14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

ADAPTIVE REGULARIZATION OF NOISY LINEAR INVERSE PROBLEMS

Lars Kai Hansen, Kristoffer Hougaard Madsen, Tue Lehn-Schiøler

Informatics and Mathematical Modelling
Technical University of Denmark B321
DK-2800 Lyngby, DENMARK

ABSTRACT

In the Bayesian modeling framework there is a close relation between regularization and the prior distribution over parameters. For prior distributions in the exponential family, we show that the optimal hyper-parameter, i.e., the optimal strength of regularization, satisfies a simple relation: The expectation of the regularization function, i.e., takes the same value in the posterior and prior distribution. We present three examples: two simulations, and application in fMRI neuroimaging.

1. LINEAR INVERSE PROBLEMS

Noisy linear inverse problems are of interest in data analysis, e.g., in astronomy, computerized tomography, early vision, electrocardiography, mathematical physics and metrology ([2]). Straightforward solutions in terms of matrix inversion often provides useless solutions dominated by noise. Variable strength 'regularization' is therefore invoked to control the signal to noise ratio of the solution. Methods for tuning the *amount* of regularization have been extensively studied and include generalized cross-validation, the so-called 'L-curve', and Bayesian approaches. In this contribution we will discuss the Bayesian approach. First we provide some general results. We recapitulate a proof of the fact that Bayesian averaging is generalization optimal under ideal conditions, i.e. that we have both the correct observation model and we have complete prior knowledge. Secondly, we derive a new theoretical result for so-called empirical Bayes, which is relevant when we only know the *functional form* of the prior, but not its *strength*. We show that operating with adaptive priors in a linear regression model can lead to parameter pruning, i.e., that the optimal regularization is *infinite*. Finally, we discuss the special case of the Bayesian general linear model with conjugate priors and give a application to data analysis of functional brain imaging.

2. BAYES OPTIMALITY

Often a learning problem has natural quantitative measure of generalization. If a loss function is defined the natural measure is the so-called *generalization error*, i.e., the *expected loss* on a random sample independent of the training set. Generalizability is a key topic of learning theory and much progress has been reported. Analytic results for a broad class of adaptive systems can be found in the literature [18, 12, 13]. Typically they estimate the asymptotic generalization ability of algorithms that are parameterized by a finite number of real values parameters. Haussler and Opper presented a very rich framework for analysis of generalization for Bayesian averaging and other schemes in [7]. Analytic generalization errors for finite training sets can only be obtained for specific learning machines such as linear regression [3].

Bayesian averaging is optimal in a number of ways (admissibility, the likelihood principle etc) [16]. It is important to note that Bayesian predictions are stochastic just like predictions of any other inference scheme that generalize from a finite sample. Here we recapitulate a proof originally presented in [5] that Bayesian averaging is *generalization optimal* if the observation model and the a prior are identical to the distributions of the processes that generate the data (the likelihood) and choose the parameters (the so-called 'teacher' distribution), respectively. To prove this let us consider a model that is smoothly parameterized and whose predictions can be described in terms of a so-called predictive density. Predictions in the model are based on a given training set: a finite sample $D = \{\mathbf{y}_\alpha\}_{\alpha=1}^N$ of the stochastic vector $\mathbf{y}$ whose density – the teacher – is denoted $p(\mathbf{y}|\theta_0)$. In other words the true density is assumed to be defined by a fixed, but unknown, teacher parameter vector θ_0. The model, denoted H, involves the parameter vector θ and its predictive density is,

$$p(\mathbf{y}|D,H) = \int p(\mathbf{y}|\theta,H)p(\theta|D,H)d\theta, \tag{1}$$

$p(\theta|D,H)$ is the posterior parameter distribution. In a maximum likelihood scenario this distribution is a delta function centered on the most likely parameters under the model for the given data set. In ensemble averaging approaches, like boosting bagging or stacking, the distribution is obtained by (re-)training on re-sampled training sets. In a Bayesian scenario, the parameter distribution is the posterior distribution,

$$p(\theta|D,H) = \frac{p(D|\theta,H)p(\theta|H)}{\int p(D|\theta',H)p(\theta'|H)d\theta'} \tag{2}$$

where $p(\theta|H)$ is the prior distribution. In the following we will only consider one model hence we suppress the model conditioning label H.

