Improved classification images with sparse priors in a smooth basis

Patrick J. Mineault Montreal Neurological Institute, McGill University, Montreal, QC, Canada

Simon Barthelmé Laboratoire de Psychologie de la Perception, Université Paris Descartes, Paris, France

Christopher C. Pack Montreal Neurological Institute, McGill University, Montreal, QC, Canada

Classification images provide compelling insight into the strategies used by observers in psychophysical tasks. However, because of the high-dimensional nature of classification images and the limited quantity of trials that can practically be performed, classification images are often too noisy to be useful unless denoising strategies are adopted. Here we propose a method of estimating classification images by the use of sparse priors in smooth bases and generalized linear models (GLMs). Sparse priors in a smooth basis are used to impose assumptions about the simplicity of observers' internal templates, and they naturally generalize commonly used methods such as smoothing and thresholding. The use of GLMs in this context provides a number of advantages over classic estimation techniques, including the possibility of using stimuli with non-Gaussian statistics, such as natural textures. Using simulations, we show that our method recovers classification images that are typically less noisy and more accurate for a smaller number of trials than previously published techniques. Finally, we have verified the efficiency and accuracy of our approach with psychophysical data from a human observer.

Keywords: classification image, detection/discrimination, linear template, generalized linear model, sparse prior

Citation: Mineault, P. J., Barthelmé, S., & Pack, C. C. (2009). Improved classification images with sparse priors in a smooth basis. *Journal of Vision, 9*(10):17, 1–24, http://journalofvision.org/9/10/17/, doi:10.1167/9.10.17.

Introduction

In recent years, the classification image approach has emerged as a powerful method of probing observers' strategies during psychophysical tasks. In a typical experiment, the observer is asked to indicate the presence or absence of a target signal masked with additive noise (Figure 1A). The resulting data are then evaluated under the assumption that the observer performs the task by linearly correlating the signal with an internal template, responding "target present" when the result exceeds a criterion, and "target absent" otherwise (Figure 1B). In this context, one can obtain an estimate of the observer's internal template by correlating the responses and the noise fields. The resulting *classification image* is a visual representation of the observer's strategy in the task.

The key advantage of the classification image approach over other psychophysical measures is that it is theoretically applicable when little is known about the stimulus parameters that are relevant in performing a task. The procedure is in this sense analogous to the reverse correlation technique commonly used in neurophysiology (Simoncelli, Pillow, Paninski, & Schwartz, 2004), and the results of classification image experiments can, in appropriate circumstances, be compared meaningfully to data obtained from single-unit recordings (Neri & Levi, 2006). Consequently, the approach has been used widely in psychophysics, including in studies of Vernier acuity (Ahumada, 1996), disparity processing (Neri, Parker, & Blakemore, 1999), motion perception (Neri & Levi, 2008), object discrimination (Olman & Kersten, 2004), and face recognition (Sekuler, Gaspar, Gold, & Bennett, 2004).

Overcoming noise in classification images with prior assumptions

Despite the power and flexibility of the approach, the utility of classification images is limited by the amount of data that can be obtained in a given experimental task. For example, in tasks involving two spatial dimensions as well as time, even a modest stimulus resolution of 16 × 16 pixels and 16 time steps requires the estimation of over 4,000 parameters, which may require each observer to perform tens of thousands of trials. In practice, there is rarely sufficient data for the number of degrees of freedom the experimenter wishes to probe, and as a result classification images can be quite noisy. Such noise limits the interpretability and usefulness of the classification image.

doi: 10.1167/9.10.17 Received March 20, 2009; published September 23, 2009 ISSN 1534-7362

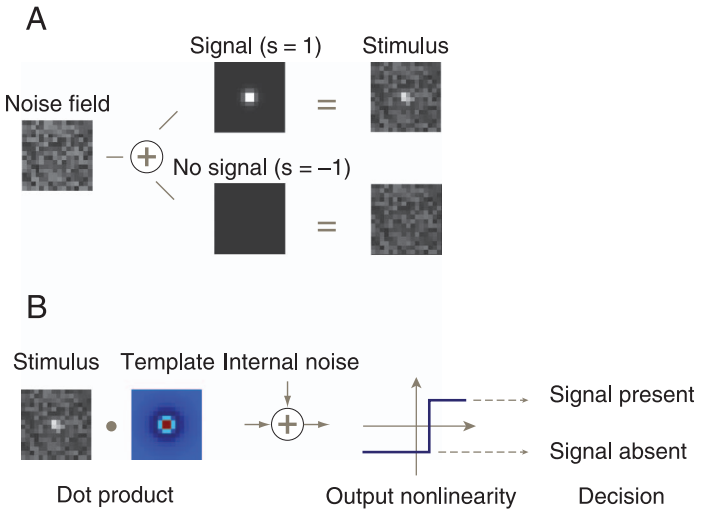

Figure 1. Outline of the classification image paradigm and the linear observer model. (A) The classification image paradigm. On every trial, the observer is presented with a stimulus which is either the sum of a randomly generated noise field and the signal or just a noise field. The observer's task is to indicate whether the stimulus was present or absent on a trial. (B) The linear observer model. The observer performs the task in panel A by correlating the stimulus with an internal template, giving a number which is corrupted by additive internal noise. The observer responds "signal present" when the number exceeds a criterion, and "signal absent" otherwise.

One solution to this problem is to supplement observer data with prior information about the observer's strategy. Such prior information reduces the effective number of free parameters that must be estimated, leading to better estimates, assuming that the internal template conforms to the constraints. This approach has long been used implicitly in the form of post hoc smoothing and thresholding, under the assumptions that the observer's internal template is smooth and sparse, respectively. Explicit prior assumptions have recently been used successfully in classification image estimation (Knoblauch & Maloney, 2008a; Knoblauch & Maloney, 2008b; Ross & Cohen, 2009).

At first glance, the use of prior information may seem to be in contradiction with classification images' stated advantage of being applicable when little is known about the visual process being probed. However, while for a given task we may know little about the *specifics* of the visual process involved, we often have *general* exploitable knowledge derived from previous classification image experiments and from neurophysiology. To give but one example (Knoblauch & Maloney, 2008b; Thomas & Knoblauch, 2005), while we may not know exactly how a human observer detects a time-modulated luminance signal in noise, we do know that humans have limited sensitivity to high temporal frequencies. This suggests that observers' internal templates will be smooth at a certain scale for such a task. This prior knowledge of smoothness can then be used to obtain more accurate classification images on this particular task (Knoblauch & Maloney, 2008b).

Our goal in this paper is to define and explore the consequences of incorporating a powerful class of prior assumptions to estimate classification images. One approach to finding good prior assumptions is to attempt to find regularity within a set of observations, an approach that has been used with great success in defining state-of-the-art image denoising techniques (Srivastava, Lee, Simoncelli, & Zhu, 2003; Portilla, Strela, Wainwright, & Simoncelli, 2003). An informal review of various articles (Abbey & Eckstein, 2002; Ahumada, 1996; Chauvin, Worsley, Schyns, Arguin, & Gosselin, 2005; Knoblauch & Maloney, 2008b; Levi & Klein, 2002; Mangini & Biederman, 2004; Neri & Levi, 2006, 2008; Neri et al., 1999; Sekuler et al., 2004; Tadin, Lappin, & Blake, 2006) that make use of the classification image technique shows that published classification images, regardless of exact protocol used, share a certain similarity: They appear to be well described by a small number of smooth features, such as lines and Gaussian blobs. In other words, many internal templates can be well described by a sparse sum of smooth basis functions, such as Gaussian blobs.

A second approach to finding good assumptions is to consider facts about the process *underlying* a set of observations. In the context of classification images, this means incorporating knowledge of visual physiology. The human visual system is trained on visual images which themselves are sparse in a basis of smooth, oriented filters (Srivastava et al., 2003). Conversely, a simple constraint of sparse representation of images is sufficient to reproduce many properties of the visual system, including V1 receptive fields (Olshausen & Field, 1996) and color opponency (Lee, Wachtler, & Sejnowski, 2002). It is thus tempting to conjecture that humans are naturally more efficient at representing sparse visual structure (Olshausen & Field, 2004), and that this induces sparse internal templates in classification image experiments.

A third, more pragmatic approach to finding good prior assumptions is to combine and extend proven prior assumptions in a synergistic manner. Smoothing and thresholding have proven useful in analyzing data from classification image experiments (Chauvin et al., 2005; Gold, Murray, Bennett, & Sekuler, 2000; Knoblauch & Kenneth, 2008; Mangini & Biederman, 2004; Rajashekar, Bovik, & Cormack, 2006; Tadin et al., 2006), and combining and extending the two might yield a more effective estimation technique. Sparseness in a smooth basis is a natural generalization of assumptions of smoothness and sparseness that has the potential to yield a strong class of prior assumptions.

In light of the ideas mentioned above, we propose imposing sparseness in a basis of smooth functions as a way of increasing the accuracy and efficiency with which classification images are estimated. We impose this assumption in a framework that can naturally accommodate prior information.

Imposing basis sparseness in generalized linear models

Our analytical framework builds upon generalized linear models (GLMs), which have been used previously to estimate classification images (Abbey & Eckstein, 2001; Knoblauch & Maloney, 2008a; Knoblauch & Maloney, 2008b; Solomon, 2002) and neuronal receptive fields (Wu, David, & Gallant, 2006). As with the linear observer model typically used in classification image experiments (Ahumada, 2002), GLMs assume that the output of a system is generated by first linearly correlating the input with a template and that the internal response is then transduced to an observed response by a fixed stochastic process (Figure 1B). GLMs have a number of desirable properties: they provide a unifying framework for estimating classification images, neuronal receptive fields, and functional imaging (Victor, 2005); they can be fit efficiently by maximum likelihood (ML) methods (Wu et al., 2006); they work with arbitrary stimuli; and extensions to the basic model can incorporate important input or output nonlinearities (Ahrens, Paninski, & Sahani, 2008). Most importantly here, constraints on classification images can naturally be imposed in the GLM context by the use of a prior, which assigns probabilities to different parameter values. For example, spatial smoothness is imposed by assuming that the spatial derivatives of the template follow a Gaussian distribution of a given width (Knoblauch & Maloney, 2008b).

The assumption of sparseness in a particular basis translates naturally into a prior distribution which gives higher probability to models with a small number of nonzero basis coefficients. The Laplace distribution, which is heavily peaked around zero, is the sparsest distribution for which the GLM estimation problem is tractable (Seeger, 2008). We thus propose to estimate classification images through GLMs by imposing such a sparseness-inducing prior on basis coefficients. This strategy provides a realistic and parsimonious account of the observer's strategy during a variety of psychophysical tasks.

We show by simulations and experiments with a real observer that classification images estimated with a sparse prior in a basis are less noisy and take less trials to converge than those estimated by other methods used in the literature. In addition, the sparse prior discards coefficients which do not contribute significantly to an observer's decision process, leading to classification images that are highly interpretable and, under appropriate circumstances, readily comparable to data obtained from single-unit recordings. All data sets and Matlab software used in this article are freely available at our Web site (http://apps.mni.mcgill.ca/research/cpack/sparseglm.zip) under the General Public License (GPL). A preliminary version of this work was presented previously (Mineault & Pack, 2008).

Methods: Statistical estimation of internal templates

The linear observer model

Consider a task in which an observer must report the presence or absence of a target. On each trial, the signal may or may not be present, and in all cases, the target or absence thereof is masked by a noise field. As in previous work (Abbey & Eckstein, 2002; Ahumada, 2002; Knoblauch & Maloney, 2008b; Murray, Bennett, & Sekuler, 2002), we assume that the observer performs the detection task by correlating the stimulus with an internal template. Trials in which the stimulus is similar to the template lead the observer to report the presence of the target. Thus, the stimulus is represented by a real vector $\mathbf{x}$ of dimension k, the template by a vector $\mathbf{w}$, the internal noise by ϵ, and the observer computes an internal decision variable, v, by

$$v = \mathbf{x}^{\mathrm{T}}\mathbf{w} + \epsilon. \quad (1)$$

Following each stimulus presentation, the observer gives a binary response $y = \pm 1$ according to whether the internal variable is larger than a criterion or offset c. The response y is thus given by

$$y = \mathrm{sign}(v - c). \quad (2)$$

We assume that the observer's internal noise ϵ is taken from a symmetric distribution with mean 0 and standard deviation σ, so that $\epsilon \sim \Phi'(0, \sigma^2)$, where Φ is the cumulative distribution function (cdf) for the distribution in question. The internal noise represents observers' inconsistency: the same physical stimulus can elicit different responses. Other formulations (Ahumada, 2002) assume that it is the threshold c that is a random variable; our formulation is equivalent.

Finding estimates for the model parameters

We use statistical inference to find the most probable internal template $\mathbf{w}$, given the data $\mathbf{y}$. From Bayes' theorem, the posterior probability distribution of the parameters $\mathbf{w}$ is obtained from

$$p(\mathbf{w}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{w})p(\mathbf{w}). \quad (3)$$

This equation captures the relationship between the *posterior* $p(\mathbf{w}\,|\,\mathbf{y})$, the *likelihood* $p(\mathbf{y}\,|\,\mathbf{w})$, and the *prior* $p(\mathbf{w})$. Our goal is to find the value of $\mathbf{w}$ that maximizes the posterior. For numerical reasons, it turns out to be simpler

to find the value of $\mathbf{w}$ that minimizes the negative log of the posterior:

$$\begin{aligned}\mathbf{w}_{\mathrm{ML/MAP}} &= \underset{\mathbf{w}}{\operatorname{argmin}}\{-\log p(\mathbf{y}|\mathbf{w}) - \log p(\mathbf{w})\} \\ &= \underset{\mathbf{w}}{\operatorname{argmin}}\{L(\mathbf{y}, \mathbf{w}) + R(\mathbf{w})\}. \end{aligned} \quad (4)$$

When the prior is flat, we obtain the *maximum likelihood* (ML) estimate of $\mathbf{w}$; otherwise, the estimate is called *maximum a posteriori* (MAP). The negative log-likelihood is denoted by L, while the negative log-prior, sometimes called the *regularizer,* is denoted by R.

