NeuroImage 56 (2011) 570–581

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Gaussian process methods for estimating cortical maps ☆

Jakob H. Macke [a,b,c,*], Sebastian Gerwinn [b,c], Leonard E. White [d], Matthias Kaschube [e], Matthias Bethge [b,c]

[a] Gatsby Computational Neuroscience Unit, University College London, Alexandra House, 17 Queen Square, London WC1N 3AR, UK
[b] MPI for Biological Cybernetics, Computational Vision and Neuroscience Group, Spemannstraße 41, 72076 Tübingen, Germany
[c] Werner Reichardt Centre for Integrative Neuroscience, University of Tübingen, Germany
[d] Duke Institute for Brain Sciences, Duke University, Durham, NC 27705, USA
[e] Lewis-Sigler Institute for Integrative Genomics and Department of Physics, Princeton University, Princeton, NJ 08544, USA

ARTICLE INFO

Article history:
Received 1 February 2010
Revised 26 April 2010
Accepted 30 April 2010
Available online 15 May 2010

Keywords:
Optical imaging
Functional imaging
Gaussian process
Bayesian statistics
Orientation preference map
Cortical map
Optimal smoothing
Visual cortex
Noise-correlations
Decoding

ABSTRACT

A striking feature of cortical organization is that the encoding of many stimulus features, for example orientation or direction selectivity, is arranged into topographic maps. Functional imaging methods such as optical imaging of intrinsic signals, voltage sensitive dye imaging or functional magnetic resonance imaging are important tools for studying the structure of cortical maps. As functional imaging measurements are usually noisy, statistical processing of the data is necessary to extract maps from the imaging data. We here present a probabilistic model of functional imaging data based on Gaussian processes. In comparison to conventional approaches, our model yields superior estimates of cortical maps from smaller amounts of data. In addition, we obtain quantitative uncertainty estimates, i.e. error bars on properties of the estimated map. We use our probabilistic model to study the coding properties of the map and the role of noise-correlations by decoding the stimulus from single trials of an imaging experiment.

Introduction

One of the characteristic features of the visual cortex in primates and other mammalian species is the spatial arrangement of neurons into functionally defined topographic maps. Map-like arrangements have been found for orientation preference (Blasdel and Salama, 1986; Bonhoeffer and Grinvald, 1991; Ohki et al., 2005), direction of motion preference (Weliky et al., 1996), ocular dominance (Hübener et al., 1997; LeVay et al., 1978), spatial frequency preference (Hübener et al., 1997; Issa et al., 2000; Sirovich and Uglesich, 2004) as well as binocular disparity (Kara and Boyd, 2009). The development of cortical maps (Goodhill, 2007), their statistical structure (Chklovskii and Koulakov, 2004; Hunt et al., 2009; Wolf and Geisel, 1998) and relationships between different maps (Hübener et al., 1997; Issa et al., 2008; Kara and Boyd, 2009) have been the focus of extensive research.

The classical method for estimating functional maps is optical imaging of intrinsic signals. More recently, voltage sensitive dye imaging, functional magnetic resonance imaging (Fukuda et al., 2006; Yacoub et al., 2008), or 2-photon calcium imaging (Li et al., 2008; Ohki et al., 2005) have been used to measure maps at higher spatial or temporal resolution, or non-invasively. For most of these methods the signal-to-noise ratio is low, i.e. the modulation of the response by the stimulus is small compared to non-specific background fluctuations. Therefore, statistical pre-processing methods play an important role for extracting topographic maps from the raw experimental data. The classical, and still most commonly used method for analyzing intrinsic signaling data is to simply average the data within each stimulus condition, and report differences between conditions. In the case of orientation preference maps (OPMs), this amounts to estimating the preferred orientation at each pixel by vector averaging the different stimulus orientations weighted according to the evoked responses. In a second step, spatial bandpass filtering is usually applied.

One disadvantage of this approach is that the frequency characteristics of the bandpass filters are free parameters which are often set ad-hoc, and may have a substantial impact on the statistics of the obtained map (Polimeni et al., 2005). In addition, the approach ignores the effect of anisotropic and correlated noise (Polimeni et al., 2005; Yokoo et al., 2001), which may result in artifacts (Sirovich and Uglesich, 2004). Analysis based on principal component analysis

☆ A brief, preliminary report of portions of this research was presented at the conference on Neural Information Processing Systems (NIPS), and has appeared as 'Bayesian estimation of orientation preference maps' in the conference proceedings 'Advances in Neural Processing Systems, 22, 2010'.

* Corresponding author. Gatsby Computational Neuroscience Unit, University College London, Alexandra House, 17 Queen Square, London WC1N 3AR, UK.
E-mail addresses: jakob@gatsby.ucl.ac.uk (J.H. Macke), sgerwinn@tuebingen.mpg.de (S. Gerwinn), white033@mc.duke.edu (L.E. White), kaschube@princeton.edu (M. Kaschube), mbethge@tuebingen.mpg.de (M. Bethge).

1053-8119/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2010.04.272

(Sirovich and Everson, 1992), linear discriminant analysis and oriented PCA (Everson et al., 1997) (and extensions thereof (Gabbay et al., 2000; Yokoo et al., 2001)) as well as variants of independent component analysis (Stetter et al., 2000) have been developed to overcome these limitations. These methods have in common that they do not make specific parametric assumptions about the relationship between stimulus and response or between different stimuli. Rather, maps are defined by filters which are maximally discriminative between different stimulus conditions. They differ from the classical approach in that they do not assume the noise to be isotropic and uncorrelated. However, as they do not include an explicit model of how stimuli are related to cortical activity, they make it hard to incorporate prior knowledge about the structure of maps, and can therefore be data-intensive. As optical imaging is inherently noisy, it is very important to have means of assessing how well map properties are constrained by the data, e.g. the location or number of pinwheel estimated. The above-mentioned models do not aim to estimate uncertainty directly.

Bayesian methods (Friston, 2002; Penny et al., 2005) attempt to combine the strengths of the classical and discriminative models by combining prior knowledge about the statistical spatial structure of the data with flexible noise models. Here, we describe a probabilistic model for functional imaging data which can be used both for the estimation of cortical maps and for decoding studies. Specifically, we will employ Gaussian process (GP) methods (Rasmussen and Williams, 2006) for modeling functional imaging data and estimating cortical maps. Prior knowledge about the statistical structure of OPMs is encoded in the covariance function of a Gaussian process prior over maps. By combining the prior with the data through an explicit generative model of the measurement process, we obtain a posterior distribution over maps.

Compared to conventional smoothing-methods for analyzing functional imaging data, this GP-approach has a number of advantages. First, the mean of the posterior distribution can be interpreted as an optimally smoothed map. In contrast to usual filtering approaches, the shape of the filter has a clear probabilistic interpretation. The filtering is adaptive and adjusts to the amount and quality of the data observed at any particular location. Second, our model allows us to model noise with non-constant variances across the cortical surface, as well as correlations across pixels. In intrinsic signal imaging data, both noise-variances and noise-correlations vary strongly across the cortical surface, in a manner which depends on layout of blood vessels across the cortical surface. Therefore, it is important to take these dependencies into account.