The generalization error will be the expected negative log probability (also known as the 'deviance') $\Gamma(\theta_0|D) = \int -\log p(\mathbf{y}|D)p(\mathbf{y}|\theta_0)d\mathbf{y},$. The expected value of the generalization error for training sets produced by the given teacher is given by

$$\Gamma(\theta_0) = \int\int -\log p(\mathbf{y}|D)p(\mathbf{y}|\theta_0)d\mathbf{y}\,p(D|\theta_0)dD. \tag{3}$$

Playing the game of "guessing a probability distribution" [7] we not only face a random training set, we also face a teacher drawn from the teacher distribution $p(\theta_0)$. The teacher averaged generalization must then be defined as

$$\Gamma = \int \Gamma(\theta_0)p(\theta_0)d\theta_0. \tag{4}$$

Γ is the expected generalization error for a random training set sampled from the randomly chosen teacher for a given learning procedure. The generalization error is minimized by Bayes averaging if the teacher distribution is used as prior. To see this, form the Lagrangian functional

$$\begin{aligned} &\mathscr{L}[q(\mathbf{y}|D)] = \\ &\int\int\int -\log q(\mathbf{y}|D)p(\mathbf{y}|\theta_0)d\mathbf{y}p(D|\theta_0)dDp(\theta_0)d\theta_0 \\ &+ \ l(\int q(\mathbf{y}|D)d\mathbf{y}-1) \end{aligned} \tag{5}$$

defined on positive normalizable functions $q(\mathbf{y}|D)$. The Lagrange multiplier l is used to ensure that $q(\mathbf{y}|D)$ is indeed a normalized density in the domain of $\mathbf{y}$. Equating the variational derivative to zero we recover the predictive distribution of Bayesian averaging,

$$q_{\mathrm{opt}}(\mathbf{y}|D) = \int p(\mathbf{y}|\theta)\frac{p(D|\theta)p(\theta)}{\int p(D|\theta')p(\theta')d\theta'}d\theta, \tag{6}$$

where $l = \int p(D|\theta)p(\theta)d\theta$ is the normalization constant.

3. EMPIRICAL BAYES' MODELING

The main challenge using Bayes is that the prior needs to be specified correctly to prove optimality. In practical applications one can invoke so-called empirical Bayes methods, in which the prior is specified with a certain set of free hyper-parameters to determine for the given instance. Here we will derive a conceptually simple relation which can be used for setting hyper-parameters for priors in the exponential family.

Let the model be finitely parameterized by the vector θ, and the likelihood of the data set D be denoted $P(D|\theta)$. Assume a prior distribution in the exponential family:

$$P(\theta|\lambda) = \frac{\exp\left(-\lambda^\top \mathbf{f}(\theta)\right)}{\int \exp\left(-\lambda^\top \mathbf{f}(\theta)\right)d\theta} \tag{7}$$

Where $\mathbf{f}(\theta)$ is a vector of regularization functions and λ the associated vector of hyper-parameters (strengths).

The optimal hyper-parameter vector can be found in an ML-II approach by maximizing the log-posterior:

$$\begin{aligned} \log P(\lambda|D) &= \log P(D|\lambda) + \log P(\lambda) - \log P(D) \\ P(D|\lambda) &= \int P(D|\theta)P(\theta|\lambda)d\theta \end{aligned} \tag{8}$$

We have denoted a possible prior on the hyper-parameters by $P(\lambda)$. The derivative of the log-posterior is given by

$$\frac{\partial \log P(\lambda|D)}{\partial\lambda} = \frac{\int P(D|\theta)\frac{\partial P(\theta|\lambda)}{\partial\lambda}d\theta}{\int P(D|\theta)P(\theta|\lambda)d\theta} + \frac{\partial \log P(\lambda)}{\partial\lambda} \tag{9}$$