Likelihood function for the linear observer Model

For a single stimulus $\mathbf{x}$, the probability of observing response $y = +1$ given $\mathbf{w}$ and offset c is

$$\begin{aligned} p(y = +1|\mathbf{x}, \mathbf{w}, c) &= \int_c^{\infty} \Phi'(z - \mathbf{x}^{\mathrm{T}}\mathbf{w}, \sigma^2)\, dz \\ &= 1 - \Phi(c - \mathbf{x}^{\mathrm{T}}\mathbf{w}, \sigma^2) \\ &= \Phi(\sigma^{-1}(\mathbf{x}^{\mathrm{T}}\mathbf{w} - c)). \end{aligned} \quad (5)$$

As probabilities sum to 1, $p(y = -1 \,|\, \mathbf{x}, \mathbf{w}, c) = 1 - p(y = +1 \,|\, \mathbf{x}, \mathbf{w}, c)$.

The experimenter's goal is to estimate the observer's internal template $\mathbf{w}$, which is assumed to be constant throughout the duration of the experiment. We therefore wish to find a template that captures the relationship between a series of stimuli $\mathbf{X} = [\mathbf{x}^{\{1\}}, \mathbf{x}^{\{2\}}, \ldots, \mathbf{x}^{\{n\}}]$ and the corresponding responses $\mathbf{y} = (y_1, y_2, \ldots, y_n)$. We make the standard assumption that an observer's response on a given trial is independent of his or her responses on other trials. The likelihood function is then

$$\begin{aligned} p(\mathbf{y}|\mathbf{X}, \mathbf{w}, c) &= \prod_{i=1}^{n} p(y_i|\mathbf{x}, \mathbf{w}, c) \\ &= \prod_{i=1}^{n} \Phi(\sigma^{-1} y_i((\mathbf{x}^{\{i\}})^{\mathrm{T}}\mathbf{w} - c). \end{aligned} \quad (6)$$

There is an ambiguity in this formulation, as for any given value of σ it is possible to multiply c and $\mathbf{w}$ by a constant factor such that the likelihood stays the same. We resolve this ambiguity by using the standard procedure of setting $\sigma = 1$; noisier observers are accommodated by a smaller overall magnitude for the weight vector. The negative log-likelihood is therefore

$$-\log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, c) = -\sum_{i=1}^{n} \log\, \Phi(y_i(\mathbf{X}\mathbf{w} - c)_i). \quad (7)$$

We make this formulation slightly more general by allowing for a set of auxiliary variables $\mathbf{U}$ with corresponding weights $\mathbf{u}$. Later, we will impose a prior on $\mathbf{w}$ but not on $\mathbf{u}$. This formulation allows for the inclusion of factors that affect the observer's responses other than the noise stimulus. For example, in Tadin et al. (2006), the question of interest is the time-varying influence of visual motion in the stimulus surround on judgments of motion direction in the center. Here, the time-varying surround stimulus would be put into the $\mathbf{X}$ matrix while the signal sign in the center would take one column of the $\mathbf{U}$ matrix. Another possibility, in detection tasks, is to include the signal sign into the $\mathbf{U}$ matrix, as in Knoblauch and Maloney (2008b, p.5, Equation 16), rather than summing noise and signal in the $\mathbf{X}$ matrix. We found that the $\mathbf{U}$ strategy generally led to faster optimization while yielding equivalent estimates of internal templates. A final possibility is to include trial-delayed responses in the $\mathbf{U}$ matrix to model observer's tendency to respond similarly or dissimilarly on subsequent trials, as in the commonly used method of modelling refractory periods and burstiness in neurons (Pillow et al., 2008). We eliminate the constant c by adding a constant column to the matrix $\mathbf{U}$. Under this new parameterization, we have the negative log-likelihood function for the linear observer model:

$$L = -\sum_{i=1}^{n} \log\, \Phi(y_i(\mathbf{X}\mathbf{w} + \mathbf{U}\mathbf{u})_i). \quad (8)$$

The above expression is known in the statistics literature as the negative log-likelihood for a binomial *generalized linear model* (GLM) (Gelman, Carlin, Stern, & Rubin, 2003). Assuming that the internal noise has a Gaussian distribution, Φ becomes the cdf of a Gaussian, and we obtain the probit model (Knoblauch & Maloney, 2008b). If instead we assume that the internal noise has a logistic distribution, $\Phi(x) = 1 / (1 + \exp(-x))$, we obtain the logit or logistic regression model. In practice, the Gaussian and logistic distributions are very similar, and empirically it is difficult if not impossible to determine which provides a more accurate description of an observer's internal noise. For computational simplicity, we adopt the logistic regression model.

Methods: Regularization of the solution

In a classification image experiment, we typically have substantial prior expectations about the internal template $\mathbf{w}$, as we have argued in the introduction. Such prior information can effectively narrow the space of possible classification images, thus improving the efficiency with which classification images can be recovered, as well as

Gaussian prior name	Regularizer $R = -\log p(\mathbf{w})$
Weight decay (ridge)	$\frac{\lambda}{2}\mathbf{w}^T\mathbf{I}\mathbf{w}$
Smoothness	$\frac{\lambda}{2}\mathbf{w}^T(\mathbf{D}^T\mathbf{D})\mathbf{w}$
Spline smoothness (Knoblauch & Maloney, 2008b)	$\frac{\lambda}{2}\mathbf{w}^T\mathbf{S}\mathbf{w}$
Arbitrary	$\frac{\lambda}{2}\mathbf{w}^T(\mathbf{A}^T\mathbf{A})\mathbf{w}$

Table 1. Gaussian priors in the context of classification images and corresponding regularizers.

their accuracy and the interpretability of the results. Here we propose the use of sparse priors on smooth basis coefficients, which together impose global sparseness and local smoothness on the recovered templates.

We briefly discuss Gaussian priors for comparison. These assume that linear combinations of weights have Gaussian distributions. This can be used to impose the condition that weights are not too large, or that they vary smoothly across space. They have been discussed in detail in the context of neurophysiology in Wu et al. (2006) and are an integral component of the spline smoothing framework proposed by Knoblauch and Maloney (2008b).

Gaussian priors

Many priors arise from the assumption that linear transformations of weights have Gaussian distributions, $p(\mathbf{w}) = \mathcal{N}(\mathbf{w};0, (\lambda\mathbf{A}^\mathbf{T}\mathbf{A})^{-1})$, where λ is a scale factor. Common cases of Gaussian priors are shown in Table 1. When the matrix $\mathbf{A}$ is the identity matrix, we get a penalty on the magnitude of coefficients known as weight decay. When $\mathbf{A}$ is a spatial derivative operator, we get the smoothness prior, which encourages spatially smooth weights. In general, any choice of $\mathbf{A}$ for which $\mathbf{A}^T\mathbf{A}$ is positive definite generates a proper Gaussian prior. Choices other than weight decay and smoothness priors are discussed in some detail in Wu et al. (2006).

Sparse priors

We introduce a constraint of sparseness by way of a prior distribution over the weights $\mathbf{w}$. If weights are sparse, then most are zero and few have large values. The corresponding prior distribution should have a peak at 0 and heavy tails. A simple prior embodying these assumptions is that of an independent sparse prior over the weights:

$$\begin{aligned} p(w_i|\lambda) &\propto \exp(-\lambda|w_i|) \\ p(\mathbf{w}|\lambda) &\propto \exp(-\lambda\textstyle\sum_i|w_i|) = \exp(-\lambda\|\mathbf{w}\|_1). \end{aligned} \tag{9}$$

Here $\|\mathbf{w}\|_1$ denotes the L_1 norm of the vector $\mathbf{w}$. The distribution $\exp(-\lambda\|\mathbf{w}\|_1)$ is known as the Laplace distribution, a term we avoid because of the potential confusion with the unrelated Laplacian pyramid which we introduce later; in the following we will refer to the prior induced by this distribution as the "sparse prior." This distribution has a sharp peak at the origin and heavy tails. In the context of Equation 3, the multiplication of this prior with the likelihood $p(y|w)$ has predictable effects on the resulting posterior density. For instance, if we assume a Gaussian likelihood (Figure 2), the resulting posterior density has a peak at the origin when the Gaussian's center is sufficiently close to 0. Hence, parameters whose presence gives only a marginal increase in the likelihood of the data will tend to be clamped at 0 and thus discarded from the model. L_1 regularization in the standard ("pixel") basis thus gives results similar to post hoc thresholding.

Of priors of the form $p(\mathbf{w}) = \exp(-\lambda\sum_i|w_i|^q)$, which includes both Gaussian and Laplace distribution priors, $p(\mathbf{w})$ is log-concave for $q \geq 1$, which leads to a convex, tractable optimization problems when estimating $\mathbf{w}$. As sparseness increases for small q, $q = 1$ yields the sparsest distribution of this form which leads to a tractable optimization problem (Seeger, 2008), which makes it generally preferable to alternative sparseness-inducing priors.

Reformulating the linear observer in terms of basis coefficients

A given model may require many weights to represent an internal template in the pixel basis yet be sparse in some other basis. We would thus like to reparameterize our problem to express our assumption that weights are sparse in an arbitrary basis $\mathbf{B}$. Denoting the basis weights as $\tilde{\mathbf{w}}$, we re-express the negative log-likelihood (Equation 7) as

$$L_\mathrm{B} = -\sum_{i=1}^{n}\log\Phi\big(y_i(\mathbf{X}\mathbf{B}\tilde{\mathbf{w}}+\mathbf{U}\mathbf{u})_i\big). \tag{10}$$

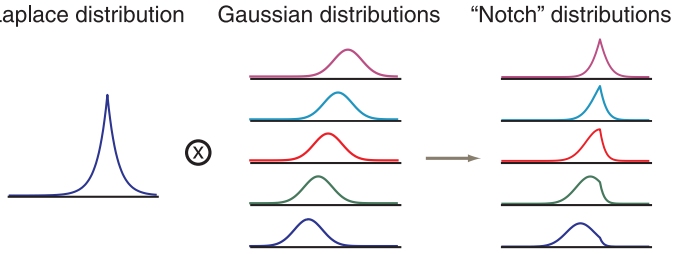

Figure 2. Effect of sparse prior on weights. When a Laplace distribution centered at 0 is multiplied by Gaussian likelihoods with varying centers, the corresponding posteriors have discontinuities in their first derivatives at 0 ("notches"). For a wide range of Gaussian centers, the MAP estimate is exactly 0, as seen in the top three posterior distributions.

The L_1 norm is not conserved under a rescaling of the columns of the basis matrix **B**, which means that a sparse prior will consider certain basis functions more likely if **B** is scaled unevenly. To avoid this, the matrix **B** should be normalized so that its component column vectors have norm 1. The associated MAP estimate is given by the vector $\tilde{\mathbf{w}}$ which minimizes:

$$L_B + \lambda \|\tilde{\mathbf{w}}\|_1. \quad (11)$$

The sparse prior is now imposed on the basis coefficients rather than on the classification image coefficients. The internal template can then be visualized by projection of the weights onto pixel space, $\mathbf{w} = \mathbf{B}\tilde{\mathbf{w}}$.

Choice of basis

We use the term *basis* loosely, as the rows of **B** need not span $\mathbf{R}^k$, as with the spline basis used in Knoblauch and Maloney (2008b), nor do they need to be linearly independent, as with overcomplete transforms. We can freely construct a basis matrix that embeds our assumptions for the particular classification image reconstruction problem at hand. In the case where **B** is overcomplete, that is, its columns span $\mathbf{R}^k$ but are not linearly independent, there are many equivalent ways of expressing the classification image. The sparse prior will tend to select the simplest way of expressing the classification image. The compatibility of sparse priors with overcomplete bases is advantageous as it is generally simpler to construct bases which have desirable properties when one removes the restriction of linear independence.

An assumption of smoothness can be embedded implicitly into the choice of smooth basis functions. In this regard, Gaussian basis functions are a natural choice, but for the analysis of real classification images it is desirable to have a basis that allows for the degree of smoothness to vary over space. Because internal templates may vary among observers and among tasks an ideal basis would not require strong *a priori* assumptions about the degree of smoothness in the classification image. Both criteria can be met by the use of a Laplacian pyramid (Burt & Adelson, 1983), which consists of multiple Gaussian functions that are smooth on various spatial scales.

In constructing a Laplacian pyramid, one typically chooses Gaussian functions with widths that are powers of 2. Each set of Gaussian basis functions that share the same width is known as a level, and within a level, the spatial separation of the basis functions is proportional to the width of the Gaussians. The power of two scheme means that in m dimensions, the decomposition is overcomplete by a factor $2^m / (2^m - 1) \leq 2$, i.e., only mildly overcomplete. The Laplacian pyramid can be extended by using more basis functions than is standard, for example by adding half-levels, or having more basis functions than standard within a level. Such undecimated Laplacian pyramids lead to greater computational cost during model fitting but can be more powerful than standard Laplacian pyramids.

Other decompositions based on using oriented basis functions at different resolutions include the steerable pyramid transform (Simoncelli & Freeman, 1995), several families of overcomplete wavelets (Selesnick, Baraniuk, & Kingsbury, 2005), and Gabor pyramids. These correspond to an assumption of sparse, smooth, oriented structure. Finally, transformations such as the discrete cosine transform (DCT) and the Fourier transform (Ahumada & Lovell, 1971; Levi & Klein, 2002) may be used to analyze the contributions of different spatial frequencies to performance on the task.

Methods: Hyperparameter selection

Using a Gaussian or sparse prior involves the specification of a free hyperparameter λ, much like smoothing involves the choice of the width of the smoothing kernel. This hyperparameter may be selected by assessing how well a model generalizes to untrained examples using k-fold cross-validation (Wu et al., 2006). Generalization performance is naturally assessed using the cross-validated log-likelihood of the data.

The k-fold cross-validation increases the computational burden of estimating a model by roughly km, where m is the number of regularization hyperparameters considered. In typical use, km is on the order of 50–100. This limits the practical applicability of this form of cross-validation to models which take a minute or so to fit, such as models with Gaussian priors.

Many methods for fitting a model with a sparse prior work by successively finding minima of subproblems of the form:

$$\underset{\mathbf{w}}{\operatorname{argmin}}\, f(\mathbf{w}) + \lambda_i \|\mathbf{w}\|_1. \quad (12)$$

Here λ_i is a decreasing sequence of regularization hyperparameters, and $f(\mathbf{w})$ is some convex function of **w**. These algorithms therefore generate an entire regularization path for roughly the same cost as fitting the model with the smallest λ considered. The cost of k-fold cross-validation is then roughly $k + 1 \ll km$ times the cost of fitting a single model, which makes cross-validation with sparse priors feasible. We chose the fixed-point continuation algorithm (Hale, Yin, & Zhang, 2007) to estimate model parameters with sparse priors. The algorithm is outlined in Appendix A.