Third, the model returns an estimate of the map-parameters at any location, not only at those at which measurements were obtained. This allows one to estimate maps from multi-electrode recordings (Swindale et al., 1987), or for artifact removal. Fourth, the posterior variances at each pixel can be used to compute point-wise error bars at each pixel location (Stetter et al., 2000; Yokoo et al., 2001). By sampling from the posterior (using the full posterior covariance), we can also get error bars on topological or global properties of the map, such as pinwheel counts or locations. Finally, the use of an explicit, generative model of the data facilitates both the interpretation and setting of parameters and quantitative model comparisons. In addition, the model can be used to study the information content of the map by using it to decode stimuli from the experimental measurements.

Mathematically speaking, estimating an orientation preference map is equivalent to estimating a vector field (the 2-dimensional vector encoding preferred orientation) across the cortical surface from noisy measurements. Related problems have been studied in spatial statistics, e.g. in the estimation of wind-fields in geo-statistics (Cornford et al., 2004; Csató et al., 2001), where Gaussian process methods for this problem are often referred to as *co-kriging* methods (Cressie, 1992; Cressie and Johannesson, 2008). In machine learning, related problems have been studied in the context of *multi-task learning* (Lawrence and Platt, 2004) or multi-output regression (Teh et al., 2005). Matlab-code for implementing the statistical methods described in the article is available for download at http://www.kyb.mpg.de/bethge/code/gpmaps.

Methods

Encoding model: relating stimulus to response

We model an imaging experiment, where at each of N trials, the activity at n pixels (or voxels) is measured simultaneously. Under the assumption of cosine-tuning curves, the expected measurement at location x in response to a stimulus with orientation θ_i can be written as $\hat{r}_i(x) = A(x)\cos(2\theta_i - \psi(x)) + c(x)$ (Swindale, 1998). Here, $\psi(x)$ is the preferred orientation of pixel x, $A(x)$ is a measure of the tuning strength, and $c(x)$ the mean activity of this pixel. Equivalently, this can be written as $\hat{r}_i(x) = m_1(x)\cos(2\theta_i) + m_2(x)\sin(2\theta_i) + m_3(x)$. Then, the preferred orientation is the argument of the complex number $m'(x) := m_1(x) + \mathbf{i}m_2(x)$, and $A(x)$ is its modulus.

More generally, we relate the response $r_i(x)$ at trial i to a stimulus parameterized by the vector $\mathbf{v}_i$ by

$$r_i(x) = \sum_{k=1}^{d} v_{ki} m_k(x) + \epsilon_i(x) = \mathbf{v}_i^\top \mathbf{m}(x) + \epsilon_i(x). \quad (1)$$

That is, the mean response at each pixel is modeled to be a linear function of some stimulus parameters $\mathbf{v}_i$. For estimating orientation preference maps with cosine-tuning curves we would have $\mathbf{v}_i^\top = (\cos(2\theta_i), \sin(2\theta_i), 1)$, and $d = 3$. For estimating both orientation and direction preference maps, an appropriate parameterization would be $\mathbf{v}_i^\top = (\cos(2\theta_i), \sin(2\theta_i), \cos(\theta_i), \sin(\theta_i), 1)$. While this parameterization assumes cosine-tuning curves, it can be generalized by including more basis functions, e.g. sines and cosines with higher frequencies. In this way, arbitrary tuning functions, e.g. the commonly used Gaussian tuning curves, can be approximated. This is especially useful for modeling experiments in which many different stimulus orientations are used, and the shape of tuning curves can be measured at a high resolution (Swindale, 1998). We refer to the coefficients $m_k(x)$ as *map components* which indicate the selectivity of pixel x to stimulus feature k. The term $\epsilon_i(x)$ models the noise in pixel x on trial i. The noise is taken to be Gaussian distributed with covariance Σ_ϵ. $\Sigma_\epsilon(x,x)$ is the noise-variance of pixel x, whereas $\Sigma_\epsilon(x,x')$ corresponds to noise covariations between x and x'.

Formula (1) can be written compactly as $\mathbf{r}_i = M\mathbf{v}_i + \varepsilon_i$, where M is an $n \times d$-dimensional matrix, and $\mathbf{r}_i$ and ε_i are n dimensional vectors. As we want to define Gaussian priors over M, it is notationally more convenient to reshape the matrix as a vector $\mathbf{m}$ of dimensionality nd, $\mathbf{m} = \mathrm{vec}(M)$. Then, stimulus and response are related by $\mathbf{r}_i = V_i^\top \mathbf{m} + \varepsilon_i$, where $V_i = v_i \otimes \mathbb{I}_n$, and $\otimes$ denotes the Kronecker-product.

Specifying a prior distribution over maps

We specify the prior distribution over maps as a d-dimensional Gaussian process. That is, we assume each of the d map components to be drawn from a multivariate Gaussian distribution. Such a Gaussian process is defined by its mean function $\mu_{\mathrm{prior}}(x, k)$, and a covariance function $K_{\mathrm{prior}}(x, x', k, l)$. The prior mean $\mu_{\mathrm{prior}}(x, k)$ is the expected value of the kth map component on pixel x. As we assume that all orientations are equally likely for each pixel, we set $\mu_{\mathrm{prior}}(x, k) = 0$ for all x and k. The covariance function K_{prior} encodes our assumptions about correlations of the map across pixels or map components, $K_{\mathrm{prior}}(x, x', k, l) = \mathrm{Cov}_{\mathrm{prior}}(m_k(x), m_l(x'))$. In our setting, all prior assumptions about the structure of the maps are encoded in the specific choice of K_{prior}.

The choice of covariance function should be guided by our prior assumptions, or ideally knowledge, about the structure of the map of

interest. Cortical maps, and in particular orientation preference maps, have been studied extensively in the past (Issa et al., 2008), and we can use these studies to guide the choice of the prior. It is known that orientation preference maps are smooth (Ohki et al., 2005) and that they have a semi-periodic structure of regularly spaced columns. Hence, filtering white noise with appropriately chosen filters (Rojer and Schwartz, 1990) yields maps which resemble measured OPMs (see Fig. 1). While it has been suggested that real OPMs differ from Gaussian random fields in their higher-order statistics (Erwin et al., 1995; Wolf and Geisel, 1998), use of a Gaussian prior can be motivated by the maximum entropy principle: we assume a prior with minimal higher-order correlations, with the goal of inferring them from the experimental data (Wolf and Geisel, 1998). For real orientation preference maps, there is a slight anisotropy (Coppola et al., 1998). However, for simplicity, we take the prior to be isotropic, i.e. not to favor any orientation over others.

Concretely, we generate prior samples by convolving two Gaussian white-noises image with Mexican Hat Filters constructed using a Difference-of-Gaussians: $f(x) = \sum_{k=1}^{2} \frac{\alpha_k}{2\pi\sigma_k^2} \exp\left(-\frac{1}{2}\frac{\|x\|^2}{\sigma_k^2}\right)$. We set $\alpha_1 = -\alpha_2$ to make sure that the filter is balanced, i.e. has zero mean. We fixed $\sigma_2 = 2\sigma_1$ to eliminate one hyper-parameter, and to avoid numerical degeneracies which can arise when $\sigma_1 \approx \sigma_2$.