The derivative of the parameter prior has two components,

$$\frac{\partial P(\theta|\lambda)}{\partial\lambda} = [\langle \mathbf{f}(\theta)\rangle_{\mathrm{prior}} - \mathbf{f}(\theta)]P(\theta|\lambda), \tag{10}$$

where we have defined the expectation in the prior distribution, $\langle \mathbf{f}(\theta)\rangle_{\mathrm{prior}} = \int \mathbf{f}(\theta)P(\theta|\lambda)d\theta$. Inserting Eq. (10) in Eq. (9) we obtain our key result

$$\frac{\partial \log P(\lambda|D)}{\partial\lambda} = \langle \mathbf{f}(\theta)\rangle_{\mathrm{prior}} - \langle \mathbf{f}(\theta)\rangle_{\mathrm{post}} + \frac{\partial \log P(\lambda)}{\partial\lambda} \tag{11}$$

with the definition of the posterior expectation

$$\begin{aligned} \langle \mathbf{f}(\theta)\rangle_{\mathrm{post}} &= \int \mathbf{f}(\theta)P(\theta|D)d\theta \\ P(\theta|D) &= \frac{P(D|\theta)P(\theta|\lambda)}{\int P(D|\theta)P(\theta|\lambda)d\theta} \end{aligned} \tag{12}$$

In the case of a non-specific (zero derivative in the vicinity of the maximum of the posterior) hyper-parameter prior this suggest that we should search for the optimal parameters among the solutions to the equation

$$\langle \mathbf{f}(\theta)\rangle_{\mathrm{prior}} = \langle \mathbf{f}(\theta)\rangle_{\mathrm{post}}, \tag{13}$$

i.e., the hyper-parameters should be tuned so that the prior and posterior expectations of the regularization function are identical.

This approach is illustrated in Figure 1. We have defined a simple simulation problem with $d = 49, N = 50$. The *design matrix* $\mathbf{W}$ is a random matrix and is designed to have a given condition number. We show the difference between left and right sides of Eq. (13) as function of the regularization parameter. For reference we also show the error - computed as the mean square distance between the obtained solution and the 'true' solution. The minimum error is obtained close to $\lambda = 1.0$, which is also close to the point were Eq. (13) is obtained.

4. A LINEAR MODEL WITH ADAPTIVE 'RIDGE' REGULARIZATION

The linear regression model $\mathbf{y} = \mathbf{W}\mathbf{x} + \mathbf{n}$ appears in many contexts, with $\mathbf{n}$ an i.i.d. Gaussian noise process with variance parameter σ^2. In the regression model $\mathbf{W}$ is unknown while the data set consists of input-output samples ($\mathbf{x}$,$\mathbf{y}$). Collecting the parameters $\mathbf{W}$ in a d-dimensional vector θ, the log-likelihood takes on the quadratic form ($N = |D|$),

$$\log P(D|\theta) = -\frac{N}{2}(\theta - \theta_{\mathrm{ML}})^\top \mathbf{A}(\theta - \theta_{\mathrm{ML}}) + \mathrm{const} \tag{14}$$

where $\mathbf{A}$ is a $d \times d$ matrix and θ_{ML} is shorthand for the maximum likelihood parameters. Assuming a Gaussian prior on the parameters which is controlled by a single hyper-parameter λ,

$$P(\theta|\lambda) = \sqrt{\lambda/2}\exp -\frac{1}{2}\lambda\theta^2 \tag{15}$$

the regularization function is $f(\theta) = \frac{1}{2}\theta^2$, the whole model then corresponding to so-called 'ridge-regression'.