Methods: Simulations

We simulated a linear observer on several detection tasks. The response y on the ith trial was given by

$$\begin{aligned} v_i &= (\mathbf{X}_i + \mathbf{S}_i)\mathbf{w} + c \\ y_i &= \text{sign}(v_i + \epsilon_i). \end{aligned} \tag{13}$$

The stimulus was composed of a signal $\mathbf{S}_i$ combined additively with a noise stimulus $\mathbf{X}_i$, which was chosen from the Gaussian distribution. The templates $\mathbf{w}$ varied depending on the task. In the one-dimensional task, the signal and templates were even Gabors. In the two-dimensional task, the signal was a Gaussian blob and the template was a Gaussian blob in one case and a difference of Gaussians in the other. The signal was normalized so that the observer correctly categorized the signal on 81% of the trials before the addition of internal noise. The internal noise ϵ_i was chosen from a Gaussian distribution $p(\epsilon) \sim \mathcal{N}(0, \sigma^2)$, after which the performance dropped to 75%. Since Gaussians are mapped to Gaussians under linear transformations, $p(v) \sim \mathcal{N}(\mu, \sigma_r^2)$ for trials of the same signal sign and the following relation holds:

$$\sigma/\sigma_r = \sqrt{\left(\frac{\Phi^{-1}(0.81)}{\Phi^{-1}(0.75)}\right)^2 - 1} \approx 0.833. \tag{14}$$

This gave an observer self-consistency on repeated simulated presentations of the stimuli of 0.76, within the range of reported values of self-consistency in various classification image experiments (Murray et al., 2002; Neri & Levi, 2008). The offset or criterion term c was adjusted such that the observer was unbiased. We estimated the observer's internal template with a logistic regression GLM with a weight decay, smoothness, and sparse prior. The optimal hyperparameter λ for each prior type was estimated through 5-fold cross-validation.

For the 1D task, the signal was sampled at a resolution of 64 pixels. The corresponding sparse prior basis for the task was a Laplacian pyramid spanning three levels, with half-levels included, overcomplete by a factor of ≈4. For the 2D task, the signal was sampled at a resolution of 17 × 17 pixels, and three bases were used: the Dirac (pixel) basis, a full Laplacian pyramid, overcomplete by a factor of ≈1.25, and a full steerable pyramid basis with two orientations, overcomplete by a factor of ≈3.25.

Methods: Real observer

We tested our procedure on a real observer (author PM), whose task was to detect the presence or absence of a Gaussian blob under conditions directly analogous to the those used in our simulations. We also used variations of this task in which the observer had to identify the null signal or four Gaussian blobs placed symmetrically around the center of the screen. The 2500 trials were performed originally for each task. Signals were masked by additive independent Gaussian noise. The noise variance was adjusted by a staircase procedure so that the observer performed at 75%. The signal and noise were sampled at a resolution of 16 × 16 pixels. The results were analyzed with a full Laplacian pyramid with half-levels, overcomplete by a factor of ≈2, in conjunction with the sparse prior. The 2500 trials were separated at random into a fit set containing 2000 trials and a validation set containing 500 trials.

Inference power estimation

For the four-blob task, an additional 2700 trials were collected. The goal was to estimate how many trials one must do to have sufficient power to reject a baseline hypothesis using models with different priors. This is equivalent to asking how many trials one must do for the cross-validated deviance of a given model to be smaller than that of a baseline model. We first pooled the additional trials with the original trials, for a total of 5200 trials. We took 100 samples each of lengths 500, 800, 1200, 2000, 3250, and 5100 trials, without replacement, from this pool. For each sample, we fit a variety of different models, discussed in the Real observer section. Cross-validated deviance was averaged across samples of the same length to obtain an estimate of the mean cross-validated deviance attained by each model for a certain number of trials.

Results

Simulated observer, one-dimensional Gabor

We first simulated an experiment in which the linear observer had to detect a one-dimensional Gabor stimulus (Figure 3, lower right) embedded in additive, Gaussian noise. Figure 3 shows estimated templates for increasing numbers of simulated trials based on the standard weighted sums formula (top row), a GLM with a smoothness prior (middle row), and a GLM with a sparse prior in a Laplacian pyramid basis (bottom row). The sparse prior template estimate is accurate even for very low numbers of trials. In terms of correlation between the real and estimated templates, the sparse prior estimate at 200 trials is comparable to the smoothness prior estimate at somewhere between 500 and 1000 trials.

The smoothness prior performs suboptimally here because the smoothness scale varies over the template. The sides of the template are flat (infinitely smooth), while

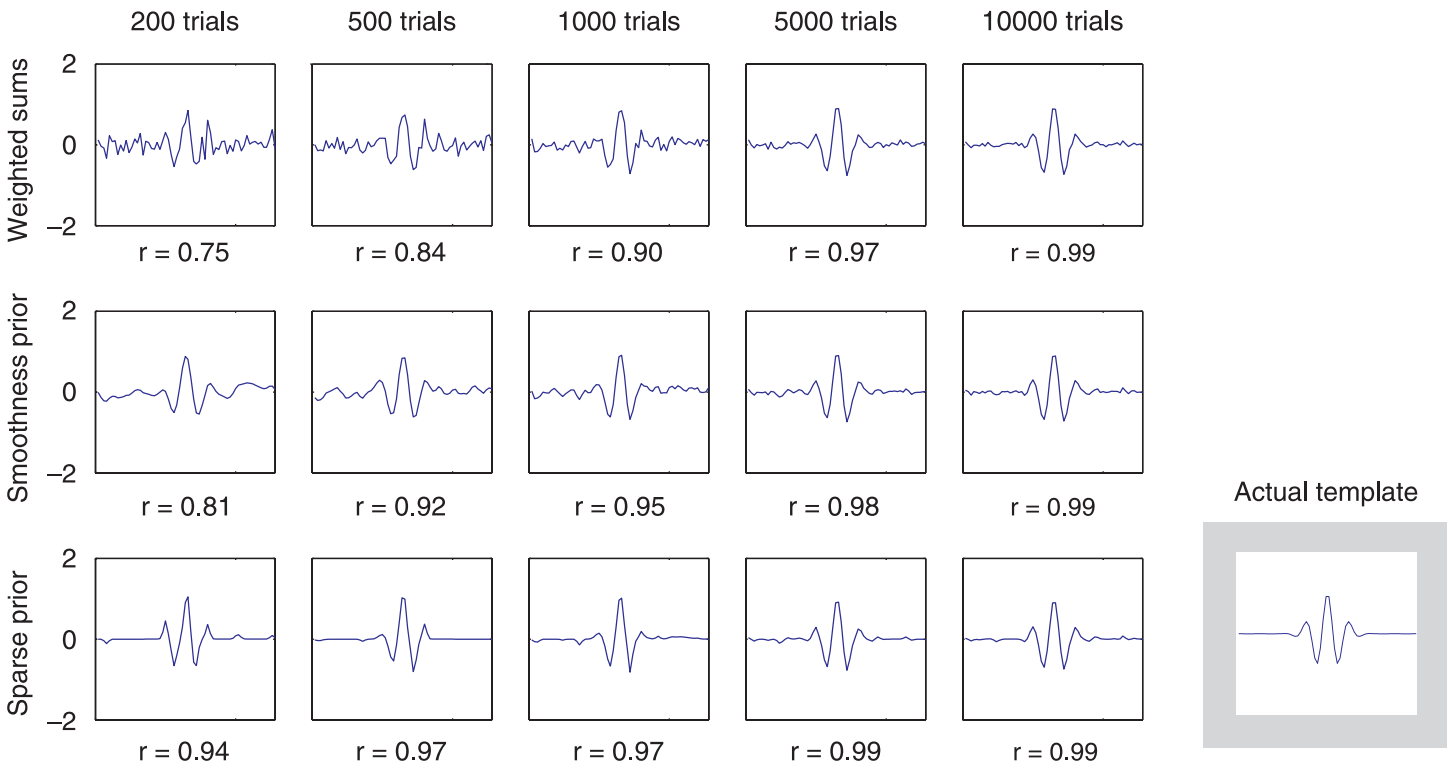

Figure 3. Estimated templates for a simulated linear observer on a 1D Gabor detection task with varying priors and numbers of trials. The correlation between the estimate and the actual template is given below each estimated template. Inset: the actual simulated internal template. With a sparse prior, 200 trials suffice to obtain an accurate estimate of the actual template.

the center of the template is smooth on a small spatial scale. Using a stronger smoothness prior to eliminate the noise on the side of the template would oversmooth the center of the template. The shortest characteristic length scale of the internal template acts as an upper bound on the level of smoothness that can be imposed on the template without significantly biasing parameter estimates. In contrast, the sparse prior in a Laplacian pyramid basis is highly effective here, as a Gabor can be represented sparsely in this basis.

More insight into the effect of a sparse prior can be gained by considering the weights as a function of the strength of the regularization. Figure 4A shows the estimated weights as a function of λ for 1000 simulated trials. For large λ, most weights are zero. As λ is decreased, more weights become active. Once a weight has become active, it tends to stay active for smaller λ.

Figure 4B shows the template estimates for different values of λ. For large λ, only the areas of the template

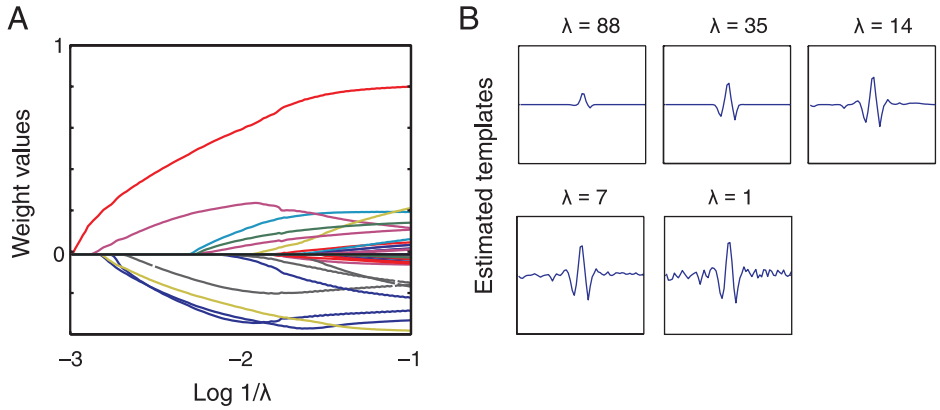

Figure 4. Weight paths with sparse priors. (A) Estimated weights as a function of inverse regularization strength. Each line represents one of the weights of the fitted GLM model as a function of log $1/\lambda$. As λ decreases, more weights are added to the model. Once a weight is active, it tends to stay active. (B) Estimated templates for different values of λ. At large values of λ, only the most important areas of the template are recovered. The reconstruction becomes more complex and accurate for smaller λ. Beyond a certain λ, the model starts fitting to noise.

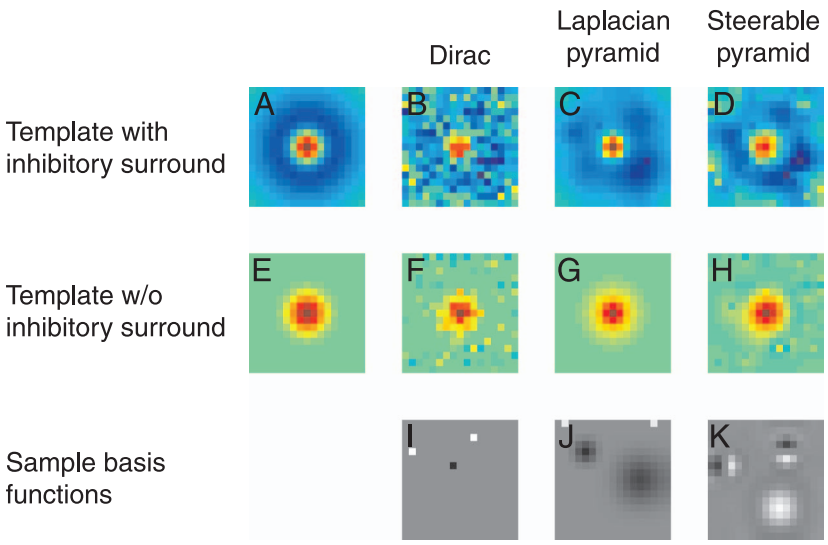

Figure 5. Estimated internal templates for simulated linear observer on detection task with 2D DOG templates. The Dirac basis cannot sparsely represent the template with inhibitory surround. The Steerable pyramid has basis functions which include positive and negative areas, and hence it is unsuitable for judging the existence of an inhibitory surround. The Laplacian pyramid is well suited for this task.

which have the most influence on the observer's decision are recovered. As λ is decreased and more weights become active, the reconstruction becomes more complex and accurate. Beyond a certain point, however, the model starts fitting to noise, and the reconstruction becomes less accurate. From this point of view, the sparse prior works by determining how influential each weight is, and only allowing weights into the model which are more influential than a cutoff that is determined by the strength of the regularization λ.

The tuning of the hyperparameter is conceptually very similar to thresholding. When an overcomplete basis is used, however, the influence of a single weight, as measured by its magnitude, no longer has a straightforward interpretation. In an overcomplete basis, there are several equivalent ways of expressing the same template, which means that the magnitude of a weight can change depending on the representation chosen. To resolve this ambiguity, the sparse prior approximately selects a best subset of model variables in a nongreedy fashion (Tibshirani, 1996). This can be viewed as an instantiation of Occam's razor: Given several models which predict the same outcome, whichever model is simplest is the preferred model. This gives a model with a sparse prior robustness against irrelevant covariates and naturally generalizes the standard practice of thresholding classification images.

Simulated observers, two-dimensional difference of Gaussians

We have shown that estimated templates are accurate when sparse priors are imposed on a suitable basis. Our proposed estimation method can accommodate any choice of basis, and it is this free choice of basis which underlies the power of the method. If, in the chosen basis, the observer's template cannot be represented sparsely, or if there is too much noise for coefficients to be well constrained, the sparse prior will tend to discard coefficients which have little influence on the outcome: the model is truncated. Model truncation can lead to artifacts whose nature depends on the basis considered; typically, artifacts look like the basis functions used.