Then, the map can be constructed by taking the filtered images as the real and imaginary parts (see Fig. 1A). This will result in a prior which is uncorrelated in the different map components, $K_{\text{prior}}(x, x', 1, 2) = \text{Cov}_{\text{prior}}(m_1(x), m_2(x')) = 0$, and a stationary covariance function which is given by

$$K_c(\tau) = K_c\left(\|x - x'\|\right) = K_{\text{prior}}(x, x', k, k) = \sum_{a,b=1}^{2} \frac{\alpha_a \alpha_b}{2\pi(\sigma_a^2 + \sigma_b^2)} \exp\left(-\frac{1}{2}\left(\frac{\tau^2}{\sigma_a^2 + \sigma_b^2}\right)\right), \quad (2)$$

with $\tau = \|x - x'\|$. The prior covariance function can be used to calculate the prior covariance matrix $K_{\mathbf{m}} = E_{\text{prior}}(\mathbf{m}\mathbf{m}^\top)$. If we assume different map components to be uncorrelated, and the same covariance for each component, the prior covariance has block structure, and can be written as $K_{\mathbf{m}} = \mathbb{I}_d \otimes K$ (Penny et al., 2005), where the matrix $K = E_{\text{prior}}(\mathbf{m}_1\mathbf{m}_1^\top)$ is the covariance of the first map component.

This prior has two hyper-parameters, namely the absolute magnitude α_1 and the kernel width σ_1. In principle, optimization of the marginal likelihood (Lawrence and Platt, 2004; Rasmussen and Williams, 2006) can be used to set hyper-parameters. In practice, it was computationally more efficient to select hyper-parameters by matching the radial component of the empirically observed auto-correlation function of the map, similar to 'variogram matching' techniques employed in geo-statistics (Cressie, 1992), see Fig. 1.

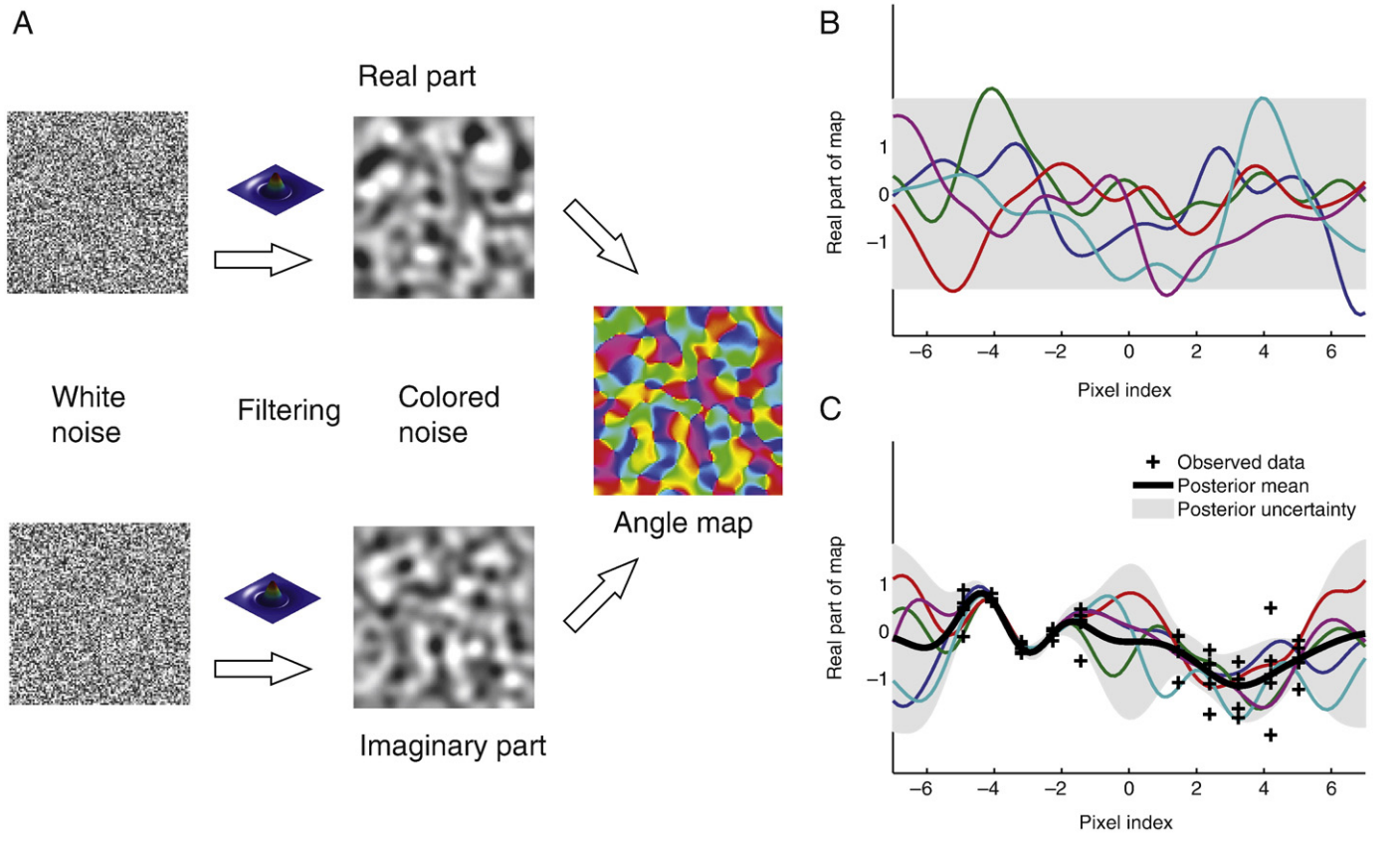

Fig. 1. Illustration of the Gaussian process method: A) Construction of the prior distribution over orientation preference maps: we assume a Gaussian process prior both over the real and imaginary parts of the map. A sample from the prior can be obtained by filtering two instances of white noise with a Difference-of-Gaussian filter, and using the resulting matrices as the real and imaginary part of the map. Different samples from the prior can be generated by using different random samples of white noise. B) Three samples from a one-dimensional Gaussian process. We can think of them as slices through one of the two map components the real-valued part of the map. In the absence of experimental data, we assume the map components to be smooth, and to be distributed around zero. C) The observed data (black crosses) constrains the possible maps. Only samples (colored) which are consistent with the data have high probability. The samples are distributed around the posterior mean (black). In locations where no data was observed, there is considerable uncertainty about the location of the posterior mean (shaded in gray). This is true at the edges of the map, and at pixel index 0. In addition, the posterior uncertainty depends on the variance of the data observed. If the variance is low (left part), the map is tightly constrained by the data. If the variance is high (right part), there will be considerable residual uncertainty about the exact location of the posterior mean.