As both the likelihood and prior are Gaussians, the posterior also becomes a Gaussian, and the mean vector is $\theta(\lambda) = \mathbf{A}(\mathbf{A} + \lambda/N)^{-1}\theta_{\mathrm{ML}}$ while co-variance matrix easily is found to be $\mathbf{\Sigma}(\lambda) = 1/N(\mathbf{A} + \lambda/N)^{-1}$. Hence, the expectations are given by

$$\langle \frac{1}{2}\theta^2\rangle_{\mathrm{prior}} = \frac{d}{2\lambda} \tag{16}$$

$$\langle \frac{1}{2}\theta^2\rangle_{\mathrm{post}} = \frac{1}{2}\mu(\lambda)^2 + \frac{1}{2}\mathrm{Trace}\mathbf{\Sigma}(\lambda) \tag{17}$$

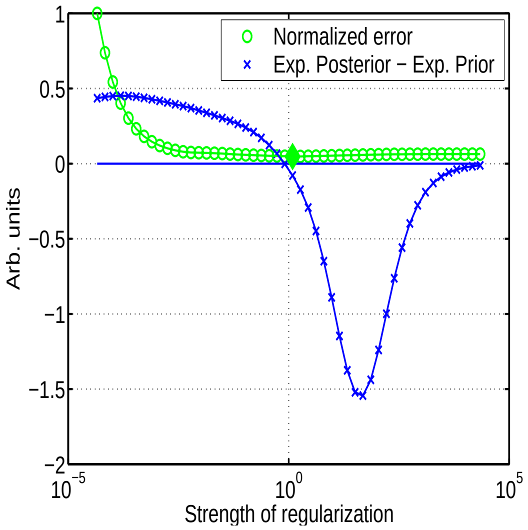

Figure 1: The difference between expectation of the regularization function in the prior and posterior distributions (difference between left and right side of Eq. (13)) is plotted versus the strength of the hyper-parameter (λ) for a simple linear model with Gaussian prior on the parameters ('ridge regression'). The quality of the estimator measured as the mean square difference between the regularized parameters and the 'true' parameters is plotted for reference. The mean square error is minimal at the ($\diamond$), approximately $\lambda = 1$, which is also close to the point were Eq. (13) is obtained.

4.1 Pruning from adaptive regularization

In [4] we discussed several scenarios for optimization of hyper-parameters, in particular, it was shown that for very low signal-to-noise ratios, the adaptive regularization leads to *pruning*, i.e., the optimal regularization strength is infinite. This can be illustrated in the present model by analyzing a the simple special case: Let the $\mathbf{A} = a\mathbf{1}$. In this situation the posterior expectation simplifies to $\langle \frac{1}{2}\theta^2 \rangle_{\text{post}} = \frac{1}{2}\frac{\theta^2_{\text{ML}} a^2}{(a+\lambda/N)^2} + \frac{1}{2N}\frac{d}{a+\lambda/N}$ and this in turn leads to the optimal hyper-parameter,

$$\lambda_{\text{opt}} = \frac{d}{\theta^2_{\text{ML}} - d/(aN)}, \tag{18}$$

hence, for $\theta^2_{\text{ML}} < d/(aN)$ the optimal regularization is infinite and the optimal parameters are pruned: $\theta_{\text{opt}} = \mathbf{0}$.

4.2 The conjugate prior for a general linear model

We now turn to the Bayesian general linear model $\mathbf{y} = \mathbf{W}\mathbf{x} + \mathbf{n}$, with $\mathbf{x}$ unknown. The so-called 'design matrix' $\mathbf{W}$ is a set of hypothesized effects of unknown strengths $\mathbf{x}$. Let the additive noise in the linear inverse problem be drawn i.i.d. from a zero mean normal distribution with (unknown) variance σ^2. For a given set of 'parameters' $\mathbf{x}, \sigma^2$ the likelihood function is given by,

$$P(y|\sigma^2, \mathbf{x}, \mathbf{W}) = \left(\frac{1}{2\pi\sigma^2}\right)^{N/2} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{W}\mathbf{x})^2\right). \tag{19}$$