A good basis for a given classification image meets two criteria: (1) plausible templates can be represented sparsely for the given problem; and (2) possible artifacts of model truncation will not bias the experimenter's interpretation of the estimated templates. For many problems, the first criterion will rule out the Dirac (pixel) basis, in which many templates are not sparse. The second criterion rules out bases whose functions are global, because artifacts of reconstruction will spread across space; this includes the Fourier basis. A good basis should be neither completely local nor completely global, which leaves a number of schemes based on using the same basis functions at different scales, such as pyramid transforms and wavelets.

We illustrate these ideas with simulated linear observers in a 2D Gaussian blob detection task. We imagine that the experimenter wants to know, in such a task, the general shape of the observer's internal template, and in particular, whether the observer's internal template has an inhibitory surround. We simulated linear observers with two different internal templates: one with an inhibitory surround (a difference of Gaussians) and one without (a single Gaussian). Figure 5 shows typical estimates of these templates after 2000 simulated trials for three choices of basis: Dirac basis, Laplacian pyramid, and steerable pyramid with two orientations.

The Dirac basis (Figure 5I) fails to meet Criterion 1 for this problem, as the template with an inhibitory surround

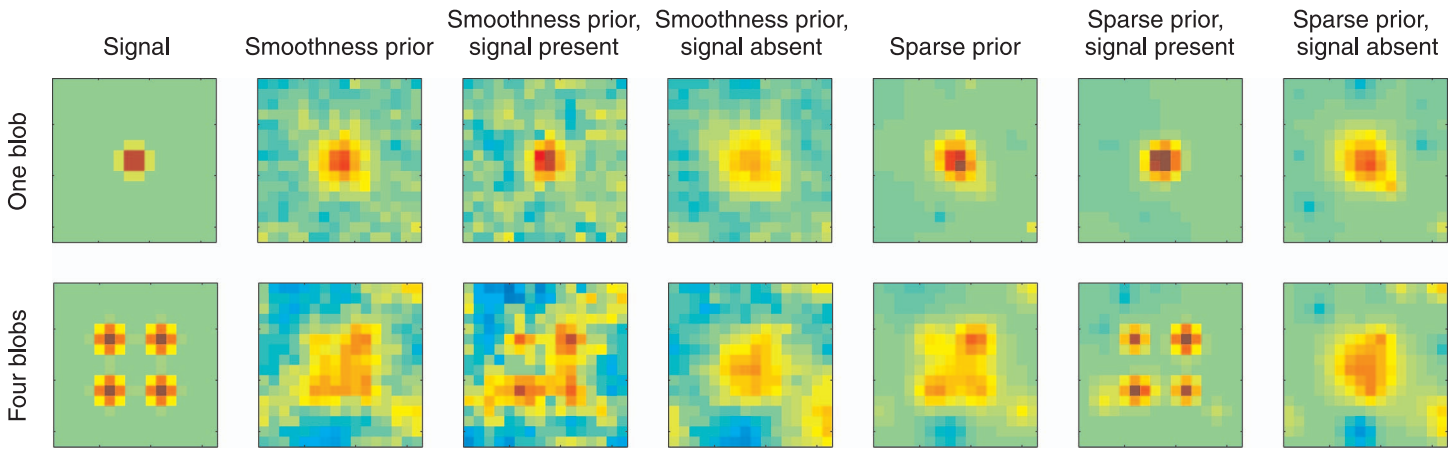

Figure 6. Estimated internal templates of real observer on one-blob and four-blob detection tasks. Models estimated with sparse priors are less noisy than those estimated with smoothness priors. This increase in efficiency allows the accurate estimation of partial classification images for the signal present/signal absent conditions.

is not sparse in pixel space. Because each pixel has a very small influence on the outcome of the model, very many have to be kept active in order to obtain a fair approximation of the template, and the sparse prior sets only 10% of the pixels to zero. The resulting templates (Figures 5B and 5F) are similar to what would be expected when using post hoc thresholding of a classification image with a low threshold.

The steerable pyramid uses Gaussian spatial derivatives at different scales as basis functions, similar to Gabors, which are appropriate for representing internal templates containing edges (Figure 5K). It fails to meet Criterion 2 for this problem, as the basis functions have both positive and negative regions. The latter can masquerade as inhibitory surrounds, and indeed the template shown in Figure 5H, contains a faint inhibitory zone around the excitatory region of the template. As the actual template has no surround, the blue region in the figure is an artifact of the choice of basis.

The Laplacian pyramid basis functions are Gaussian blobs of different sizes (Figure 5J). This is an appropriate basis here, as plausible templates for this task can be sparsely described in the basis, and the main artifact of reconstruction, excessive smoothing, is not critical in judging the existence of an inhibitory surround.

More generally, the right choice of basis depends on the specific problem at hand. Pyramid bases, composed of stereotyped smooth basis functions at different scales, are effective basis choices when the observer's internal template contains sparse smooth structure at an unknown and potentially spatially varying scale. The Laplacian pyramid, in particular, may be a safe choice for many classification image paradigms where smoothing is an effective denoising strategy. For problems in which the features of the classification image are oriented and localized, a wavelet-like basis such as the steerable pyramid may be useful. When frequency response is of importance, the discrete Fourier transform (DFT) basis can be a good choice, although one should not attempt to spatially reconstruct the template in such a basis because of ripple artifacts. Finally, bases that exploit the geometry of the task can be used to obtain the most power in answering specific questions. For example, to judge the existence or absence of an inhibitory surround, one could exploit the radial symmetry of the problem by using a basis composed of concentric rings blurred at different scales, which would give effects similar to radial averaging (Abbey & Eckstein, 2002).

Real observer

Real observers can display a number of behaviors not accounted for by a linear model: nonstationarity, lapses in attention, input nonlinearities, spatial uncertainty, and so forth. A good model should be robust to model misspecification. We therefore tested our approach with a real observer on a blob detection task similar to the one used in the simulations. In the first task, an observer was asked to indicate whether the target, a Gaussian blob, was present in the center of the screen. The target or absence thereof was masked by additive Gaussian noise. The second task was similar, but the target was now four Gaussian blobs placed around the center of the screen. Figure 6 shows the target stimulus (column 1), along with internal templates estimated under a GLM model with smoothness (columns 2–4) and sparse (columns 5–7) priors. The sparse prior was defined in a Laplacian pyramid basis as described in the Methods section.

For the one-blob task (top row), the sparse prior estimate shows a shape similar to that obtained with the smoothness prior, although noticeably less noisy. This increase in efficiency allows us to estimate with some fidelity two partial classification images, one for when the signal is present and one for when the signal is absent, by splitting the design matrix in two as described in Knoblauch and

Model type	*D*	*df*	AIC	D^{CV}	D^{val}
Ideal observer	2016	3	2022	2023	488.7
Weight decay prior	1748	166	2080	2109	503.2
Smoothness prior	1818	102	2022	2025	480.9
Sparse prior	1888	38	1964	1965	474.9
Pseudo-ideal observer	1985	4	1993	1994	474.0
Weight decay with signal effect	1722	203	2129	2134	492.3
Smoothness with signal effect	1680	173	2026	2028	461.9
Sparse prior with signal effect	1863	32	1927	1917	456.8

Table 2. Summary of fit results for several models in the 1-blob task. *D,* deviance of the estimated model; *df,* degrees of freedom; AIC, Akaike Information Criterion; D^{CV}, cross-validated deviance; D^{val}, deviance of predictions on validation set.

Maloney (2008b). The template corresponding to conditions in which the signal was present is highly spatially localized and quite similar to the signal, while the *signal-absent* template is more blurry. This is likely due to spatial uncertainty, which strongly affects recovered signal-absent templates but not signal-present templates in detection tasks (Tjan & Nandy, 2006). This effect is visible in a similar task in Figure 4 of (Abbey & Eckstein, 2002), and an analogous effect was shown in the time domain in Knoblauch and Maloney (2008b). Both the signal-present and signal-absent templates show hints of a large, weak inhibitory surround.

In the four-blob scenario (bottom row), the template estimated through the sparse prior again appears less noisy than that obtained with a smoothness prior. The partial templates show a pattern consistent with spatial uncertainty, with the estimated signal-present template looking very much like the signal, and the signal-absent template being just a blur. Again, signal-present and signal-absent templates show hints of a weak inhibitory surround. Estimated templates in both tasks are thus compatible with a spatially uncertain observer who judges the presence or absence of a blob by comparing the luminance around one or several reference points to the luminance of the surround.

It is noteworthy that the sparse prior estimates show the same kind of features which are visible, with hindsight, in the smoothness prior estimates; the sparse prior simply enhances the visibility of these features and denoises the estimated classification images. The sparse prior achieves this denoising by setting parameters which contribute little to the model to 0, as can be seen by the large areas of pure green in the estimated classification images. This in turn permits better resolution of subtle effects, such as the difference in the size of the excitatory blob in signal-present and signal-absent conditions in the one-blob task (top row, columns 3–4 versus columns 6–7), indicative of spatial uncertainty. This enhancement in the interpretability of the results is a key advantage of the sparse prior over other choices of prior.

There are several ways of performing quantitative comparisons of different models within the GLM framework, and these are discussed at length in Appendix B. In general goodness-of-fit is measured by the *deviance,* defined as twice the negative log-likelihood. Since the deviance always decreases with increasing number of free parameters, deviance values must be corrected for number of free parameters. Here we use the cross-validated deviance (D^{CV}), the Akaike Information Criterion (AIC), and the validated deviance (D^{val}) as measures of goodness-of-fit. All three metrics measure the generalization performance of a model; the first by repeated partitions of the data, the second using an asymptotic result, and the last with a set-aside validation set. In all cases, lower values are better.

Tables 2 and 3 compare the various models according to their goodness-of-fit and estimated degrees of freedom. We have included two models that can be considered as baselines. One is the ideal observer model, in which the response is given by the signal multiplied by an undetermined constant, with a signal presence/absence effect, plus an offset, embedded in a GLM with a weight decay prior. A second baseline is a pseudo-ideal observer similar to the first, but with separate gains for the signal-present and signal-absent cases, corresponding to a signal-sign dependent efficiency. The others models considered are GLMs with a weight decay prior, a smoothness prior, and a sparse prior in the Laplacian pyramid basis, either using a single template or partial templates for the signal-present and signal-absent conditions. In all cases, 5-fold cross-validation was used to determine optimal hyperparameters.

In both the one-blob (Table 2) and the four-blob (Table 3) cases, there is a clear pattern in which the model with a weight decay prior exhibits the worst performance, while the smoothness prior performs better, and the sparse is prior better still. Furthermore, the decrease in AIC and CV deviance when considering separate templates for signal-present and signal-absent conditions is most dramatic when using a sparse prior. The reason is simple: while the observer is behaving sufficiently nonlinearly for separate templates to be warranted for the signal-present and signal-absent cases, models with weight decay and smoothness priors must expend a large number of degrees of freedom to deal with the two conditions separately, as shown in the *df* column. The decreased deviance is thus overshadowed by a large increase in degrees of freedom.

Classification image experiments are often designed to detect deviations from ideal observer behavior; these

Model type	D	df	AIC	D^{CV}	D^{val}
Ideal observer	2157	3	2163	2168	545.7
Weight decay prior	1989	112	2213	2218	571.4
Smoothness prior	2051	61	2174	2179	556.1
Sparse prior	2110	36	2182	2161	545.3
Pseudo-ideal observer	2127	4	2134	2140	542.9
Weight decay with signal effect	1980	114	2207	2202	569.2
Smoothness with signal effect	1956	106	2169	2161	560.6
Sparse prior with signal effect	2041	45	2131	2130	541.9

Table 3. Summary of fit results for several models in the four-blob task. Legend as in Table 2.

deviations are then considered as signs of hard-wired strategies or mechanisms. In many classification experiment paradigms, a failure to detect a reliable departure from ideal observer behavior would warrant running more trials or changing the experimental paradigm. Table 2 shows that in the one-blob detection experiment, an ideal observer or pseudo-ideal observer null hypothesis cannot be safely rejected on the basis of the weight decay prior. In the smoothness prior case, the situation is complicated as D^{val} favors the smoothness prior over the null hypotheses, but not D^{CV} nor the AIC, which suggests that the models estimated with the smoothness prior are performing sufficiently close to the ideal and pseudo-ideal observers to warrant doing more trials. Table 3 shows a similar effect in the four-blob task, as neither the weight decay nor the smoothness prior is close to rejecting the null hypotheses. In contrast, the model that makes use of a sparse prior supports the idea that the observer is behaving nonideally and nonlinearly, decisively in the one-blob task and to a lesser extent in the four-blob task. Again, looking at the *df* column, it is clear that the sparse prior achieves this by aggressively searching for low-dimensional models.

These results indicate that the sparse prior yields improvements in goodness-of-fit and inference power in comparison to smoothness or weight decay priors. We next sought to determine how these improvements depend on the total number of trials by collecting additional data in the four-blob task (total of 5200 trials). For each of the eight models considered above, we computed the cross-validated deviance for different numbers of trials, using a resampling procedure described in the Methods section. Based on this, we derived an estimate of the number of trials required for models with different priors to perform significantly better than a baseline.

Figure 7 plots the cross-validated deviance per trial for all eight models, for different numbers of trials. Under a model with no free parameters, the CV deviance per trial should be a straight line as a function of the number of trials. An ideal observer model (left, blue line) has few free parameters, so it asymptotes fast at a high value of CV deviance per trial (roughly 1.043). In contrast, linear observer models with various priors are more flexible, but they require more data to be well constrained. With a weight decay prior, it takes more than 5000 trials to

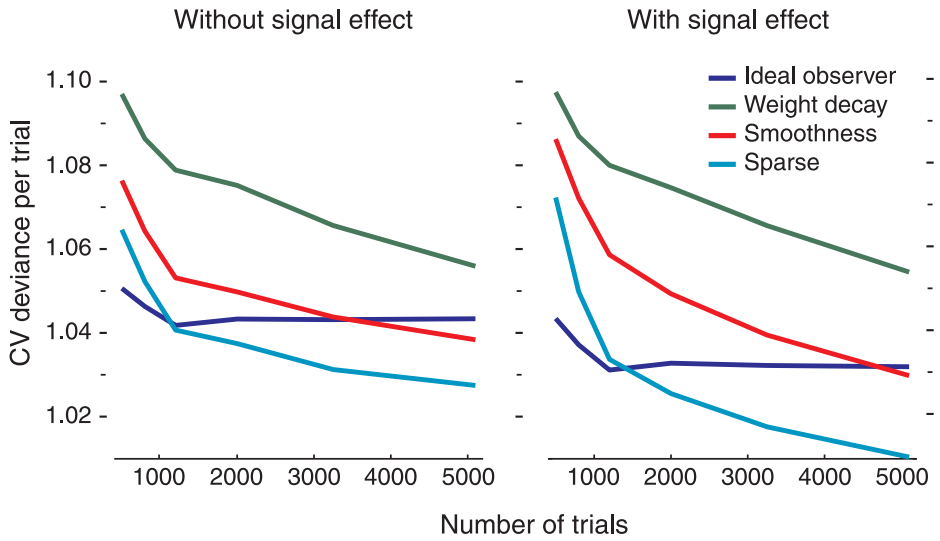

Figure 7. CV deviance per trial in the four-blob task estimated with different number of trials. Left: without a signal effect. Right: with a signal effect. The model estimated with a sparse prior reaches significance against a null model appreciably faster, requiring less trials for the same quality of fit.

disprove the ideal observer null hypothesis; with a smoothness prior, about 3400 trials; and with a sparse prior, about 1100 trials. In other words, we can obtain the same inference power under the sparse prior as under a smoothness prior with less than a third of the trials.