Bayesian inference: calculating the posterior distribution

As we have assumed both a Gaussian prior and a Gaussian likelihood, the posterior is Gaussian again, and therefore fully characterized by its mean μ_{post} and covariance Σ_{post}. By Bayes rule, the posterior distribution is proportional to the product of prior and likelihood,

$$P(\mathbf{m}|\{\mathbf{r}\}) \propto P_{\text{prior}}(\mathbf{m}) \prod_{i=1}^{N} P(\mathbf{r}_i|\mathbf{m}) \propto \exp\left(-\frac{1}{2}\mathbf{m}^\top K_{\mathbf{m}}^{-1}\mathbf{m}\right) \prod_{i=1}^{N} \exp\left(-\frac{1}{2}\left(\mathbf{r}_i - V_i^\top \mathbf{m}\right)^\top \Sigma_\epsilon^{-1}\left(\mathbf{r}_i - V_i^\top \mathbf{m}\right)\right).$$

Rearranging terms, this can be written as

$$P(\mathbf{m}|\{\mathbf{r}\}) \propto \exp\left(-\frac{1}{2}\left(\left(\mathbf{m}-\mu_{\text{post}}\right)^\top \Sigma_{\text{post}}^{-1}\left(\mathbf{m}-\mu_{\text{post}}\right)\right)\right),$$

where the posterior covariance Σ_{post} is given by

$$\Sigma_{\text{post}}^{-1} = K_{\mathbf{m}}^{-1} + \sum_i V_i^\top \Sigma_\varepsilon^{-1} V_i = K_{\mathbf{m}}^{-1} + \left(\sum_i v_i v_i^\top\right) \otimes \Sigma_\varepsilon^{-1}, \tag{3}$$

and the posterior mean μ_{post} by

$$\mu_{\text{post}} = \Sigma_{\text{post}}\left(\sum_i V_i \Sigma_\varepsilon^{-1} \mathbf{r}_i\right) = \Sigma_{\text{post}}\left(\mathbb{I}_d \otimes \Sigma_\varepsilon^{-1}\right) \sum_i v_i \otimes \mathbf{r}_i. \tag{4}$$

The posterior covariance is the harmonic mean of the prior covariance $K_{\mathbf{m}}$ and a second term which grows with each stimulus. Hence, as more data is observed, the relative importance of the prior covariance $K_{\mathbf{m}}$ to the data is reduced. The posterior mean can be obtained by taking a vector average of all stimuli ($\sum_i v_i \otimes \mathbf{r}_i$), dividing by the noise covariance ($\mathbb{I}_d \otimes \Sigma_\varepsilon^{-1}$), and finally multiplying by the posterior covariance Σ_{post}. Division by the noise covariance effectively re-weights the value of every pixel by its reliability. Multiplication by the posterior covariance couples the components of nearby pixels, and therefore acts as a spatial smoothing term. While the functional form of the posterior mean of the map (vector average, then smooth) is related to that in the classical approach of first vector averaging and then applying a fixed linear smoothing filter (Blasdel, 1992), it differs in three important ways: first, the filter shape is explicitly linked to our prior assumptions about the statistical structure of map (Sollich and Williams, 2005). Second, both anisotropic noise-variances and correlations are taken into account. Third, the smoothing is automatically adapted to the data.

We note that the posterior covariance will have block structure provided that the prior covariance K_m has block structure, and different feature maps are statistically independent *a priori*. That is, the stimuli are uncorrelated on average, i.e. $\sum_i v_i v_i^\top = D_v$ is diagonal. Hence, inference for different maps decouples, and we do not have to store the full joint covariance over all d maps.

Learning the noise model

In the formulas above, we assumed that the noise covariance Σ_ϵ is given. In practice, we have to infer both the noise-variance at each pixel, as well as correlations in the noise across pixels. We initialize the noise covariance by calculating the covariance of the responses in each stimulus condition, and averaging this estimate across stimulus conditions. Using this initial estimate of the noise covariance, we then derive the posterior mean and covariance. Using this estimate of the mean, we then update the estimate of Σ_ϵ by fitting it to the residuals $\mathbf{z}_i = \mathbf{r}_i - V_i^\top \mu_{\text{post}}$. We iterate between calculating the posterior mean (using the current estimate of Σ_ϵ) and then calculating a new point-estimate of the most likely noise covariance Σ_ϵ given the posterior mean (Kersting et al., 2007). In all cases, a very small number of iterations lead to convergence.

Typically, the number of stimulus presentations is much smaller than the number of pixels, and it is not possible to fully infer the noise covariance. Therefore, we assume the noise covariance either to be diagonal (i.e. noise to be uncorrelated), or of the form $\Sigma_\varepsilon = D_\epsilon + G_\epsilon G_\epsilon^\top$. Here, G_ϵ is of size $n \times q_\epsilon$, $q_\epsilon \ll n$, and D_ϵ is a diagonal matrix. The low-rank term $G_\epsilon G^\top$ models noise-correlations across pixels, whereas the diagonal matrix D_ϵ corresponds to noise which is independent in different pixels. The matrix G_ϵ can be used to visualize the spatial structure of correlations across the image.

Our noise model is equivalent to *factor-analysis* models (Everitt, 1984; Robertson and Symons, 2007; Yu et al., 2009), and can therefore be fit to the residuals using an Expectation Maximization algorithm (EM). We iteratively update both the noise-variances D_ϵ and the matrix G_ϵ modeling the noise-correlations:

$$D_\epsilon^{\text{new}} = diag\frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{z}_i\mathbf{z}_i^\top - 2\mathbf{z}_i\bar{\mu}^\top G_\epsilon + G_\epsilon\left(\bar{\Sigma} + \bar{\mu}_i\bar{\mu}_i^\top\right)G_\epsilon^\top\right) \tag{5}$$

$$G_\epsilon^{\text{new}} = \frac{1}{N}\left(\sum_{i=1}^{N}\mathbf{z}_i\bar{\mu}^\top\right)\left(\bar{\Sigma} + \frac{1}{N}\sum_i \bar{\mu}_i\bar{\mu}_i^\top\right)^{-1} \tag{6}$$

$$\text{where} \quad \bar{\mu}_i = G_\epsilon^\top\left(G_\epsilon G_\epsilon^\top + D_\epsilon\right)\mathbf{z}_i, \tag{7}$$

$$\bar{\Sigma} = \mathbb{I}_{q_\epsilon} - G_\epsilon^\top\left(G_\epsilon G_\epsilon^\top + D_\epsilon\right)^{-1}G_\epsilon. \tag{8}$$

This algorithm is guaranteed to converge whenever $q_\epsilon < N$ (Robertson and Symons, 2007), and can be applied without ever having to calculate the full noise covariance across all pixels. In addition, this model choice ensures that the noise covariance has full rank even when the number of data-points is less than the number of pixels.

