

Technical Appendix on Sparse Bayesian Regression

Loic Le Folgoc, Hervé Delingette, Antonio Criminisi, Nicholas Ayache

► **To cite this version:**

Loic Le Folgoc, Hervé Delingette, Antonio Criminisi, Nicholas Ayache. Technical Appendix on Sparse Bayesian Regression. MICCAI - 17th International Conference on Medical Image Computing and Computer Assisted Intervention, Sep 2014, Boston, United States. <hal-01002861>

HAL Id: hal-01002861

<https://hal.inria.fr/hal-01002861>

Submitted on 6 Jun 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Technical Appendix on Sparse Bayesian Regression

Loïc Le Folgoc¹, Hervé Delingette¹, Antonio Criminisi², and Nicholas Ayache¹

¹ Asclepios Research Project, INRIA Sophia Antipolis, France

² Machine Learning and Perception Group, Microsoft Research Cambridge, UK

Abstract. We report the technical details for a sparse bayesian approach to regression. It can be seen as an extension of the Relevance Vector Machine of Tipping *et al* [1] to a more general setting that can handle vector-valued regression and generic quadratic priors.

1 Quadratic Energies & Marginal Likelihood of the Data

We want to minimize in \mathbf{w} the following cost:

$$E(\mathbf{w}) = (\mathbf{t} - \Phi\mathbf{w})^\top \beta \mathbf{H} (\mathbf{t} - \Phi\mathbf{w}) + \mathbf{w}^\top (\mathbf{A} + \lambda \mathbf{R}) \mathbf{w} \quad (1)$$

and jointly optimize in $\mathbf{A} = \text{diag}(\mathbf{A}_i)$, β , λ . $\mathbf{w}|\mathbf{t}, \mathbf{A}, \lambda, \beta$ follows a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$, where

$$\boldsymbol{\mu} = \Sigma \Phi^\top \beta \mathbf{H} \mathbf{t}, \quad \Sigma = (\mathbf{A} + \lambda \mathbf{R} + \Phi^\top \beta \mathbf{H} \Phi)^{-1} \quad (2)$$

A key element is that the distribution of $\mathbf{t}|\mathbf{A}$ is also Gaussian, $\mathbf{t}|\mathbf{A} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$, with

$$\mathbf{C} = (\beta \mathbf{H})^{-1} + \Phi (\mathbf{A} + \lambda \mathbf{R})^{-1} \Phi^\top \quad (3)$$

Indeed, we see that:

$$\begin{aligned} p(\mathbf{t}|\mathbf{A}) &= \int p(\mathbf{t}|\mathbf{w})p(\mathbf{w}|\mathbf{A})d\mathbf{w} \\ &\propto \int \exp -\frac{1}{2}(\mathbf{t} - \Phi\mathbf{w})^\top \beta \mathbf{H} (\mathbf{t} - \Phi\mathbf{w}) \cdot \exp -\frac{1}{2}\mathbf{w}^\top (\mathbf{A} + \lambda \mathbf{R}) \mathbf{w} \cdot d\mathbf{w} \\ &\propto \exp -\frac{1}{2}\boldsymbol{\mu}^\top \Sigma^{-1} \boldsymbol{\mu} \cdot \exp -\frac{1}{2}\mathbf{t}^\top \beta \mathbf{H} \mathbf{t} \cdot \int \exp -\frac{1}{2}(\mathbf{w} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{w} - \boldsymbol{\mu}) d\mathbf{w} \end{aligned}$$

where the integral sums to 1. Thus using (2),

$$p(\mathbf{t}|\mathbf{A}) \propto \exp -\frac{1}{2}\mathbf{t}^\top (\beta \mathbf{H}) \mathbf{t} - \frac{1}{2}\mathbf{t}^\top (\beta \mathbf{H}) \Phi \Sigma \Phi^\top (\beta \mathbf{H}) \mathbf{t}$$

$\mathbf{t}|\mathbf{A}$ is Gaussian since the distribution is proportional to a Gaussian, and by identification it ensues that $\mathbf{C}^{-1} = \beta \mathbf{H} - (\beta \mathbf{H}) \Phi \Sigma \Phi^\top (\beta \mathbf{H})$. We then get the desired result using a matrix inversion identity. The fast RVM algorithm proceeds by iteratively implementing a single change to one of the \mathbf{A}_i 's on the block-diagonal matrix \mathbf{A} ; specifically the one that maximizes the increase of a quantity known as the *evidence*, $\log p(\mathbf{t}|\mathbf{A})$, then re-estimating the parameters of the conditional posterior $\mathbf{w}|\mathbf{t}, \mathbf{A}, \lambda, \beta$ using (2). The algorithm starts with all \mathbf{A}_i set to ∞ . The computations involve rank-one "block" updates; it also turns out that the optimal \mathbf{A}_i 's are rank one matrices (so we actually have rank-one updates).