With a signal effect, which requires the estimation of twice as many parameters, the conclusion is similar; the model with a sparse prior reaches significance against the pseudo-ideal observer null hypothesis with roughly 30% of the trials required to reach the same conclusion under a smoothness prior (1400 versus 4700 trials). The GLM with a sparse prior thus uses the observer's data more efficiently, leading both to an appreciable increase in inference power for a fixed number of trials, and to an appreciable decrease in the number of trials required to reach a certain inference power.

Sparse priors are thus helpful in real experiments in two main ways. First, they yield clear, noise-reduced internal template estimates which facilitate interpretation, in the same way that thresholding does. Second, by aggressively searching for low-complexity interpretations of the data, sparse priors recover better models, which allows one to conclusively show certain properties of observer's strategies without the need to gather data from an impractically large number of trials. The improvements in model fit brought by the use of a sparse prior are appreciable in a real sense due to the high noise inherent in classification image protocols and the high dimensionality of the models involved. These results show that the proposed method can supplement traditional methods even in simple tasks.

Discussion

In this work we have argued that a defining quality of classification images is simplicity, in that they may be expressed as a sparse sum of smooth basis functions. We have derived a method for imposing this condition through a sparse prior on smooth basis coefficients in a GLM framework. We showed in simulations that classification images estimated with sparse priors are often more accurate for a given number of trials than those estimated through other methods.

A key advantage of sparse priors over Gaussian priors is enhanced interpretability, as coefficients which do not contribute significantly to an observer's decision process are discarded from the model. We showed that a sparse prior allowed efficient estimation of the internal template of a real observer on a blob detection task. This increase in efficiency allowed for the accurate estimation of partial classification images for the signal-present and signal-absent conditions, revealing important differences thought to be due to spatial uncertainty. This increase in efficiency also lead to an appreciable decrease in the number of trials necessary to reach a certain inference power.

Sparse priors are especially useful in high-dimensional tasks, such as tasks involving both time and space. The number of classification image pixels to estimate in such models can be very high, and estimated classification images can be too noisy to be useful. Even in cases where the extrinsic dimensionality of the classification image is very high, the *intrinsic dimensionality* of the observer's template, that is, the number of basis coefficients needed to accurately represent it, may be quite low. Sparse priors in bases use this fact to effectively estimate the parameters of seemingly complex models.

Prior assumptions

A common argument against using priors is that they might bias the estimated internal templates and therefore lead to erroneous interpretations of observers' strategies in performing a task. A more general statement would be that priors introduce a bias-variance tradeoff: By restricting the space of possible models to some subset of all models, bias is introduced, but variance (noise) in the estimated models is reduced. In this sense, classic psychophysics, in which a handful of parameters are allowed to change, can be seen as having high bias and low variance, while unprocessed classification images have low bias but high variance. Our approach can be seen as a balance between these two extremes, with modest bias and variance.

Classic denoising methods such as smoothing and thresholding also implement a bias-variance tradeoff. The main difference between using a prior and classic denoising methods is that assumptions are made explicit in the prior approach. Hence, the question is not so much whether it is appropriate to use assumptions, unless one wants to reject classic denoising methods as well, but whether the assumptions *that one uses* are appropriate. If the assumptions are indeed warranted, then a reduction in noise during model estimation can lead to more powerful inference. The Real observer section gives an example of this, where the model fitted with a sparse prior allows us to forcefully infer that the observer is using a strategy that is both nonideal and nonlinear, in contrast to the conclusion obtained via models fitted with a weight decay and a smoothness prior.

The validity of one's choice of prior, and more generally one's choice of model, can be measured by the AIC and the cross-validated deviance, provided that the choice of prior or model was made before the examination of data. When a new model form is suggested by fitting a baseline model to data, complexity-corrected deviance measures will be overly optimistic about the new model; this process is known, derisively, as data dredging or "double dipping" (Kriegeskorte, Simmons, Bellgowan, & Baker, 2009). There are straightforward ways to address this issue. The first is to collect a leave-aside validation set, which is not used to suggest model form, and base judgments of model validity exclusively on the probability of the validation data under the different models. By Bayes' theorem, this approach provides an estimate of the probabilities of the

models themselves. This can be regarded as a probabilistic version of the validation model suggested by (Neri & Levi, 2006). While this most certainly works, as shown in the Real observer section, binary responses have high variance, and much validation data must be collected for the results to be reliable. A more efficient use of the data is to pool the validation and fit sets into one large data set.

A more practical solution is to determine model form using preliminary data. It is customary, in classification image experiments, to collect a limited quantity of pilot data based on one or two conditions with a single subject to judge whether the complete experiment with multiple conditions and subjects is likely to yield meaningful results. Model form, including the prior up to a limited number of hyperparameters, is settled at this point, and the pilot data are discarded from subsequent analysis. AIC and cross-validated deviance scores in the full experiment based on the settled models are then reliable measures of their validity.

Assuming one decides to use a sparse prior in a particular basis, the choice of basis is determined at the preliminary stage. The choice of good bases is heavily constrained, as explained in the results section, so that the experimenter's choice reduces to a handful of bases related to a pyramid decomposition, save for a few special cases. The experimenter chooses whichever of these few options works best on the preliminary data and continues to use this basis to analyze the results of the full experiment. Note that these remarks are also applicable to standard methods of classification image estimation; the choice of smoothing kernel and threshold should not be determined with the same data that is used to judge the result of the experiment unless corrections for multiple comparisons are used (Chauvin et al., 2005; Worsley, Marrett, Neelin, & Evans, 1996). Hence, prior choice need not be subjective.

Relationship to previous work

Sparse GLM models

In this article, we have suggested using a sparse prior on smooth basis coefficients of GLM model parameters as a method of accurately estimating classification images. Similar techniques have been used to induce sparseness of GLM parameters in the context of neurophysiology.

Sahani and Linden (2003) suggest using Gaussian priors in a linear model with one hyperparameter per parameter, optimizing over the marginal likelihood of the model, a technique called automatic relevancy determination (ARD). In practice, ARD sets many coefficients in the chosen basis to zero, leading to sparse solutions, as with the sparse prior. The authors also suggest an alternative method which combines automatic smoothness determination and ARD. This gives results similar to imposing a sparse prior in a basis composed of Gaussian blobs which all have the same size, this size being automatically determined. In principle, ARD could be used with an overcomplete basis such as the Laplacian pyramid. However, ARD involves a nonconvex optimization problem, and so there can be multiple local minima in the induced error function, and the algorithm may not converge. This is not a major issue when the chosen basis is orthonormal, but with an overcomplete basis, multiple local minima are likely to form, as there are many ways of expressing the same vector. In addition, it can be unwieldy to extend ARD to more complex GLMs than the linear regression model (Wipf & Nagarajan, 2008; Bishop, 2006), in particular the logistic and probit regression models relevant to classification image experiments.

David, Mesgarani, and Shamma (2007) suggest using boosting as a method of estimating sparse spectro-temporal receptive fields in the context of auditory coding. In this method, the model is fit in steps, starting with an empty model. At each iteration, the residual between the fitted model and the data is computed. The weight whose associated regression function is most correlated with the residual is then modified. Initial iterations fit to prominent effects while later iterations tend to fit to noise. Regularization is done by early stopping, which halts the procedure at an optimal number of iterations estimated through cross-validation. The stepwise fitting procedure means that several weights never enter into the model, resulting in a sparse model. Indeed, boosting can be shown to implicitly and approximately find the MAP estimate for a GLM with a sparse prior (Rosset, Zhu, Hastie, & Schapire, 2004).

As explained in Appendix A, the cost per boosting iteration is the same as the cost per fixed-point continuation iteration, and the number of iterations required by boosting with overcomplete bases can be substantially larger than with our proposed method. Our proposed method thus offers a way of estimating a sparse GLM model that is, under appropriate circumstances, computationally more efficient than boosting.

Seeger, Gerwinn, and Bethge (2007) have recently proposed using GLM models with sparse priors to estimate receptive fields. In contrast to our approach, they attempt to estimate the entire posterior rather than just its mode. This is done by finding a Gaussian density which minimizes the Kullback–Leibler divergence between the real and approximated posterior, a technique called expectation propagation (EP).

Although EP and MAP methods use the same underlying model and prior, they have rather different properties. Since the EP technique obtains an estimate of the full posterior distribution rather than just a point estimate, it can be used for selecting the stimulus that is most likely to help constrain model parameters at a given point in an experiment, a technique called active design. While the MAP method retrieves a point estimate in which several coefficients are exactly zero, the EP method retrieves the

mean of the approximated posterior for which all coefficients are nonzero. This is motivated by the fact that it is impossible, with any finite amount of data, to conclude that a coefficient is exactly zero, and that hard subset selection is incompatible with active design (Seeger et al., 2007). However, we suggest that the subset selection effect of the MAP method is highly desirable for interpretation purposes, automatically removing from the model coefficients that are likely to be fitting to noise, and thus letting the experimenter focus on the more prominent effects visible in a classification image. Finally, the EP optimization problem is more involved, numerically unstable, and computationally more expensive than the MAP estimation problem. Hence, although EP can be used to estimate receptive fields and classification images, in general EP and MAP methods are complementary techniques with different target applications.

A key point is that none of the proposed sparsity-inducing methods (including ours) have so far shown a decisive edge in the *quality* of estimated models compared to other sparsity-inducing methods. In contrast, sparsity-inducing methods do have a decisive edge over using Gaussian priors on some problems (David et al., 2007; Seeger et al., 2007; Rosset et al., 2004). Thus, it is advisable to use a sparsity-inducing method in general, while choosing the particular method which is most mathematically convenient and computationally efficient for a given application. Our proposed estimation method is competitive in terms of implementation speed, as explained in Appendix A. For example, the model with the largest design matrix considered in this article ($10{,}000 \times 256$) took roughly 40 seconds to fit on a recent desktop computer, including 5-fold cross-validation.

Basis projections in classification images

Basis projections have been used implicitly as an intermediate in estimating classification images. For example, radial averaging can be viewed as a projection of a classification image onto a basis of concentric rings (Abbey & Eckstein, 2002). Projection onto a cubic spline basis of lower dimensionality than the stimulus is possible in the generalized additive model (GAM) framework of (Knoblauch & Maloney, 2008b). A radial discrete Fourier transform (DFT) basis has been used in Abbey, Eckstein, Shimozaki et al. (2002). In all these cases, the basis is undercomplete; that is, it does not span the full vector of possible templates. The purpose of the projection was thus *hard* dimensionality reduction. In contrast, a sparse prior in a basis decides which basis functions to keep, yielding *soft* dimensionality reduction.

Others have used one basis in both presentation and analysis, for example a Fourier basis (Ahumada & Lovell, 1971; Levi & Klein, 2002). A particularly interesting example of such a technique was used in Mangini and Biederman (2004) in the context of face classification. Noise fields were generated by taking weighted sums of truncated sinusoids of varying orientation and spatial frequency. The authors then estimated classification images by applying the weighted sums formula in this truncated sinusoid basis. They applied a threshold corresponding to a given p-value in the truncated sinusoid basis and visualized the results by projecting the coefficients back onto the pixel basis. Only a fraction of the coefficients (less than 5%) were kept in the process. The resulting estimated templates were similar to the non-thresholded templates, but with less high-frequency noise, and improved interpretability. The quality of the truncated reconstruction is likely due to the fact that faces can be represented sparsely in the truncated sinusoid basis, which, like wavelet and pyramid bases, is passband and local.

Importantly, the truncated sinusoid basis was used in both the construction of the noise fields and the estimation of classification images. In contrast, our method does not require presenting a special type of noise to the observer; white noise, colored noise, or natural textures may be used. The basis used in the analysis may be chosen after the classification image experiment has been performed. Finally, we do not advocate using a truncated sinusoid basis as ringing artifacts are apparent in reconstruction. The steerable pyramid basis would be appropriate for the task used in Mangini and Biederman (2004).

Directions for future work and conclusion

Directions for future work

One might wonder whether a more powerful prior could be used to obtain even more accurate estimates of classification images in a lower number of trials. While this is a possibility, a potentially more fruitful approach is to consider better experimental designs. For example, it has been shown in the context of neurophysiological reverse correlation that different noise fields cannot be expected to yield the same amount of information about a visual neuron's receptive field (Paninski, 2005).

In the context of classification images, rather than randomly generating noise fields on each trial, the field which maximizes the expected information can be shown. Without necessarily using a fully adaptive design, optimal design theory could potentially indicate which of several classes of stimuli, such as white noise, colored noise, or natural textures should be used for a given task. Optimal design strategies depend on the prior used; an optimal design method for a linear model with a sparse prior is presented in Seeger, 2008.

The main challenge is to extend the classification image approach to capture nonlinear behavior. Two methods have been commonly used so far for that purpose: using partial internal templates for signal-present and signal-absent conditions and computing second-order classification images (Knoblauch & Maloney, 2008b). In the reverse correlation literature, a successful approach has been to consider *expanded input spaces*: the stimuli, instead of being described simply as intensity over time or over time and space, are redescribed in terms of redundant features. For example, an auditory stimulus $x(t)$ can be redescribed using a windowed Fourier transform as a spectrogram $x(f, t)$, which augments the representation with the evolution of the different component frequencies f over time. A neuron's response can be modelled as a linear weighting of $x(s, t)$, and the resulting two-dimensional weight pattern is then known as a spectro-temporal receptive field (STRF). The STRF is able to capture a range of nonlinear behaviors, such as time-varying frequency sensitivity.