Using low-rank approximations to make inference computationally tractable

The formulas for the posterior mean and covariance involve covariance matrices over all pixels and map components, or (if the prior decouples as described above), at least over all pixels. For a map of size $n = n_x \cdot n_y$, we would have to store and compute with matrices of size $n \times n$, and exact inference would become intractable (or at least impractical) for map sizes of more than 100 by 100. A number of approximation techniques have been proposed to make large-scale inference feasible in models with Gaussian process priors (see Rasmussen and Williams (2006) for an overview). Here, we utilize the fact that the spectrum of eigenvalues drops off quickly for many kernel functions (Bach and Jordan, 2002; Williams and Seeger, 2001), including the Difference-of-Gaussians used here. This means that the covariance matrix K can be approximated well by a low-rank matrix product $K \approx GG^\top$, where G is of size $n \times q$, $q \ll n$ (see Cressie and Johannesson (2008) for a related idea). To find G, we perform an incomplete Cholesky factorization on the matrix K. This can be done without having to store K in memory explicitly (Bach and Jordan, 2002). Alternatively, if the prior covariance is stationary, Fourier-methods could be used for finding a useful approximating G. In each case, the posterior covariance can be calculated without ever having to store (or invert) the full prior covariance:

$$\Sigma_{\text{post}} = \mathbb{I}_d \otimes \left(K - \beta^{-1}K\left(\Sigma_\varepsilon^{-1} - \Sigma_\varepsilon^{-1}G\left(\beta\mathbb{I}_q + G^\top\Sigma_\varepsilon^{-1}G\right)^1 G^\top\Sigma_\epsilon^{-1}\right)K\right),$$

where $\beta = 2/N$.

Because this prior does not have full rank, the posterior mean is constrained to lie in the column space of G, and might therefore be biased even in the limit of very large data-sets. To overcome this potential bias while still keeping the memory requirements in check,

one could use an approximation of the form $K = GG^\top + D$, where the diagonal matrix D is used to make K full rank, and that the diagonal of K is constant.

Decoding stimuli from imaging measurements

We defined a probabilistic model for relating stimuli to experimentally measured high-dimensional imaging measurements. Hence, we can use the model for decoding the presented stimulus from the measured responses. We can calculate the likelihood for each response under each stimulus, and use this for maximum-likelihood decoding. For each measured response, we can calculate the (unnormalized) log-likelihood

$$L(\mathbf{v}) = -\frac{1}{2}\left(\mathbf{r}^\top - \mu_{\text{post}}^\top \mathbf{v}\right)\Sigma_\varepsilon^{-1}\left(\mathbf{r} - \mathbf{v}^\top \mu_{\text{post}}\right) \tag{9}$$

for each stimulus under consideration, and pick the stimulus which maximizes it. If one wanted to decode stimulus orientations that the model was not trained on, one could optimize Eq. (9) over all possible stimuli $\mathbf{v}$, subject to some additional constraints or prior assumptions on $\mathbf{v}$.

When the model is used for a binary decoding task with 2 stimuli $\mathbf{v}$ and $\mathbf{w}$, the decision function is given by $d(\mathbf{r}) = \mathbf{m}^\top (v - w)\Sigma_\varepsilon^{-1}\mathbf{r}$, i.e. we pick $\mathbf{v}$ whenever $d(\mathbf{r})$ is greater than some decision criterion. In this case, the decoding model is closely related to linear discriminant analysis (Everson et al., 1997). However, the parameter vector $\mathbf{m}$ is obtained using the GP-prior over maps, and the noise covariance Σ_ε is estimated using the factor-analysis model. In addition, the parameters of the model can be estimated using all stimuli in the experiment, and not just $\mathbf{v}$ and $\mathbf{w}$. Thus, provided that the model assumptions are satisfied, the model can be fitted using a smaller number of trials per stimulus. Finally, rather than constructing a decoding algorithm for each pair of stimuli, as is commonly done for decoding stimuli from brain activity patterns, we have an integrated, interpretable model which can be used for decoding.

Experimental methods

To see how well the method works on experimental data, we used it to analyze data from an intrinsic signal optical imaging experiment. All experimental procedures were approved by the Duke University Institutional Animal Care and Use Committee and performed in compliance with guidelines published by the National Institutes of Health (USA). The surgical preparations of the animals and the optical imaging methods were described in detail previously (White et al., 1999). The central portion of the visuotopic map in visual areas V1 and V2 (White et al., 2001) of an anesthetized ferret (1% isoflurane in nitrous oxide/oxygen) was imaged with red light (wavelength 700 nm) while square wave gratings (spatial frequency 0.1 cycle/degree) were presented on a screen. Gratings were presented in 4 different orientations (0°, 45°, 90° and 135°), and moving along one of the two directions orthogonal to its orientation (temporal frequency 3.2 Hz).Each of the 8 possible directions was presented 100 times in a pseudo-random order for a duration of 5 s each, with an interstimulus interval of 8 s. Intrinsic signals were collected using a digital camera with pixel-size 30 μm. The response $\mathbf{r}_i$ was taken to be the average activity in a 5 second window relative to baseline. Each response vector $\mathbf{r}_i$ was normalized to have mean 0 and standard deviation 1, no spatial filtering was performed. We did not subtract the mean of individual pixels across trials. For all analyses in this paper, we concentrated on two regions of size 100 by 100 pixels each (see Fig. 3, labeled as 'region 1' and 'region 2'). Region 2 had a higher signal-to-noise ratio than region 1.

Results

Illustration on synthetic data: Gaussian process smoothing

To illustrate the ability of our method to recover orientation preference maps from noisy recordings, we simulated an imaging experiment. We generated a hypothetical map by generating one sample from the prior, i.e. by convolving a 100 by 100 matrix of Gaussian white noise with a Difference-of-Gaussians with parameters $\alpha_1 = 2$ and $\sigma_1 = 6$. Fig. 2A shows the angular components of the resulting orientation preference map. Then, we simulated imaging responses to each of 8 different oriented gratings by sampling responses according to the Eq. (1). The parameters were chosen to be roughly comparable with the experimental data (see below). Vector averaging the first 80 such responses (10 for each direction) yielded a noisy approximation to the 'true' map depicted in Fig. 2A.

We reconstructed the map using our GP-method (low-rank approximation of rank $q = 1600$, noise-correlations of rank $q_\varepsilon = 4$) on data-sets of different sizes (N between 16 and 960). In each case, we did not use the values of the hyper-parameters α and σ used to generate the prior (as they would not be known in a real experiment), but rather fitted them to the simulated experimental responses. Fig. 2C shows the angular components of the posterior mean of the GP, our reconstruction of the map for $N = 48$ (6 presentations for each direction). As can be seen, the GP-reconstruction looks much more similar to the true map than the 'naive' reconstruction shown in B. We used the posterior variances to also calculate a point-wise 95% confidence interval on the preferred orientation at each location, shown in Fig. 2D. As expected, the confidence intervals are biggest near pinwheels, where the orientation selectivity at a pixel-level (pixels were of size 30 μm) is low, and therefore the preferred orientation is not well defined.