Since, however, these parameters are unknown we will integrate them out using a *prior distribution* $P(\mathbf{x}, \sigma^2)$. Here we choose the conjugate prior which is defined to be a prior that produces a posterior of the same functional form, but with updated - data dependent - parameters,

$$\begin{aligned} P(\mathbf{y}) &= \int\!\!\int P(\mathbf{x}, \sigma^2) P(\mathbf{y}|\sigma^2, \mathbf{x}, \mathbf{W}) d\sigma^2 d\mathbf{x} \\ &= \int\!\!\int \frac{P(\mathbf{x}, \sigma^2)}{(2\pi\sigma^2)^{N/2}} \exp -\frac{(\mathbf{y} - \mathbf{W}\mathbf{x})^2}{2\sigma^2} d\sigma^2 d\mathbf{x}, \end{aligned} \tag{20}$$

see e.g., [16] .

The conjugate prior for the linear model with additive gaussian noise is the so-called *normal-inverse-gamma*, see [14], or NIG$(a, d, \mathbf{m}, \mathbf{V})$, distribution,

$$\begin{aligned} P(\mathbf{x}, \sigma^2) = \frac{(a/2)^{d/2}(\sigma^2)^{-(d+p+2)/2}}{(2\pi)^{p/2}|\mathbf{V}|^{1/2}\Gamma(d/2)} \times \\ \exp\left(-(\mathbf{x} - \mathbf{m})'(2\sigma^2\mathbf{V})^{-1}(\mathbf{x} - \mathbf{m}) - \frac{a}{2\sigma^2}\right). \end{aligned} \tag{21}$$

The hyper-parameters, collected in the vector $\lambda = \{d, a, \mathbf{m}, \mathbf{V}\}$, have the following meaning: The marginal prior distribution of $\mathbf{x}$, is a multivariate t-distribution with mean $\mathbf{m}$ and covariance $(a/(d-2))\mathbf{V}$. This distribution is unimodally centered at $\mathbf{m}$, with heavier 'tails' than a normal distribution. The marginal prior distribution of σ^2 is given by

$$P(\sigma^2|m, l) = \frac{(a/2)^{-d/2}(\sigma^2)^{-(d+2)/2}}{\Gamma(d/2)} \exp\left(-a/(2\sigma^2)\right). \tag{22}$$

Hence an inverse gamma distribution (meaning: $1/\sigma^2$ is gamma distributed) of mean $a/(d-2)$, $d > 2$.

Here we will simplify the prior by the following assumptions. First, the prior on $\mathbf{x}$ is zero mean, hence, $\mathbf{m} = 0$. Secondly, we will let the prior on the noise variance have finite variance, but otherwise vague, e.g., $d = 3$. Finally, we will let the covariance be proportional to the unit matrix $\mathbf{V} = v\mathbf{1}$. The parameter v plays a role similar to the regularizer λ above.

The relation (13) can be simplified for exponential family distributions where is possible to compute the normalization constant C. In the present case both the prior and the posterior are NIG-distributions, hence, we can normalize both prior and posterior to obtain $C_{\text{prior}}, C_{\text{post}}$. In this case the relation reads in terms of the vector of regularization parameters

$$\frac{\partial \log C_{\text{post}}(\lambda)}{\partial \lambda} = \frac{\partial \log C_{\text{prior}}(\lambda)}{\partial \lambda}, \tag{23}$$

which is simply equivalent to optimization of the 'evidence' [9],

$$\frac{\partial}{\partial \lambda} \log \frac{C_{\text{prior}}(\lambda)}{C_{\text{post}}(\lambda)} = \frac{\partial}{\partial \lambda} \log P(D|\lambda) = \mathbf{0}, \tag{24}$$

The evidence can be computed analytically using, with the following relations between the prior and posterior NIG parameters

$$\begin{aligned} \mathbf{V}_P^{-1} &= \mathbf{V}^{-1} + \mathbf{W}'\mathbf{W}, \\ \mathbf{m}_P &= \mathbf{V}_P \mathbf{W}'\mathbf{y}, \\ a_P &= a + \mathbf{y}'\mathbf{y} - \mathbf{m}'_P \mathbf{V}_P^{-1} \mathbf{m}_P, \\ d_P &= d + N. \end{aligned} \tag{25}$$

and reads,

$$P(D|a,\mathbf{V},d,\mathbf{W}) \propto \frac{|\mathbf{V}_P|^{1/2} a^{d/2} \Gamma(d_P/2)}{|\mathbf{V}|^{1/2} a_P^{d_P/2} \Gamma(d/2)} \qquad (26)$$

In Figure 2 we show the evidence as function of the two control parameters for the noise variance prior and for the prior on the solution to the linear inverse problem with the optimal Bayes parameters located in the vicinity of the point $v = 1, a = 1$.