Extensions of this technique are described in Ahrens, Paninski, and Sahani (2008) and applied to auditory neurons in Ahrens, Linden, and Sahani (2008). Our framework is directly applicable to these and related problems, by a simple preprocessing of the input. One could also leverage existing physiological models of low-level cortical neurons to describe linear observers that operate on the output of simulated neurons. An observer in a detection task could be modelled as taking a weighted sum of V1 complex cells, and again our framework can be used here without substantial modification. However, a large number of nonlinear behaviors, such as spatial uncertainty, cannot be described by input nonlinearities (Tjan & Nandy, 2006), so future research will have to establish whether this approach is indeed fruitful.

A recent article by Ross and Cohen (2009) suggests an alternative path towards modelling nonlinear classification image observers, based on the idea that observers individually match features of the classification image and nonlinearly combine them to form a decision. It is assumed that the observer performs a classification or detection task by matching the input with several linear templates. Matches are nonlinearly transformed by a logistic function, and these transformed matches are linearly combined and fed through a final logistic function, which drives the response of the observer. This model can be viewed as a robust Bayesian reinterpretation of an artificial neural network (ANN) with a single hidden layer and is reminiscent of linear-nonlinear cascade models commonly used in neurophysiology (Rust, Mante, Simoncelli, & Movshon, 2006). One of the main challenges of the approach is the proliferation of parameters, which is roughly the product of the number of hidden units and the number of pixels in the stimulus. A Markov random field prior is used to impose a type of piecewise smoothness which encourages template values to spatially cluster around values of ±1. We believe that a sparseness-inducing prior on template parameters, together with an appropriate basis, as advocated in our approach, could prove quite potent in the context of such data-hungry multi-feature classification images.

Conclusion

Classification images are of great interest to visual psychophysicists, both because they can often be compared in a straightforward manner to neuronal receptive fields, and because they provide a powerful means of measuring observers' strategies in visual tasks. However, their use has been limited by the quality and quantity of the data that can be collected. In this work, we have described an analytical framework that allows the experimenter to formalize important assumptions that are necessary to the interpretation of classification images. Our method relies on the reasonable assumption that the internal templates used by psychophysical observers are sparse and locally smooth. These constraints can be represented by appropriate choices of a basis set to represent the space of possible images and a prior that constrains the number of parameters that contribute to the recovered image. We have shown through simulation and through experiments with a real observer that estimating classification images with our method is more efficient and accurate than previous methods.

Appendix A

Fitting algorithm

Outline

To find a MAP estimate of $\mathbf{w}$ under a sparse prior in a GLM, we need to minimize an objective function E:

$$\underset{\mathbf{w}}{\operatorname{argmin}}\ E(\mathbf{w},\lambda) = f(\mathbf{w}) + \lambda\|\mathbf{w}\|_1. \tag{A1}$$

Here f is a convex, differentiable function of $\mathbf{w}$. This optimization is nontrivial as $\|\mathbf{w}\|_1$ is nondifferentiable at the axes, where $w_i = 0$ for some i. The fixed-point continuation (FPC) algorithm (Hale et al., 2007) solves Equation A1 efficiently by combining two insights. First, the fixed-point iterations:

$$\mathbf{w} \leftarrow \operatorname{shrink}(\mathbf{w} - \tau\nabla f, \tau\lambda), \tag{A2}$$

where *shrink* is the soft-threshold operator:

$$\text{shrink}(x, \alpha) = \text{sign}(x)\ \max(|x| - \alpha, 0), \tag{A3}$$

and τ, the step size, is a small number, eventually converge to the solution of Equation A1, with each iteration coming at a very moderate cost.

Convergence is however quite slow when iterations are started from an arbitrary value of $\mathbf{w}$, known as a *cold-start*. The second insight is that if once we have found $\mathbf{w}^* = \text{argmin}_{\mathbf{w}}\ E(\mathbf{w}, \lambda_1)$, we may use $\mathbf{w}^*$ as a *warm-start* estimate for minimizing $E(\mathbf{w}, \lambda_2)$ when $|\lambda_1 - \lambda_2| / (\lambda_1 + \lambda_2)$ is small. This suggests solving the series of optimization problems:

$$\underset{\mathbf{w}}{\text{argmin}}\ E(\mathbf{w}, \lambda_i), \tag{A4}$$

with $\lambda_1 > \lambda_2 > \ldots > \lambda_{\text{end}}$, using the fixed-point iterations (Equation A2), using the result of the previous optimization to start the next optimization, a process known as *continuation*. The entire regularization path is generated as part of the process, which may be used to determine the optimal λ.

The basic form of this algorithm is elegant and is reasonably fast. Here we add three insights to form a final algorithm which, while a bit less elegant, is frequently an order of magnitude faster than this basic version. First, we add a line search over τ. Second, we use insights from (Park & Hastie, 2007) to find good values of λ to sample the regularization path. Finally, we use a blockwise implementation of cross-validation which saves iterations when λ_{optimal} is unknown and large.

Inner iterations—finding the MAP estimate for fixed λ

Here we present an elementary plausibility argument for the FPC algorithm; rigorous mathematical treatment and proofs of convergence are available in Hale et al. (2007). Consider the derivatives of the objective function E with respect to $\mathbf{w}$:

$$\frac{\partial E}{\partial \mathbf{w}} = \nabla f + \lambda\, \text{sign}(\mathbf{w}). \tag{A5}$$

Corresponding gradient descent iterations, for a step size τ take the form:

$$\mathbf{w} \leftarrow \mathbf{w} - \tau \frac{\partial E}{\partial \mathbf{w}} = \mathbf{w} - \tau(\nabla f + \lambda\, \text{sign}(\mathbf{w})). \tag{A6}$$

Here the step size τ is some small number. Such iterations will reduce the objective function E for a sufficiently small τ as long as we avoid the discontinuities in the derivatives of the penalty. One way to avoid these discontinuities is simply to set w_i to 0 as it attempts to pass through the origin:

$$\mathbf{w} \leftarrow \text{sign}(\mathbf{w})\ \max(\text{sign}(\mathbf{w})(\mathbf{w} - \tau(\nabla f + \lambda\, \text{sign}(\mathbf{w})), 0). \tag{A7}$$

An issue with this new iteration is that for a weight $w_i = 0$ before the iteration, for any $\tau > 0$ the weight leaves the axis and the derivative of the penalty changes. This suggests evaluating the derivative of the penalty at $\mathbf{w} - \tau\nabla f$ instead of at $\mathbf{w}$. We thus obtain:

$$\begin{aligned}\mathbf{w} \leftarrow\ & \text{sign}(\mathbf{w} - \tau\nabla f)\cdot \\ & \max(\text{sign}(\mathbf{w} - \tau\nabla f) \\ & (\mathbf{w} - \tau(\nabla f + \lambda\, \text{sign}(\mathbf{w} - \tau\nabla f)), 0) \\ =\ & \text{shrink}(\mathbf{w} - \tau\nabla f, \tau\lambda).\end{aligned} \tag{A8}$$

Thus, the fixed-point iterations can be viewed as gradient descent iterations with special considerations to avoid issues with the discontinuities at the axes. A thorough analysis shows that these iterations indeed converge to the minimum of $E(\mathbf{w}, \lambda)$ (Hale et al., 2007).

Line search

Fixed-point iterations of Equation A2 are guaranteed to converge for fixed τ as long as $\tau < \tau_{\max} = 2/\max_i \Lambda_{ii}$, where $\mathbf{Q}\Lambda\mathbf{Q}^T = \mathbf{H}$ is an eigendecomposition of the matrix of second derivatives of the negative log-likelihood with respect to the weights, the Hessian $\mathbf{H}$. Note that this is the same condition as in gradient descent. The fixed-point iteration's relationship with gradient descent hints that convergence may be substantially faster if τ is chosen optimally on every iteration such that:

$$\tau_{\min} = \arg\min_{\tau > 0} E(\text{shrink}(\mathbf{w}^0 - \tau\mathbf{g}, \tau\lambda)). \tag{A9}$$

Over a series of informal experiments, we have noticed that performing a line search can reduce the number of inner iterations required for convergence by an order of magnitude or more over an optimal, fixed τ. Unfortunately, a naive line-search over τ involves multiple products of the form $\mathbf{Xw}$ and repeated computations of E, which are expensive; hence, overall using a naive line search does not yield a large performance improvement over a fixed τ. Here we propose a line search algorithm which, although rather involved, is much more efficient than a naive line search.

First notice that $E(\tau)$ has a peculiar form: its derivatives of all orders with respect to τ are continuous outside of a finite number of "cutpoints" $\tau_0 = 0 < \tau_1 < \tau_2 < \ldots < \tau_n$

which occur when a weight enters or leaves the model. Thus, we approximate $E(\tau)$ as a piecewise quadratic, convex function, whose minimum is inexpensively found.

Let $\boldsymbol{\eta}(\mathbf{w}) \equiv \mathbf{Xw} + \mathbf{Uu}$ and $\mathbf{w}(\alpha) = \text{shrink}(\mathbf{w}^0 - \alpha\mathbf{g}, \alpha\lambda)$. We may Taylor-expand f as a function of α up to second order, which by the chain rule gives:

$$f(\alpha) \approx f(0) + \alpha \frac{\partial f}{\partial \boldsymbol{\eta}} \cdot \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\alpha}} + \frac{1}{2}\alpha^2 \left(\frac{\partial f}{\partial \boldsymbol{\eta}} \cdot \frac{\partial^2 \boldsymbol{\eta}}{\partial \boldsymbol{\alpha}^2} + \frac{\partial^2 f}{\partial \boldsymbol{\eta}^2} \cdot \left(\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\alpha}} \right)^2 \right). \quad \text{(A10)}$$

Here $(\cdot)$ denotes the dot product. All the derivatives are computed at $\alpha = \epsilon$, $\epsilon > 0$ arbitrarily small, to give right-sided derivatives at $\alpha = 0$. This approximation is valid between $0 < \alpha < \alpha_{\max}$ assuming no weights enter or leave the model in this range. The derivatives of f with respect to $\boldsymbol{\eta}$ are straightforward to compute; they also occur in the iteratively reweighted least-squares (IRLS) algorithm often used to fit GLMs (Wood, 2006). The derivatives of $\boldsymbol{\eta}$ are given by

$$\begin{aligned} \frac{\partial \boldsymbol{\eta}}{\partial \alpha}|_\epsilon &= \mathbf{X} \frac{\partial \mathbf{w}}{\partial \alpha}|_\epsilon \\ &= \mathbf{X}\left(\left(-\mathbf{g} - \lambda \operatorname{sign}\left(\mathbf{w}^0 - \epsilon \mathbf{g} \right) \right) \cdot (\mathbf{w}(\epsilon) \neq 0) \right) \frac{\partial^2 \boldsymbol{\eta}}{\partial \alpha^2}|_\epsilon = 0. \end{aligned} \quad \text{(A11)}$$

Here $(\mathbf{w}(\epsilon) \neq 0)$ is a vector with value 1 if $\mathbf{w}_i(\epsilon) \neq 0$ and 0 otherwise. Similarly, the regularizer $R = \lambda \|\mathbf{w}\|_1$ may be Taylor expanded between $0 < \alpha < \alpha_{\max}$ to give:

$$\begin{aligned} R(\alpha) &= R(0) + \alpha \frac{\partial R}{\partial \alpha}|_\epsilon \\ &= R(0) \\ &\quad + \alpha\lambda \sum_i (-\mathbf{g} \operatorname{sign}(\mathbf{w}^0 - \epsilon \mathbf{g}) - \lambda) \cdot (\mathbf{w}(\epsilon) \neq 0). \end{aligned} \quad \text{(A12)}$$

Note that this last expansion is exact. Thus, $E(\alpha) = f(\alpha) + R(\alpha) = A + B\alpha + \frac{1}{2}C\alpha^2$ is quadratic in α and it has an extremum at $\alpha_{\min} = -B/C$. Now:

$$C = \frac{\partial^2 f}{\partial \boldsymbol{\eta}^2} \cdot \left(\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\alpha}} \right)^2. \quad \text{(A13)}$$

For GLMs, $\frac{\partial^2 f}{\partial \boldsymbol{\eta}^2} \geq 0$ (Wood, 2006); hence, $C > 0$ and E has a minimum at $\alpha_{\min} = -B / C$. Thus, either $0 < \alpha_{\min} < \alpha_{\max}$, in which case we have found a minimum for E, or not, in which case the minimum of E is located elsewhere and the approximations are no longer valid.

Given the second-order Taylor expansion of E, it follows that E is approximately piecewise quadratic, continuous, and convex away from the cutpoints. At a cutpoint τ_i, $E(\tau_i)$ has a local maximum if and only if the left-sided derivative of $E(\tau_i)$ is positive and the right sided derivative is negative; it is straightforward to show that this is never the case. We conclude that $E(\tau)$ is piecewise quadratic, continuous, and convex, and hence it has a single minimum in the range $0 < \tau < \infty$. This suggests attempting to find a minimum of E between $0 < \tau < \tau_1$. If the minimum is not in that range, we search for the minimum in the range $\tau_1 < \tau < \tau_2$ using a new Taylor expansion at $E(\mathbf{w}(\tau_1))$, and so forth until we find the minimum of E. This gives Table A1.

Notice that the iterations for $i \geq 2$ are rather inexpensive, save for computing $\frac{\partial \boldsymbol{\eta}}{\partial \alpha}$. However, at a cutpoint, $\frac{\partial \mathbf{w}_j}{\partial \alpha}$ changes for exactly one j, corresponding to the weight which enters or leaves the model. This implies that $\frac{\partial \boldsymbol{\eta}}{\partial \alpha}$ changes by a multiple of a single column of $\mathbf{X}$ at cutpoints, and hence $\frac{\partial \boldsymbol{\eta}}{\partial \alpha}$ is updated at very little cost.

The expensive computations in the initial iteration are that of f, $\frac{\partial f}{\partial \boldsymbol{\eta}}$, $\frac{\partial^2 f}{\partial \boldsymbol{\eta}}$, $\frac{\partial \boldsymbol{\eta}}{\partial \alpha}$, $\frac{\partial R}{\partial \alpha}$. $\boldsymbol{\eta}$ is saved from the previous line search, which avoids computing a product of the form $\mathbf{Xw}$. To compute f and its derivatives, roughly N logarithms, N exponentials, and a few element-wise products of vectors of length N must be performed, where is the number of trials. For $\frac{\partial \boldsymbol{\eta}}{\partial \alpha}$, one product $\mathbf{X}\frac{\partial \mathbf{w}}{\partial \alpha}$ is formed; since the right-hand side is a vector which contains mostly zeroes, this is rather inexpensive. Thus, the proposed line search algorithm, while rather involved, avoids computing $\mathbf{Xw}$ and f repeatedly, and thus is much less expensive than a naive line search.