To evaluate and compare different reconstruction methods, we need a metric for quantifying the similarity of different maps. We here chose to quantify reconstruction performance by computing the correlation coefficient of the reconstruction and the true map, each represented as a long vector with $2n$ elements. We wanted to investigate to what extent the GP-reconstruction is superior to a more conventional reconstruction based on vector averaging and subsequent smoothing with a fixed linear filter. This approach has free parameters, namely the shape of the filter (e.g. Gaussian) and its parameters (e.g. the filter width). In order to give the smoothing reconstruction the best chance, we optimized the filter width by maximizing the similarity (measured by the correlation coefficient) with the true map. This yields an optimistic estimate of the performance of this approach, as setting the filter-size requires access to the ground truth, which would not be possible in a real experiment. This approach yielded the reconstruction shown in Fig. 2E. This map has a correlation of $c = 0.80$ with the true map, in contrast with the GP-map, which has a correlation of $c = 0.90$. For the simple smoothing method, more than 160 stimulus presentations would be required to achieve this performance level. Using the smoothing approach with a Difference-of-Gaussian filter instead of a Gaussian did not lead to better performance, but rather had a correlation of $c = 0.72$. When we ignore noise-correlations (i.e. assume Σ_ε to be diagonal), GP still outperforms simple smoothing, although by a smaller amount (Fig. 2F). In general, we can see that the GP-map approaches the true map more quickly than the smoothed map (Fig. 2F). For example, using only 16 (2 for each direction) stimulus presentations, the smoothed map has a correlation with the ground truth of 0.59, whereas the correlation of the GP-map is 0.85. The GP with correlations showed superior performance not only for this particular synthetic map, but also was significantly better in the average across several samples from the prior for all data-set sizes (see Supplementary Fig. 1A. When using very many stimulus presentations, both methods will eventually converge to the true map. However, the GP-

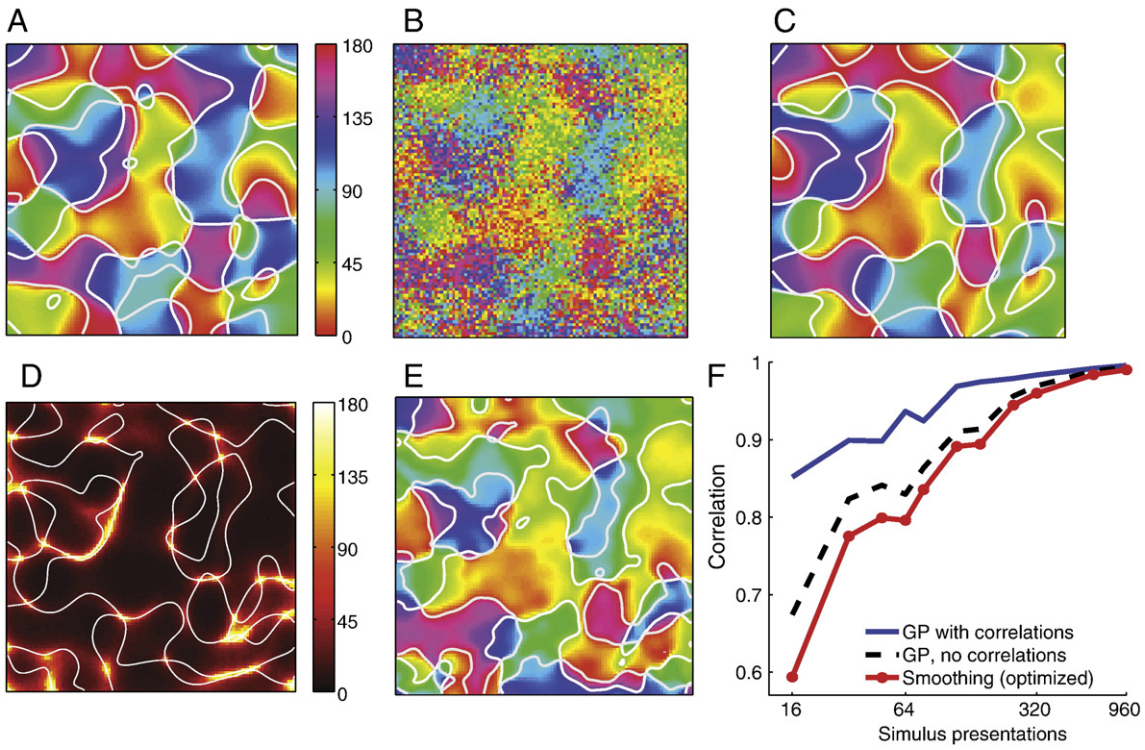

Fig. 2. Illustration of the GP estimation method on synthetic data: A) A synthetic OPM was generated by sampling from the Difference-of-Gaussian prior, and was used to simulate responses from an imaging experiment. B) A raw map was calculated by vector averaging the simulated responses to 6 stimulus presentations of 8 directions. C) The GP was used to reconstruct the map using the same data as in A). D) The posterior variance of GP is visualized as the size of 95% confidence intervals on preferred orientations. Superimposed are the zero-crossings of the GP-map. E) Reconstruction of the map by smoothing the raw map in B) with a fixed Gaussian filter, where the filter width was optimized by maximizing correlation with ground truth. F) Reconstruction performance as a function of stimulus presentations used: we evaluated the performance of the GP-method (with noise-correlations), the GP without noise-correlations, and the smoothing approach on data-sets of varying sizes. The GP outperforms the smoothing approach for all data-set sizes.

map with correlation converged at a much faster rate than the alternative methods, as can be seen from Supplementary Fig. 1B. For example, 800 stimulus presentations were necessary to reach an average reconstruction performance of $c = 0.98$ for the smoothing method, but only 190 using the full GP-method.

Application to optical imaging data from ferret V1

We used our GP-method to analyze and model data obtained using optical imaging of intrinsic signals of the visual cortex of anesthetized ferrets. Our large data-set with a total of 800 stimulus presentations made it possible to quantify the performance of our model by comparing it to unsmoothed maps. However, the GP-method itself is designed to also work robustly on smaller data-sets, and we are primarily interested in its performance in estimating maps using only few stimulus presentations. In the following, we will first describe the statistical structure of the data. Then, we will use the GP-method to estimate orientation preference and direction preference maps from it. Finally, we will use the model to decode stimulus identity from measured responses, and for reconstructing the map from sparse measurements.

Structure of data and choice of prior

Fig. 3A shows the map obtained by vector averaging all 800 measurements. The preferred orientation is color coded, and the selectivity is indicated by shading, where white corresponds to no selectivity. In order to evaluate whether our choice of a radial Difference-of-Gaussian covariance function is appropriate, we calculated the auto-correlation function of the empirical map, assuming stationarity across the whole map. For both the real and complex component of the map, the auto-correlation function could be approximated reasonably well by our Difference-of-Gaussian model (Carreira-Perpinan and Goodhill, 2004) (see Fig. 3B). However, it should be noted that our approach does not actually require this particular choice of covariance function: it just requires specification of *some* appropriate covariance. For example, one could simply take the empirically estimated auto-correlation function.

