamma_j = 1, \ \forall j = \mathrm{index}(\hat{\mathrm{w}}^2_{(i)}) : \ i = 1, \cdots, k^*\}$
7. Refine $\hat{\mathbf{w}}_{\boldsymbol{\gamma}^*} = \sigma^{-2}\mathbf{M}^{-1}_{\boldsymbol{\gamma}^*\boldsymbol{\gamma}^*}\mathbf{K}^{\mathsf{T}}_{\boldsymbol{\gamma}^*}\mathbf{y}$.
```

Conjecture 1 *Assume that the variance matrix is well conditioned. Then there exist* $\mathrm{g} > 0$ *and* $\eta > 0$ *such that the solution yielded by* NEST *coincides with*

$$(5) \qquad \left[\begin{array}{l} \boldsymbol{\gamma}^{*} = \arg \max\limits_{\boldsymbol{\gamma} \in \{0,1\}^{p}} \log \det \left(\mathbf{M}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\right) \\ \textit{subject to } \|\sigma^{-2}\mathbf{M}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}^{-1}\mathbf{K}_{\boldsymbol{\gamma}}^{\top}\mathbf{y}\|_{2}^{2} < \eta. \end{array}\right.$$

The concavity of the objective function turns out to be of paramount importance with regard to computational efficiency, as one does not need to traverse the entirety of the basis functions in order to determine the most relevant ones. It suffices to stop when the value of the objective function decreases.

Numerical explorations

For a simple illustration of the merit of our proposed method, we consider recovering the univariate sinc function from noisy observations. Specifically, we have $f^{*}(\mathbf{x}) = \frac{\sin 10\mathbf{x}}{10\mathbf{x}}$, $\mathbf{x} \in [-1, +1]$. We generate $n = 199$ points with noise variance $\sigma^{2} = 0.2^{2}$. We also find the bandwidth for the Gaussian kernel to be 0.25. Note that by standardizing, we simply mean using $\tilde{\mathbf{y}} = \mathbf{y} - \mathrm{v}\mathbf{1}_{n}$ in the derivations in place of $\mathbf{y}$, which in practice corresponds to using $\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}_{n}$ in the computations. We find it convenient to set $\mathrm{g} = 1/(cn)$ to reflect the inherent relationship between g and model complexity, which in this case is $p = n$. With that our tuning parameter is now c which is found to be a positive number. We consider the simplest of regularization schemes, namely the ridge regularizer, which corresponds to using *isotropic* prior variance matrix $\mathbf{Q} = \mathrm{g}^{-1}\mathbf{I}_{n}$ for some $\mathrm{g} > 0$. Until [3], it was strongly held that the contours of the gaussian distribution make it impossible for ridge regularizer to be sparse. [3] essentially demonstrated the equivalence between the LASSO and an adaptive ridge approach.

Whereas [3] was based on traditional linear regression models for which LASSO is mostly used, [1] demonstrated that one can obtain very sparse solution with a ridge regularizer in the kernel regression context via posterior simulation.

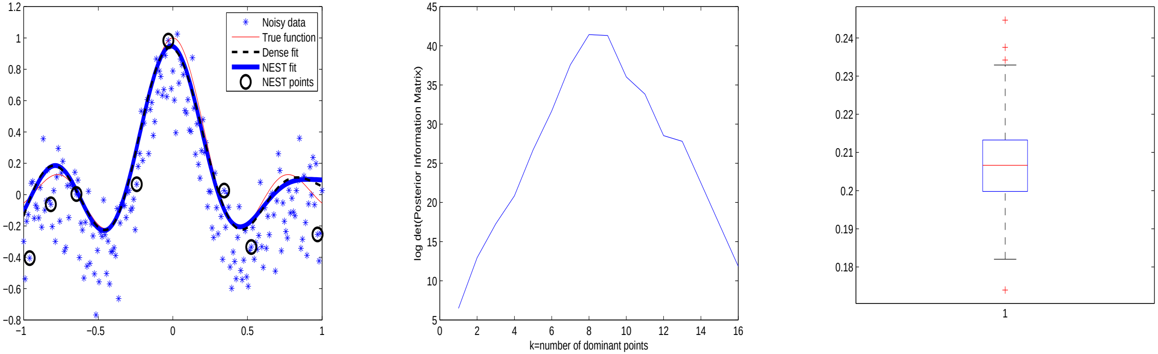

Figure 1: (left) The fit is excellent with only 8 relevant points out of $n = 199$ (center) the size of the information matrix is convex and seen here peaking around 8 (right) This box plot shows the estimated predictive root mean squared error over 25 replications. The median in is just under 0.21, for a data set with noise standard deviation of 0.2, revealing the excellent performance of NEST.

Conclusion

We have proposed a novel technique for finding sparse representations in radial basis function regression. It is somewhat surprising that sparsity can be achieved with an inherently non-sparsity inducing

prior distribution like the isotropic Gaussian. Our take on this is that sparsity at its core is really basis selection, which in turn might well depend more on the selection algorithm than the shape of the penalty. We claim, at the least in the radial basis function regression setting, that our proposed technique, by virtue of the fact that it is based on a suitably chosen convex objective function, is computationally more efficient than its our predecessors when it comes to finding sparse solutions. Besides, our technique has a very solid theoretical foundation inherited from the theory of D-optimality. A more thorough and comprehensive description of NEST can be found in [2] where the technique is explored from a variety of aspects like the effect of sample size, the effect of the prior variance matrix - Zellner's g-prior [4], and Silverman's g-prior [6] - just to name a few.

References

[1] Fokoué, E. (2008). Estimation of Atom Prevalence for Optimal Prediction. *Contemporary Mathematics*. Vol **443**, pp 103-129, The American Mathematical Society.

[2] Fokoué, E. (2009). Naturally Efficient Sparsity Tuner for Radial Basis Function Regression, *Technical Report*. Center for Quality and Applied Statistics, Rochester Institute of Technology, 98 Lomb Memorial Drive, Rochester, New York, USA.

[3] Grandvalet, Y. (1998). Least absolute shrinkage is equivalent to quadratic penalization. In L. Niklasson,M. Boden, and T. Ziemske, editors, ICANN98, volume 1 of Perspectives in Neural Computing, pages 201206. Springer, 1998.

[4] F. Liang and R. Paulo and G. Molina and M. Clyde and J. O. Berger (2008). Mixtures of g-priors for Bayesian Variable Selection. *J. Amer. Statist. Assoc.*, **103**, 410-423.

[5] Tipping, M. E. (2001). Sparse Bayesian Learning and the Relevance Vector Machine. *Journal of Machine Learning Research*, **1**, 211-244.

[6] Zhang, Z., Jordan, M. I. and Yeung, D. (2008). Posterior Consistency of the Silverman g-Prior in Bayesian Model Choice. *Technical Report*, Number xx, Department of Electrical Engineering and Computer Science, University of California, Berkeley, California, USA.