2 Computation of the gain in evidence for a given action

We want to evaluate the change in $\log p(\mathbf{t}|\mathbf{A}_{-i}, \mathbf{A}_i)$ when a single prior weight \mathbf{A}_i is changed. Recall that $\mathbf{t}|\mathbf{A}_{-i}, \mathbf{A}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$, thus

$$\log p(\mathbf{t}|\mathbf{A}_{-i}, \mathbf{A}_i) = -1/2 \{ \log |\mathbf{C}| + \mathbf{t}^\top \mathbf{C}^{-1} \mathbf{t} \} \quad (4)$$

Now let's single out the contribution of \mathbf{A}_i , for each of the two terms above. Noting the form of eq. (3), let $\mathbf{L} = (\mathbf{A} + \lambda \mathbf{R})^{-1}$. We can single out the contribution of the i th basis to \mathbf{L} first, as a rank one update:

$$\mathbf{L} = \begin{pmatrix} \mathbf{L}_{-i} & 0 \\ 0 & 0 \end{pmatrix} + \mathbf{U}_i l_i \mathbf{U}_i^\top \quad (5)$$

with $\mathbf{U}_i^\top = ((\lambda \mathbf{L}_{-i} \mathbf{R}_i)^\top \mathbf{I}^\top)$ and $l_i = \{\mathbf{A}_i + \lambda \mathbf{R}_{ii} - (\lambda \mathbf{R}_i)^\top \mathbf{L}_{-i} (\lambda \mathbf{R}_i)\}^{-1} \triangleq \{\mathbf{A}_i + \kappa_i\}^{-1}$. Note also that any basis for which $\mathbf{A}_j = \infty$ can be disregarded, since its corresponding line and column in \mathbf{L} and \mathbf{L}_{-i} will be null. We can interpret \mathbf{L} as a square matrix of dimension the number of active bases (including the basis under consideration), and the algorithmic complexity of matrix operations involving \mathbf{L} or \mathbf{L}_{-i} will indeed be related to the reduced sized of these matrices. This is also true for \mathbf{U}_i , as a direct consequence, and for all of the other quantities involved.

Injecting (5) into (3) gives a decomposition of \mathbf{C} into the sum of a term that does not depend on the i th basis and of a rank-one term:

$$\mathbf{C} = \mathbf{C}_{-i} + \Phi^i \mathbf{U}_i l_i \mathbf{U}_i^\top \Phi^{i\top} \quad (6)$$

Φ is superscripted with i to recall that along with all the other active bases, the i th basis ϕ_i is present in this matrix. Using rank-one updates for, respectively, the determinant and the inverse, and letting $\mathbf{C}_{-i}^{-1} \triangleq (\mathbf{C}_{-i})^{-1}$, we get the two following expressions:

$$|\mathbf{C}| = |\mathbf{C}_{-i}| |l_i| |l_i^{-1} + \mathbf{U}_i^\top \Phi^{i\top} \mathbf{C}_{-i}^{-1} \Phi^i \mathbf{U}_i| \quad (7)$$

$$\mathbf{t}^\top \mathbf{C}^{-1} \mathbf{t} = \mathbf{t}^\top \left(\mathbf{C}_{-i}^{-1} - \mathbf{C}_{-i}^{-1} \Phi^i \mathbf{U}_i \left\{ l_i^{-1} + \mathbf{U}_i^\top \Phi^{i\top} \mathbf{C}_{-i}^{-1} \Phi^i \mathbf{U}_i \right\}^{-1} \mathbf{U}_i^\top \Phi^{i\top} \mathbf{C}_{-i}^{-1} \right) \mathbf{t} \quad (8)$$

These quantities rewrite more compactly if we introduce appropriate notations. Namely, let $q_j(i) \triangleq \phi_j^\top \mathbf{C}_{-i}^{-1} \mathbf{t} \in \mathbb{R}^d$ and $s_{jk}(i) \triangleq \phi_j^\top \mathbf{C}_{-i}^{-1} \phi_k \in \mathcal{M}_{d,d}$. The concatenation of these q_j 's for all active bases plus the basis under scrutiny (total of m bases), a.k.a $\mathbf{q}_i \in \mathbb{R}^{d \times m}$, will come in helpful. Similarly $\mathbf{s}_i \in \mathcal{M}_{d \times m, d \times m}$ will denote the matrix with $s_{kl}(i)$ as (k, l) th coefficient, where indices span the set of active bases plus the i th basis. Now, let $q^i \triangleq \mathbf{U}_i^\top \Phi^{i\top} \mathbf{C}_{-i}^{-1} \mathbf{t} = \mathbf{U}_i^\top \mathbf{q}_i \in \mathbb{R}^d$, and $s^i \triangleq \mathbf{U}_i^\top \Phi^{i\top} \mathbf{C}_{-i}^{-1} \Phi^i \mathbf{U}_i = \mathbf{U}_i^\top \mathbf{s}_i \mathbf{U}_i \in \mathcal{M}_{d,d}$. With these notations in hand and recalling that $l_i^{-1} = \mathbf{A}_i + \kappa_i$, we can rewrite eq. (7) and eq. (8) as:

$$\log |\mathbf{C}| = \log |\mathbf{C}_{-i}| - \log |\mathbf{A}_i + \kappa_i| + \log |\mathbf{A}_i + \kappa_i + s^i| \quad (9)$$

$$\mathbf{t}^\top \mathbf{C}^{-1} \mathbf{t} = \mathbf{t}^\top \mathbf{C}_{-i}^{-1} \mathbf{t} - q^{i\top} \{ \mathbf{A}_i + \kappa_i + s^i \}^{-1} q^i \quad (10)$$