In the above, we stated that one of the advantages of our Gaussian process model is that it can deal with noise-variances that are not constant across the map. Here, we show that the noise in our hemodynamic imaging data is indeed highly non-uniform across the map. For each pixel x, we calculated the noise-variance as the pixel-variance for each stimulus averaged across all stimuli. Fig. 3C shows the layout of noise-variances across the map. As expected, large blood vessels show up clearly in this noise map, as the noise-variances are larger at blood vessels. The histogram of estimated noise-variances (Fig. 3D confirms that the measured noise-variances are not consistent with uniform noise: if the noise-variance was indeed constant across the map, 99% of the measured variances would fall between the two bars marked in gray. However, the spread of variances is much bigger than one would expect for isotropic noise.

Noise-correlations in functional imaging data can arise from a variety of sources, including optical blur (Polimeni et al., 2005), correlations in the hemodynamic responses as well as correlations in the underlying neural activity. To illustrate that the noise is correlated across pixels, we calculated the correlation between each two pixels per stimulus, and averaged this across all stimuli. Fig. 3E shows the average noise-correlation of each pixel with all of its neighbours within a 2-pixel radius. The blood vessels show up clearly in this map, indicating that the noise-correlations are substantially higher at blood vessels. Fig. 3C shows the histogram of noise-correlations. The average noise-correlation between neighbouring pixels was 0.37.

In summary, both the absolute magnitudes of the noise as well as the noise-correlations are much stronger at blood vessels. This shows

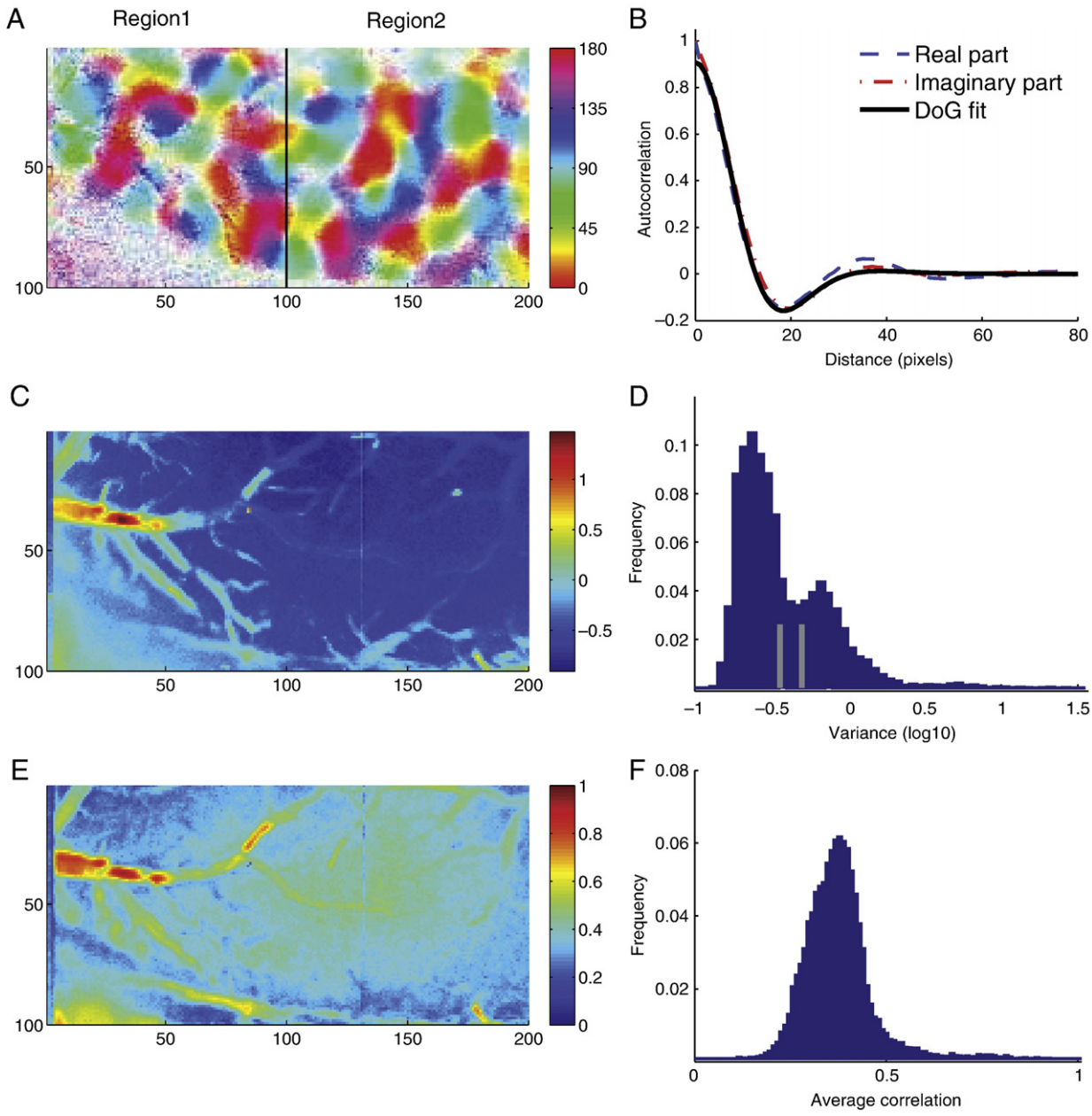

Fig. 3. Statistical structure of the optical imaging data. A) Raw orientation preference map, obtained by vector averaging data from all 800 trials. Colors code for preferred orientation (0° to 180°) and intensity for the absolute selectivity value $A(x)$, where white corresponds to no orientation selectivity, $A(x) = 0$. Selectivities are normalized such that 20% of the pixels have full saturation. B) Auto-correlation function of the real and imaginary parts of the empirical map in A. Superimposed is the best fitting Difference-of-Gaussian covariance function (see Methods). C) Noise-variances. The image shows the logarithm (base 10) of the residual variances across the map. D) Histogram of noise-variances (log basis 10). The two gray bars show spread (99% confidence region) of measured noise-variances that one would expect if the noise-variances were isotropic across the map. E) Noise-correlations. The map shows the average noise-correlation that each pixel has with its neighbours within a radius of 2 pixels. Similar to the noise-variances, the magnitude of correlations varies strongly across the map, and is highest at blood vessels. F) Histogram of noise-correlations, showing that the residuals are strongly correlated for neighbouring pixels.

that appropriate statistical models of hemodynamic imaging data must be able to model non-constant noise-variances as well as noise-correlations.

Estimating orientation preference maps

We want to use the GP-model to extract orientation preference maps from the optical imaging data described above. For real measured data, we do not know ground truth to estimate the performance of our model. Therefore, we used 5% of the data for estimating the map, and compared this map with the (unsmoothed) map estimated on the other 95% of data, which served as our proxy for ground truth. Fig. 4A shows the vector average of the first 760 trials of region 2, and Fig. 4B the average of the remaining 40 trials. Our goal was to reconstruct map A using only the data used to obtain B. Fig. 4C shows the reconstruction of the map using the GP-model, using a low-rank approximation of order $q = 1600$ and a one-dimensional noise model ($q_\epsilon = 1$). The correlation of the posterior mean with the 'ground truth' map was 0.89. As above, we compared the GP-map against one obtained by smoothing with a Gaussian kernel, where the kernel width of the smoothing kernel was chosen by maximizing its correlation with (our proxy for) the ground truth. Fig. 4D shows the smoothing reconstruction, which had a correlation of 0.84. Again, this is an optimistic estimate, as the width of the smoothing kernel was optimized using the 'ground truth' map.