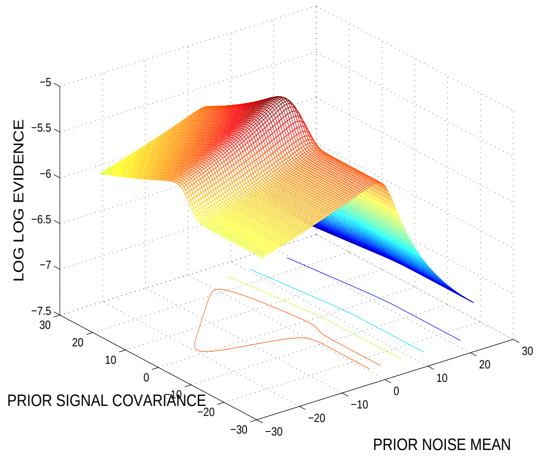

Figure 2: Evidence of a general linear model with the normal-inverse-gamma conjugate prior as function the two hyperparameters a and v controlling the noise variance and the parameters respectively. The evidence optimum is achieved in the vicinity of the point $v = 1, a = 1$.

4.3 Detecting the response to periodic stimulus in fMRI

Using functional magnetic resonance imaging (fMRI) is it possible to obtain a mapping from the visual field to the cortex. This mapping can be used for investigating the subject's placement visual field, e.g., as a spatial guide for interpretation of more complex visual stimuli. Specifically a visual field sign map is often obtained by either a polar mapping experiment (rotating wedge) and an eccentricity mapping experiment (expanding ring) [15], the two can be combined in one efficient measurement as shown recently in [10]. The idea is that every location in the visual field is activated in a periodic pattern with a certain phase which is a characteristic of given spatial locations. Thus we need an efficient detector of periodic signals for rather short signals. The use of the Bayes general linear model for this task was first presented in [6]. We here illustrate the results of using the general linear model with the NIG prior for detecting model orders and reconstructing the detected signal using its phase to link visual field and cortex.

Data: Using a 3T MRI scanner (Magnetom Trio, Siemens, Erlangen, Germany) 528 GRE EPI volumes were acquired. The functional volumes consisted of 20 slices with 3 mm thickness, oriented along the calcarine sulcus, TR=1.2s, FOV=192, 64x64 matrix, flip angle = 67^o. The visual stimulation consisted of a wedge rotating either clockwise (CW) or counter clockwise (CCW) at two different cycle times (time for a full rotation) on a grey background (as shown to the left). Simultaneously an expanding/contracting ring was shown also at two different cycle rates (time for one full expansion/contraction). Within both the ring and wedge a black-white checkerboard flickered at a reversal rate of 8 Hz, both stimuli covered a maximum of 18.4^o of the subject's visual field. The stimulation can be summarized as follows:

- 90s CCW wedge cycle rate of 25 s and expanding ring cycle rate of 30s (1 and 2)
- 90s CCW wedge cycle rate of 30 s and expanding ring cycle rate of 25s (3 and 4)
- 25s pause with fixation point only
- 90s CW wedge cycle rate of 25 s and contracting ring cycle rate of 30s (5 and 6)
- 90s CW wedge cycle rate of 30 s and contracting ring cycle rate of 25s (7 and 8)

The change in cycle rate assured that the same number of cycles where completed for both the ring and the wedge. The response was modelled with sine and cosine predictors (1. and 2. order harmonics) with the frequency corresponding to each of the cycle rates.