Ignoring the terms that do not depend on the i th basis, we see that the contribution to the evidence of the model for a given value of \mathbf{A}_i is directly related to:

$$l(\mathbf{A}_i) = \log |\mathbf{A}_i + \kappa_i| - \log |\mathbf{A}_i + \kappa_i + s^i| + q^{i\top} \{\mathbf{A}_i + \kappa_i + s^i\}^{-1} q^i \quad (11)$$

Naturally if $\lambda = 0$ (no additional regularization) this comes down to the regular RVM, with $q^i = q_i$, $s^i = s_{ii}$ and $\kappa_i = 0$.

3 Maximization of the gain in evidence

If $q^i q^{i\top} - s^i$ has no positive eigenvalue, the maximum \mathbf{A}_i lies at infinity and the basis should remain inactive, or be removed. Otherwise the gradient of Eq. (11) provides ground to look for rank-one maximizers $\mathbf{A}_i = \alpha_i \eta_i \eta_i^\top$. To that aim we compute $n_i = s^{i-1} q^i$ and

$$a_i = \frac{(n_i^\top s^i n_i)^2}{(n_i^\top q^i)^2 - n_i^\top s^i n_i} - n_i^\top \kappa_i n_i \quad (12)$$

If $a_i \geq 0$ the maximizer is given by $\alpha_i = a_i$ and $\eta_i = n_i$. Otherwise ($a_i < 0$) we set $\alpha_i = 0$ and numerically solve over the optimal orientation η_i . This latter case arises when the regularization level alone is sufficient to make additional "shrinkage" unnecessary.

4 Update of λ

We derive an update rule via an expectation-maximization procedure. Knowing \mathbf{w} , it would be straightforward to derive an estimate of λ by maximizing the log-likelihood of \mathbf{w} or the posterior of λ given \mathbf{w} . However \mathbf{w} is hidden in our model. Instead, we try to maximize the log-likelihood on average (i.e. to minimize the average loss): $\max_\lambda \mathbb{E}_{\mathbf{w} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})} [\log p(\mathbf{t}, \mathbf{w} | \mathbf{A}, \lambda, \beta) | \mathbf{A}^*, \beta^*]$, where \mathbf{A}^* and β^* are our current estimates of the respective quantities. Discarding terms which are constants of λ , we obtain the following:

$$\lambda^* = \arg \max_\lambda -\frac{\lambda}{2} \text{tr}(\boldsymbol{\Sigma} \mathbf{R}) + \frac{1}{2} \log |\mathbf{A} + \lambda \mathbf{R}| - \frac{\lambda}{2} \boldsymbol{\mu}^\top \mathbf{R} \boldsymbol{\mu} \quad (13)$$

Deriving leads to:

$$\frac{\partial}{\partial \lambda} f(\lambda) \propto \text{tr}(\{\mathbf{A} + \lambda \mathbf{R}\}^{-1} \mathbf{R}) - \text{tr}(\boldsymbol{\Sigma} \mathbf{R}) - \boldsymbol{\mu}^\top \mathbf{R} \boldsymbol{\mu} \quad (14)$$

This is a decreasing function of λ and thus has at most one zero. If $\frac{\partial}{\partial \lambda} f(\lambda)$ is negative at the origin, $\lambda^* = 0$. Otherwise, we optimize by using the Newton method on a log scale. This is motivated by the fact that the function of interest is not only decreasing, but also smooth and convex. Lastly note that $\{\mathbf{A} + \lambda \mathbf{R}\}^{-1} \mathbf{R} = \{\mathbf{R}^{-1} \mathbf{A} + \lambda \mathbf{I}\}^{-1}$, so we can compute the eigenvalues δ_k of $\mathbf{A}^{1/2} \mathbf{R}^{-1} \mathbf{A}^{1/2}$ once and rely on the fact that $\text{tr}(\{\mathbf{A} + \lambda \mathbf{R}\}^{-1} \mathbf{R}) = \sum_k 1/(\delta_k + \lambda)$ to avoid repeated matrix inversions.

References

1. Tipping, M.E., Faul, A.C., et al.: Fast marginal likelihood maximisation for sparse bayesian models. In: Workshop on artificial intelligence and statistics. Volume 1., Jan (2003)