- Compute the ordered set of cutpoints of $E(\tau)$, giving $\tau_0 = 0 < \tau_1 < \tau_2 < \ldots < \tau_{n-1} < \tau_n = \infty$
- For i from 1 to n
  - If $i = 1$
    - Compute $\boldsymbol{\eta}$, f, $\frac{\partial f}{\partial \boldsymbol{\eta}}$, $\frac{\partial^2 f}{\partial \boldsymbol{\eta}}$, $\frac{\partial \boldsymbol{\eta}}{\partial \alpha}$, $\frac{\partial R}{\partial \alpha}$ evaluated at $\alpha = \epsilon$
  - Else
    - Let $\Delta\alpha \leftarrow (\tau_{i-1} - \tau_{i-2})$
    - Update $\boldsymbol{\eta} \leftarrow \boldsymbol{\eta} + \Delta\alpha \frac{\partial \boldsymbol{\eta}}{\partial \alpha}$
    - Update $f \leftarrow f + \Delta\alpha \frac{\partial f}{\partial \boldsymbol{\eta}} \cdot \frac{\partial \boldsymbol{\eta}}{\partial \alpha} + \frac{1}{2}\Delta\alpha^2 \frac{\partial^2 f}{\partial \boldsymbol{\eta}^2} \cdot \left(\frac{\partial \boldsymbol{\eta}}{\partial \alpha} \right)^2$
    - Update $\frac{\partial f}{\partial \boldsymbol{\eta}} \leftarrow \frac{\partial f}{\partial \boldsymbol{\eta}} + \Delta\alpha \frac{\partial^2 f}{\partial \boldsymbol{\eta}^2} \frac{\partial \boldsymbol{\eta}}{\partial \alpha}$
    - Update $\frac{\partial \boldsymbol{\eta}}{\partial \alpha}$, $\frac{\partial R}{\partial \alpha}$
  - End If
  - Compute $\alpha_{\min}$
  - If $0 < \alpha_{\min} < (\tau_i - \tau_{i-1})$
    - $\tau_{\min} = \tau_{i-1} + \alpha_{\min}$; Exit
  - Else if $\alpha_{\min} < 0$
    - $\tau_{\min} = \tau_{i-1}$; Exit
  - End If
- End For

Table A1. Line search for $\tau_{\min}$.

Outer iterations

The FPC algorithm works by solving Equation A1 by iterative thresholding for decreasing values of λ. After each set of inner iterations, $\mathbf{u}$ is reoptimized and a smaller value of λ is chosen. The next value of λ is set to the largest λ for which a new weight should appear in the model:

$$\tilde{\lambda}_{\text{new}} = \max_{i \in \{j | w_j = 0\}} |(\nabla f(\mathbf{w}_{\text{est}}))_i|. \quad \text{(A14)}$$

Simulations show that the optimization can stall when $\tilde{\lambda}_{\text{new}}$ is continually chosen to be very close to the current λ, and that the estimate is inaccurate when $\tilde{\lambda}_{\text{new}}$ is far from the current λ. We thus bracket $\tilde{\lambda}_{\text{new}}$ to be neither too close nor too far from λ_{current}:

$$\lambda_{\text{new}} = \begin{cases} \alpha_{\min}\lambda_{\text{curr}} & \text{when } \tilde{\lambda}_{\text{new}} < \alpha_{\min}\lambda_{\text{curr}} \\ \tilde{\lambda}_{\text{new}} & \\ \alpha_{\max}\lambda_{\text{curr}} & \text{when } \tilde{\lambda}_{\text{new}} > \alpha_{\max}\lambda_{\text{curr}} \end{cases}. \quad \text{(A15)}$$

We found that using a nonuniform step size is more efficient than any one fixed step size, and used $\alpha_{\min} = 0.9$, $\alpha_{\max} = 0.98$.

Initial and final iterations

At the very start of the optimization, $\mathbf{w}$ is set to 0, and $\mathbf{u}$ is optimized. ∇f is computed, and λ_{start} is set to be slightly smaller than the first value of λ for which a weight should appear in the model:

$$\lambda_{\text{start}} = \max_i |\nabla_i f|. \quad \text{(A16)}$$

Outer iterations are stopped when $\lambda_{\text{next}} < \epsilon\lambda_{\text{start}}$ for some user-defined value of ϵ, which is set, by default, to 10^{-3}.

Cross-validation

The number of λ values considered should be minimized within reason to obtain an efficient algorithm. To ensure that fitting is done over a tight range of λ values, we implemented an efficient version of k-fold cross-validation, which we call blockwise cross-validation. Rather than fitting each fold from $10^{-3}\lambda_{\text{start}} < \lambda < \lambda_{\text{start}}$ in turn, we first fit the model for all folds from $0.7\lambda_{\text{start}} < \lambda < \lambda_{\text{start}}$ and compute the cross-validated deviance. If a minimum is found, we stop. If not, we restart the fitting process from $0.5\lambda_{\text{start}} < \lambda < 0.7\lambda_{\text{start}}$, and so on until either $10^{-3}\lambda_{\text{start}}$ is reached or a minimum of the cross-validated deviance is found. All information related to fitting is saved in a structure after each block; hence, restarting an optimization has little overhead. We determine that a minimum has been found by asking that the cross-validated deviance at the minimum λ probed so far is higher than the minimum cross-validated deviance by a certain number of units, by default 10. Finally, we fit the full model down to the minimum λ found by cross-validation.

Our algorithm succeeds in finding the first non-shallow local minimum $D^{\text{CV}}(\lambda)$ between λ_{start} and $10^{-3}\lambda_{\text{start}}$. While we cannot rule out the possibility that $D^{\text{CV}}(\lambda)$ has several non-shallow local minima, this does not appear to be an issue in practice; $D^{\text{CV}}(\lambda)$ is typically quite smooth and almost quadratic in shape near its minimum, and we have not observed non-shallow local minima in any of the fits performed for the purposes of this article. In the pathological scenario where $D^{\text{CV}}(\lambda)$ has several non-shallow local minima, the cross-validation algorithm will select the one associated with the largest λ, or the largest level of regularization, which we consider to be a safe fallback.

Of crucial importance in cross-validation is that different values of λ across different folds correspond to the same level of regularization. Tibshirani (1996) and Park and Hastie (2007) suggest that corresponding regularization levels are found when the ratio $\lambda/\lambda_{\max}^{\text{fold}}$ is the same across folds. We thus perform cross-validation to find the optimal ratio $\lambda/\lambda_{\max}^{\text{fold}}$ rather than λ. The regularization paths are sampled differently across each fold. We solved this issue by linearly interpolating cross-validated deviance values at all $\lambda/\lambda_{\max}^{\text{fold}}$ used by a fold.

Complexity analysis and memory and time requirements

Computing ∇f, a $O(mn)$ operation, typically accounts for 50%–90% of the time spent during optimization; roughly 5% is accounted for by miscellaneous overhead; and the remainder time is spent in the line search. The ratio of inner iterations to outer iterations varies with λ, being typically equal to 1 for large λ and 2–4 for small λ. Given the algorithm for determining the next λ and assuming we evaluate the model at λ values from $0.001\lambda_{\text{start}}$ to λ_{start}, the number of outer iterations is bounded above by $\log(0.001)/\log(0.98) \approx 340$. Given typical ratios of inner iterations to outer iterations, this might come out to about 600 evaluations of ∇f. This number can be cut down by a factor 2 or 3 by stopping iterations when λ reaches beyond λ_{optimal} determined by cross-validation; our software includes a cross-validation implementation, outlined above, that accomplishes this. In total, perhaps 300 evaluations of ∇f may be required in a typical application; with overhead this may balloon up to a cost equivalent to computing ∇f 500–600 times, multiplied by $(k + 1)$ during k-fold cross-validation. This compares favorably and in some cases may be appreciably

faster than boosting where the cost of each iteration is equal to the cost of computing ∇f, in addition to some overhead such as computing f and performing a line search in some variants (Buhlmann & Hothorn, 2007).

Memory requirements are typically equal to the cost of holding two or three copies of the design matrix $\mathbf{X}$ in memory. With a 64-bit version of Windows or Linux and 4 GBs of RAM, one can thus expect to able work with design matrices as large as 25,000 × 5,000 (1 GB in memory) before running into out-of-memory errors. For the largest design matrices tested here, which are of size 10,000 × 256, optimization including 5-fold cross-validation takes about 40 seconds on our test computer running on 4 Intel Xeon CPUs running at 2 Ghz and 3 GB of RAM.

Software

Our software package, implemented in Matlab, includes two main functions, with function signatures:

```
[thefit]
=glmfitsparseprior(y,X,U,stopcrit,varargin)
[thefit]
= cvglmfitsparseprior(y,X,U,folds,varargin).
```
(A17)

Here $\mathbf{y}$, $\mathbf{X}$, and $\mathbf{U}$ are response and design matrices and `stopcrit` is the fraction of λ_{start} after which to stop optimization. `folds` is a matrix of booleans which defines which portion of the data is assigned to the fit set versus the validation set for each cross-validation fold. Such a matrix may be generated by the included auxiliary function `getcvfolds`. A structure is returned in both cases which contains the entire regularization path, deviances, AIC values, the values of λ used, and, if applicable, cross-validated deviances and $\mathbf{w}$ and $\mathbf{u}$ at the optimal value of λ. In addition to the binomial/logit model presented here, the software supports two other GLMs: a Poisson model with exponential inverse link, for count data, e.g., binned spike trains, and a Gaussian noise model with identity link, e.g., least-squares. These latter may be invoked through supplementary arguments whose calling sequence is available from the Matlab command-line help.

Appendix B

Inference for the GLM

Goodness-of-fit is assessed in GLMs through the scaled deviance of a fitted model (Wood, 2006):

$$D^* = 2(L - L_{\max}). \tag{B1}$$

The deviance *proper* is defined as $D = D^*\phi$, where ϕ is a scale parameter that depends on the noise distribution used. For the binomial distribution, $\phi = 1$ (Wood, 2006); hence, we use the terms deviance and scaled deviance interchangeably. L is the negative log-likelihood of the fitted model, while $L_{\max}$ is the negative log-likelihood for a saturated model which contains one parameter per data point y_i. For the logistic regression model used here, $L_{\max} = 0$, although this is not always the case, for example with Poisson regression. By construction, $D \geq 0$; a small D implies a good fit to the data while a large D implies a poor fit.

Standard linear regression with Gaussian residuals can be cast as a special case of a GLM (Wood, 2006), for which the deviance is given by

$$D = \sum_{ij} (y_i - X_{ij}\widehat{w_j})^2. \tag{B2}$$

Here $\hat{\mathbf{w}}$ is the estimated weight vector. This expression comes from the fact that the negative log of a Gaussian is a sum-of-squares. It is thus helpful to think of deviance as analogous to the residual the sum-of-squares in linear regression. Like the residual sum-of-squares, the deviance always becomes smaller as predictors are added to the design matrix; hence, it is not useful by itself for model selection purposes. Rather, the deviance is used as a basis for model-complexity independent measures of goodness-of-fit, such as the cross-validated deviance, the validated deviance, and the Akaike Information Criterion. The deviance may also be used for classical hypothesis testing in log-likelihood ratio tests.

Cross-validation and validation

As explained in the Methods section, k-fold cross-validation works by splitting the data into k randomly chosen nonoverlapping partitions of equal size, fitting the model with data in all but the partition and computing the likelihood or deviance of the data in the ith partition, and repeating the process for $i = 1 \ldots k$. This yields a cross-validated deviance score D^{CV}. D^{CV} is a measure of how well a model predicts out-of-sample observations and therefore estimates its generalization performance.

How do we interpret D^{CV} measured for a model $\mathcal{M}_j$? By Bayes' theorem, the probability of the model given the data $\mathbf{y}$ is given by

$$p(\mathcal{M}_j|\mathbf{y}) \propto p(\mathbf{y}|\mathcal{M}_j)p(\mathcal{M}_j). \tag{B3}$$

Assuming a flat prior on models $p(\mathcal{M}_j) = 1$ for all j, we have:

$$p(\mathcal{M}_j|\mathbf{y}) \propto p(\mathbf{y}|\mathcal{M}_j). \tag{B4}$$

The probability of a data set $\mathbf{y}$ under the model is approximately given by cross-validated likelihood of the data (Geisser, 1975; Geisser & Eddy, 1979). Hence, the relative likelihood of two models is approximately given by

$$PSBF(\mathcal{M}_1, \mathcal{M}_2) = \frac{p(\mathcal{M}_1|\mathbf{y})}{p(\mathcal{M}_2|\mathbf{y})} \approx \exp\left(\frac{1}{2}\left(D_2^{CV} - D_1^{CV}\right)\right). \tag{B5}$$

This has the same form as a Bayes factor, hence the name Pseudo-Bayes factor (PSBF) (Gelfand & Dey, 1994). Bayes factors are a common technique of model comparison. An often-used interpretation scale for Bayes factors states that $PSBF\ (\mathcal{M}_1,\ \mathcal{M}_2) > 150$, corresponding to $D_2^{\mathrm{CV}} - D_1^{\mathrm{CV}} > 10$ is "very strong" (Kass & Raftery, 1995) or "decisive" (Jeffreys, 1961) evidence for $\mathcal{M}_1$ over $\mathcal{M}_2$.

The variance of D^{CV} can be large; its value depends on the exact choice of folds. Therefore, when differences in the cross-validated deviances between two models are relatively small, for example less than 10 units, we do not recommend concluding that either model is better based on these scores; lower variance estimates of D^{CV} can be obtained by repeatedly computing D^{CV} for different folds.

In the same spirit, the deviance of a model based on predictions on a hold-out validation set can be useful to assess model performance. However, since binomial data are quite noisy, much validation data must be collected to obtain reliable results.