Analysis A 3D rigid body transformation was used to correct for motion. The effects of interest were modelled with harmonics of the stimulation cycle rate as shown in Figure 3. To determine the phase of the activation the time to peak was found from the reconstructed signal (using the maximum a posteriori parameter estimates) for each of the 8 different activation types. The hemodynamic lag was determined and accounted for by comparing the phase of stimuli running in opposite directions with the same frequency.

Furthermore we have used the evidence for detecting the cycle time combinations of the two periodic stimuli. In Figure 4 we show the evidence of different models (different $\mathbf{W}$'s) in a region of interest (delignated in the visual cortex). Different models were constructed assuming the cycle time combinations indicated on the axes. The correct combination 25s and 30s are identified as the most probable.

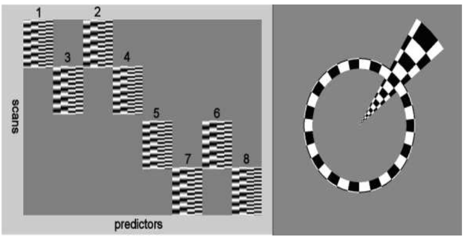

Figure 3: Left: Design matrix $\mathbf{W}$ for linear modeling of the response to simultaneous periodic visual stimulus. Right: Snapshot of the visual stimulus.

5. DISCUSSION

Adaptive regularization - aka 'empirical Bayes' - can be a useful tool for solving noisy linear inverse problems, see e.g., [11]. While the issue here has been to discuss some generic aspects of adaptive regularization, such as the possibility of pruning and the fact that the optimal hyperparameters solve a simple equation (13) we also note that this method is only one among several existing methods for controlling the amount of regularization. The so-called L-curve

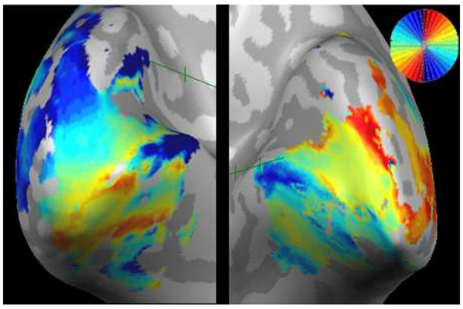

Figure 4: Mapping from visual field to primary visual cortices: Phase maps of the activation shown on an inflated version of the right and left hemispheres. The legend in the top left corner indicates corresponding maps in the visual field.

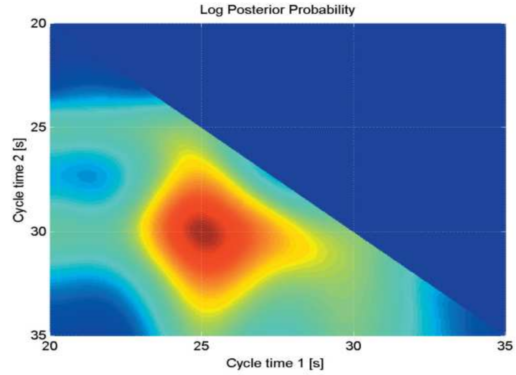

Figure 5: Detecting cycle time combinations. The figure show the mean posterior probabilities of different models in a region of interest (in the visual cortex). Different models were constructed assuming the cycle time combinations indicated on the axes. The correct combination 25s and 30s is identified as the most probable. The probabilities are shown in log colour scale.

method is based on a parametric plot of the regularized solution norm $||\mathbf{L}\mathbf{x}_\lambda||_2$ and the residual norm $||\mathbf{A}\mathbf{x}_\lambda - \mathbf{b}||$. Both of these are parameterized by λ. The optimal regularization occur when both norms are small as mentioned in [2] and also discussed in [8]. The L-curve method proposes to balance the terms by finding the location of maximum positive curvature in the parametric plot, at this choice of regularization the solution changes nature from being dominated by regularization errors (over smoothing) to being dominated by errors in the right hand side. The regularization parameter can also be estimated using so-called generalized cross validation (GCV) [1]. The basic idea is: if any data point y_i is left out and a solution x_i is computed to the reduced problem of one dimension less, then the estimate of y_i computed from x_i must be a good estimate. Methods for estimation of regularization have been compared in [17]. There is no clear consensus about the relative merits of these approaches, so a principled statistical approach like the Bayesian approach suggested here may helpful for understanding the properties of regularization functions and hyperparameter tuning.

Acknowledgments

We thank an anonymous reviewer for useful pointers to early literature. This work is supported by the Danish Technical Research Council, through the framework project 'Intelligent Sound', www.intelligentsound.org (STVF No. 26-04-0092).

REFERENCES

[1] G. Golub, M. Heath and G. Wahba, *Generalized cross-validation as a method for choosing a good ridge parameter*, Technometrics 21:215–223, (1979).

[2] P.C. Hansen. *Analysis of discrete ill-posed problems by means of the L-curve.* SIAM Review, 34(4):561-580, (1992).

[3] L.K. Hansen: *Stochastic Linear Learning: Exact Test and Training Error Averages*. Neural Networks 6:393-396, (1993)

[4] L.K. Hansen and C.E. Rasmussen: *Pruning from Adaptive Regularization*. Neural Computation 6:1223-1232 (1994).

[5] L.K. Hansen: *Bayesian Averaging is Well-Temperated* In S.S. Solla et al. (eds.) Proceedings of NIPS∗99, Denver, November 29 - December 4, 1999, 265-271 (2000).

[6] L.K. Hansen, F.Å. Nielsen, and J. Larsen. *Exploring fMRI data for periodic components*. Artificial Intelligence in Medicine, 25(1): 35-44, (2002).

[7] D. Haussler and M. Opper: *Mutual Information, Metric Entropy, and Cumulative Relative Entropy Risk* Annals of Statistics 25:2451-2492 (1997)

[8] C.L. Lawson and R.J. Hanson. *Solving Least Squares Problems*. Prentice-Hall, Englewood Cliffs, New Jersey, 1974.

[9] D.J.C MacKay: *Bayesian Interpolation*, Neural Computation 4:415-447, (1992).

[10] K.H. Madsen and T.E. Lund: *Simultaneous acquisition of polar and eccentricity mappings of the human visual cortex using fMRI*. In proceedings of the 13th ISMRM, (2005).

[11] R. Molina, A.K. Katsaggelos, J. Mateos: *Bayesian and regularization methods for hyperparameter estimation in image restoration*. IEEE Transactions on Image Processing, 8(2):231–246, (1999).

[12] J. Moody: "Note on Generalization, Regularization, and Architecture Selection in Nonlinear Learning Systems," in B.H. Juang, S.Y. Kung & C.A. Kamm (eds.) *Proceedings of the first IEEE Workshop on Neural Networks for Signal Processing*, Piscataway, New Jersey: IEEE, 1–10, (1991).

[13] N. Murata, S. Yoshizawa & S. Amari: *Network Information Criterion — Determining the Number of Hidden Units for an Artificial Neural Network Model.* IEEE Transactions on Neural Networks, 5(6):865–872, (1994).

[14] A. Ohagan: *Bayesian Inference*. Kendall's Advanced Theory of Statistics. Vol 2B. The University Press, Cambridge (1994).

[15] M.I. Sereno, A.M. Dale, J.B. Reppas, K.K. Kwong, J.W. Belliveau, T.J. Brady, B.R. Rosen, R.B. Tootell *Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging.* Science 268:889–893, (1995)

[16] C.P. Robert: *The Bayesian Choice - A Decision-Theoretic Motivation*. Springer Texts in Statistics, Springer Verlag, New York (1994).

[17] A.M. Thompson, J.C. Brown, J.W. Kay, D.M. Titterington, *A study of methods of choosing the smoothing parameter in image restoration by regularization*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 13(4):326 - 339 (1991).

[18] H. White, "Consequences and Detection of Misspecified Nonlinear Regression Models," *Journal of the American Statistical Association*, 76(374):419–433, (1981).