AIC and log-likelihood ratio tests

The effective degrees of freedom for the MAP estimate of a GLM with a sparse prior is defined as (Park & Hastie, 2007; Hastie, 2007; Zou, Hastie, & Tibshirani, 2007):

$$df = |\{k|\mathbf{w}_k \neq 0\}|. \tag{B6}$$

That is, it is equal to the number of nonzero elements in the estimated weight vector $\mathbf{w}$. For a model with a Gaussian prior, on the other hand (Wood, 2006):

$$df = tr((\mathbf{H} + \lambda \mathbf{A}^T\mathbf{A})^{-1}\mathbf{H}). \tag{B7}$$

Here $\mathbf{H}$ denotes the Hessian of the negative log-likelihood and *tr* the trace. Note that this reduces to n, the number of parameters in the model, as $\lambda \to 0$. Given df for the GLM with either a sparse or Gaussian prior, we may define an AIC (Akaike Information Criterion) analogue (Wood, 2006):

$$\mathrm{AIC} = D + 2\,df. \tag{B8}$$

For the purposes of hyperparameter selection in a sparse GLM, we do not recommend the use of the AIC, as its integer-valued nature means AIC(λ) is discontinuous and has several shallow minima. However, it may be useful in comparing our results with other GLMs with binomial endpoints that, for practical reasons, avoid cross-validation (Knoblauch & Maloney, 2008b; Ross & Cohen, 2009). We note that the AIC is equivalent, up to a linear transformation, to the unbiased risk estimator (UBRE) used in Knoblauch and Maloney (2008b). As with cross-validated deviance, a lower AIC is better, and large differences in AIC values between two models are interpreted as support for the model with the lower AIC. Since the AIC is based on asymptotic results, its validity is dubious when low numbers of trials are used. In the case where the AIC and the cross-validated deviance point towards incompatible conclusions, we recommend averaging cross-validated deviance over several different random choices of folds and base conclusions on this low-variance cross-validated deviance estimate.

It has been shown (Wood, 2006) that for a simple GLM model $\mathcal{M}_1$ nested inside a more complex GLM model $\mathcal{M}_2$:

$$D_1 - D_2 \sim \chi^2_{df_2 - df_1}. \tag{B9}$$

A p-value may be obtained from this expression. This is known as a log-likelihood ratio test. Again, the χ^2 approximation is based on asymptotic results and has dubious validity for a small number of trials (Wood, 2006). It is also important to keep in mind that the test is only valid for *nested* models. We recommend the use of log-likelihood ratio tests when the importance of a single predictor or ensemble of related predictors are in question. For nonnested models, Vuong's test may be used (Vuong, 1989).

Acknowledgments

We would like to thank two anonymous reviewers for their insightful comments. Simon Barthelmé wishes to thank Pascal Mamassian for his support. This work was supported by grants from the CIHR (MOP-79352) and Le ministre du Développement économique, de l'Innovation et de l'Exportation du Québec.

Commercial relationships: none.
Corresponding author: Patrick Mineault.
Email: patrick.mineault@mail.mcgill.ca.
Address: 3801 University Street #896, Montreal, Quebec H3A 2B4, Canada.

References

Abbey, C. K., & Eckstein, M. P. (2001). Maximum-likelihood and maximum-a-posteriori estimates of human-observer templates. In E. A. Krupinski & D. P. Chakraborty (Eds.), *Proceedings of SPIE* (vol. 4324, pp. 114–122).

Abbey, C. K., & Eckstein, M. P. (2002). Classification image analysis: Estimation and statistical inference for two-alternative forced-choice experiments. *Journal of Vision, 2*(1):5, 66–78, http://journalofvision.org/2/1/5/, doi:10.1167/2.1.5. [PubMed] [Article]

Abbey, C. K., Eckstein, M. P., Shimozaki, S. S., Baydush, A. H., Catarious, D. M., & Floyd, C. E. (2002). Human-observer templates for detection of a simulated lesion in mammographic images. In D. P. Chakraborty & E. A. Krupinski (Eds.), *Proceedings of SPIE* (vol. 4686, pp. 25–36).

Ahrens, M., Paninski, L., & Sahani, M. (2008). Inferring input nonlinearities in neural encoding models. *Network: Computation in Neural Systems, 19,* 35–67. [PubMed]

Ahrens, M. B., Linden, J. F., & Sahani, M. (2008). Nonlinearities and contextual influences in auditory cortical responses modeled with multilinear spectrotemporal methods. *Journal of Neuroscience, 28,* 1929–1942. [PubMed] [Article]

Ahumada, A. A., & Lovell, J. (1971). Stimulus features in signal detection. *Journal of the Acoustical Society of America, 49,* 1751–1756.

Ahumada, A. J. (1996). Perceptual classification images from Vernier acuity masked by noise. *Perception, 26,* 18.

Ahumada, A. J. (2002). Classification image weights and internal noise level estimation. *Journal of Vision, 2*(1):8, 121–131, http://journalofvision.org/2/1/8/, doi:10.1167/2.1.8. [PubMed] [Article]

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer.

Buhlmann, P., & Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science, 22,* 477.

Burt, P., & Adelson, E. (1983). The Laplacian Pyramid as a compact image code. *IEEE Transactions on Communications, 31,* 532–540.

Chauvin, A., Worsley, K. J., Schyns, P. G., Arguin, M., & Gosselin, F. (2005). Accurate statistical tests for smooth classification images. *Journal of Vision, 5*(9):1, 659–667, http://journalofvision.org/5/9/1/, doi:10.1167/5.9.1. [PubMed] [Article]

David, S. V., Mesgarani, N., & Shamma, S. A. (2007). Estimating sparse spectro-temporal receptive fields with natural stimuli. *Network, 18,* 191–212. [PubMed]

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association, 70,* 320–328.

Geisser, S., & Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association, 74,* 153–160.

Gelfand, A. E., & Dey, D. K. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society B, Statistical Methodology, 56,* 501–514.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall.

Gold, J. M., Murray, R. F., Bennett, P. J., & Sekuler, A. B. (2000). Deriving behavioural receptive fields for visually completed contours. *Current Biology, 10,* 663–666. [PubMed]

Hale, E. T., Yin, W., & Zhang, Y. (2007). *A fixed-point continuation method for L1-regularized minimization with applications to compressed sensing* (vol. TR07-07; Tech. Rep.). Rice University.

Hastie, T. (2007). Comment: Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science, 22,* 513.

Jeffreys, H. (1961). *Theory of probability*. Oxford: Oxford University Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90,* 773–795.

Knoblauch, K., & Maloney, L. (2008a). Classification images estimated by generalized additive models [Abstract]. *Journal of Vision, 8*(6):344, 344a, http://journalofvision.org/8/6/344/, doi:10.1167/8.6.344.

Knoblauch, K., & Maloney, L. (2008b). Estimating classification images with generalized linear and additive models. *Journal of Vision, 8*(16):10, 1–19, http://journalofvision.org/8/16/10/, doi:10.1167/8.16.10. [PubMed] [Article]

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., & Baker, C. I. (2009). Circular analysis in systems neuroscience: The dangers of double dipping. *Nature Neuroscience, 12,* 535–540. [PubMed]

Lee, T.-W., Wachtler, T., & Sejnowski, T. J. (2002). Color opponency is an efficient representation of spectral properties in natural scenes. *Vision Research, 42,* 2095–2103. [PubMed]

Levi, D. M., & Klein, S. A. (2002). Classification images for detection and position discrimination in the fovea and parafovea. *Journal of Vision, 2*(1):4, 46–65, http://journalofvision.org/2/1/4/, doi:10.1167/2.1.4. [PubMed] [Article]

Mangini, M. C., & Biederman, I. (2004). Making the ineffable explicit: Estimating the information

employed for face classifications. *Cognitive Science, 28,* 209–226.

Mineault, P. J., & Pack, C. C. (2008). Getting the most out of classification images [Abstract]. *Journal of Vision, 8*(6):271, 271a, http://journalofvision.org/8/6/271/, doi:10.1167/8.6.271.

Murray, R. F., Bennett, P. J., & Sekuler, A. B. (2002). Optimal methods for calculating classification images: Weighted sums. *Journal of Vision, 2*(1):6, 79–104, http://journalofvision.org/2/1/6/, doi:10.1167/2.1.6. [PubMed] [Article]

Neri, P., & Levi, D. (2006). Receptive versus perceptive fields from the reverse-correlation viewpoint. *Vision Research, 46,* 2465–2474. [PubMed]

Neri, P., & Levi, D. (2008). Temporal dynamics of directional selectivity in human vision. *Journal of Vision, 8*(1):22, 1–11, http://journalofvision.org/8/1/22/, doi:10.1167/8.1.22. [PubMed] [Article]

Neri, P., Parker, A. J., & Blakemore, C. (1999). Probing the human stereoscopic system with reverse correlation. *Nature, 401,* 695–698. [PubMed]

Olman, C., & Kersten, D. (2004). Classification objects, ideal observers & generative models. *Cognitive Science: A Multidisciplinary Journal, 28,* 227–239.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature, 381,* 607–609. [PubMed]

Olshausen, B. A., & Field, D. J. (2004). Sparse coding of sensory inputs. *Current Opinion in Neurobiology, 14,* 481–487. [PubMed]

Paninski, L. (2005). Asymptotic theory of information-theoretic experimental design. *Neural Computation, 17,* 1480–1507. [PubMed]

Park, M. Y., & Hastie, T. (2007). L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society B, Statistical Methodology, 69,* 659–677.

Pillow, J. W., Shlens, J., Paninski, L., Sher, A., Litke, A. M., Chichilnisky, E. J., et al. (2008). Spatio-temporal correlations and visual signaling in a complete neuronal population. *Nature, 454,* 995–999. [PubMed] [Article]

Portilla, J., Strela, V., Wainwright, M., & Simoncelli, E. (2003). Image denoising using scale mixtures of gaussians in the wavelet domain. *IEEE Transactions on Image Processing, 12,* 1338–1351. [PubMed]

Rajashekar, U., Bovik, A. C., & Cormack, L. K. (2006). Visual search in noise: Revealing the influence of structural cues by gaze-contingent classification image analysis. *Journal of Vision, 6*(4):7, 379–386, http://journalofvision.org/6/4/7/, doi:10.1167/6.4.7. [PubMed] [Article]

Ross, M. G., & Cohen, A. L. (2009). Using graphical models to infer multiple visual classification features [Abstract]. *Journal of Vision, 9*(3):33, 1–24, http://journalofvision.org/9/3/23/, doi:10.1167/9.3.23.

Rosset, S., Zhu, J., Hastie, T., & Schapire, R. (2004). Boosting as a regularized path to a maximum margin classifier. *Journal of Machine Learning Research, 5,* 941–973.

Rust, N. C., Mante, V., Simoncelli, E. P., & Movshon, J. A. (2006). How MT cells analyze the motion of visual patterns. *Nature Neuroscience, 9,* 1421–1431. [PubMed]

Sahani, M., & Linden, J. F. (2003). Evidence optimization techniques for estimating stimulus-response functions. In K. O. S. Becker & S. Thrun (Eds.), *Advances in neural information processing systems* (vol. 15, pp. 317–324). Cambridge, MA: MIT Press.

Seeger, M., Gerwinn, S., & Bethge, M. (2007). Bayesian inference for sparse generalized linear models. In J. N. K. et al. (Ed.), *ECML 2007* (vol. 4701, pp. 298–309).

Seeger, M. W. (2008). Bayesian inference and optimal design for the sparse linear model. *Journal of Machine Learning Research, 9,* 759–813.

Sekuler, A. B., Gaspar, C. M., Gold, J. M., & Bennett, P. J. (2004). Inversion leads to quantitative, not qualitative, changes in face processing. *Current Biology, 14,* 391–396. [PubMed]

Selesnick, I. W., Baraniuk, R. G., & Kingsbury, N. C. (2005). The dual-tree complex wavelet transform. *IEEE Signal Processing Magazine, 22,* 123–151.

Simoncelli, E. P., & Freeman, W. T. (1995). The steerable pyramid: A flexible architecture for multi-scale derivative computation. In *Proceedings of the 1995 International Conference on Image Processing* (vol. 3). IEEE Computer Society.

Simoncelli, E. P., Pillow, J., Paninski, L., & Schwartz, O. (2004). Characterization of neural responses with stochastic stimuli. In M. Gazzaniga (Ed.), *The cognitive neurosciences* (3rd ed., pp. 327–338). Cambridge, MA: MIT Press.

Solomon, J. A. (2002). Noise reveals visual mechanisms of detection and discrimination. *Journal of Vision, 2*(1):7, 105–120, http://journalofvision.org/2/1/7/, doi:10.1167/2.1.7. [PubMed] [Article]

Srivastava, A., Lee, A. B., Simoncelli, E. P., & Zhu, S.-C. (2003). On advances in statistical modeling of natural images. *Journal of Mathematical Imaging and Vision, 18,* 17–33.

Tadin, D., Lappin, J. S., & Blake, R. (2006). Fine temporal properties of center-surround interactions in motion revealed by reverse correlation. *Journal of Neuroscience, 26,* 2614–2622. [PubMed] [Article]

Thomas, J. P., & Knoblauch, K. (2005). Frequency and phase contributions to the detection of temporal luminance modulation. *Journal of the Optical Society of America A, Optics, Image Science, and Vision, 22,* 2257–2261. [PubMed] [Article]

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B, Statistical Methodology, 58,* 267–288.

Tjan, B. S., & Nandy, A. S. (2006). Classification images with uncertainty. *Journal of Vision, 6*(4):8, 387–413, http://journalofvision.org/6/4/8/, doi:10.1167/6.4.8. [PubMed] [Article]

Victor, J. (2005). Analyzing receptive fields, classification images and functional images: Challenges with opportunities for synergy. *Nature Neuroscience, 8,* 1651–1656. [PubMed] [Article]

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica, 57,* 307–333.

Wipf, D., & Nagarajan, S. (2008). A new view of automatic relevance determination. In J. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Advances in neural information processing systems 2008* (pp. 1625–1632). Cambridge, MA: MIT Press.

Wood, S. N. (2006). *Generalized additive models: An introduction with R.* Boca Raton, FL: Chapman & Hall/CRC.

Worsley, K. J., Marrett, S., Neelin, P., & Evans, A. C. (1996). Searching scale space for activation in pet images. *Human Brain Mapping, 4,* 74–90.

Wu, M. C.-K., David, S. V., & Gallant, J. L. (2006). Complete functional characterization of sensory neurons by system identification. *Annual Review of Neuroscience, 29,* 477–505. [PubMed]

Zou, H., Hastie, T., & Tibshirani, R. (2007). On the "degrees of freedom" of the lasso. *Annals of Statistics, 35,* 2173–2192.