The GP-map outperformed the smoothing map consistently: For 18 out of 20 different splits of the data from region 2 into training and

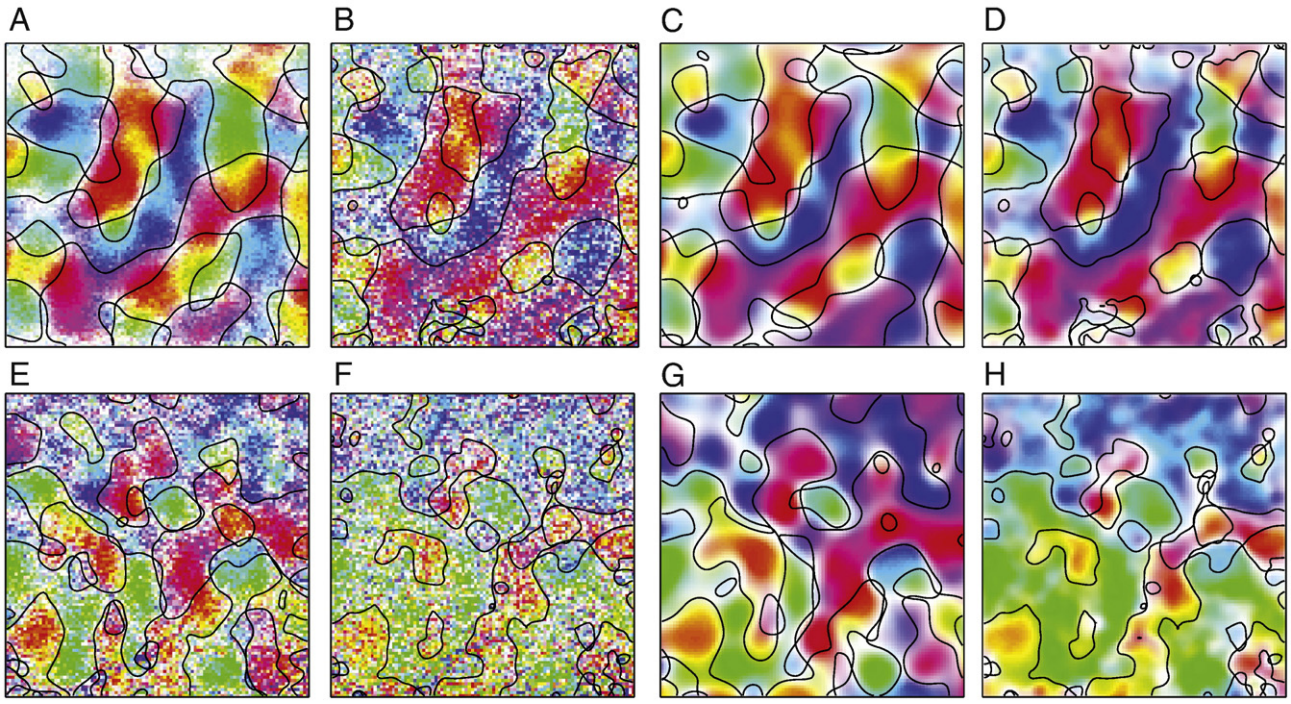

Fig. 4. Orientation preference and direction preference maps in ferret visual cortex. A) OPM estimated from 760 out of 800 stimuli by vector averaging. Zero-crossings plotted in black were smoothed with a 2 pixel window for display purposes. This map was taken to be our proxy for ground truth. Color coding as in Fig. 3. B) Vector average of the remaining 40 trials. C) Gaussian process reconstruction of the map, using the same data as in B). The correlation of this map with the 'ground truth' map in A is 0.89. D) For comparison, we show the map obtained by smoothing the raw map in C), where the filter width is obtained by maximizing the correlation to map A. The correlation with the map in A is 0.84. E–H) Same figures as above, but showing estimated direction preference maps. In this case, the color-code runs from 0 to 360°, and using 80 trials for estimation. The correlation of the GP-reconstruction with the 'ground truth' is 0.65, the correlation for the smoothed map is 0.60.

test data, the correlation of the GP-map was higher ($p<2\times10^{-4}$, average correlations $c=0.89\pm0.006$ S.E.M. for GP, $c=0.87\pm0.007$ for smoothing). The same held true when we smoothed maps with a Difference-of-Gaussians filter rather than a Gaussian (20 out of 20, average correlation $c=0.81\pm0.01$). Region 1 had a lower signal-to-noise ratio, and therefore estimation performance was worse on it. However, GP still outperformed the standard smoothing approach: the GP-map had a higher correlation on 20 out of 20 different splits (($p<2\times10^{-7}$, average correlations $c=0.67\pm0.016$ for GP, $c=0.61\pm$ 0.017 for smoothing). When using a Difference-of-Gaussian filter for the smoothing map, the average correlation was 0.53 ± 0.02, and the GP-map won on 20 out of 20 folds.

Estimating direction preference maps

Next, we wanted to use the GP-model to also extract direction preference maps from the imaging data. For this, we chose a representation of the tuning curve as $t(\theta)=v_1\cdot\cos(2\theta)+v_2\cdot\sin(2\theta)+v_3\cdot\cos(\theta)+v_4\cdot\sin(\theta)+v_5$.The complex valued direction preference map can then be constructed from the angle θ_D that maximizes the tuning curve by $m_D(x)=t(\theta_D)\exp(\mathbf{i}\theta_D)$). However, we here plot the map constructed from the directional sine and cosine-components v_3 and v_4, as the performance of the model on the components v_1 and v_2 has already been evaluated above, when looking at the OPM. It is known that DPMs in ferret V1 are somewhat weaker than OPMs. Therefore, we estimated the maps on 10% rather than 5% of the data, and used a 4 dimensional noise model.

Fig. 4E shows the map obtained using vector averaging on 90% of the data, our proxy for the ground truth, and Fig. 4F the vector average of the remaining 5%. In panel G, we can see the reconstruction of the map using the GP-method, and in panel H the map obtained using a Gaussian filter with optimized standard deviation. Again, the GP-approach has a better reconstruction performance, as its correlation with map in panel A is 0.65, whereas the correlation of the smoothing map is 0.60. On all 10 splits into test and training data, the GP-reconstruction performed better ($p<0.001$), and was also better on average (average correlation 0.57 ± 0.002 vs. 0.49 ± 0.03).

One of the strengths of the GP-model is that the filter-parameters are inferred by the model, and do not have to be set ad-hoc. The analyses above show that, even when we optimized the filter width for smoothing (which would not be possible in a real experiment), the GP still outperforms the approach of smoothing with a Gaussian window. In addition, it is important to keep in mind that using the posterior mean as a clean estimate of the map is only one feature of our model. In the following, we will use the GP-model to optimally interpolate a sparsely sampled map, and for decoding stimulus identity from the imaging measurements.

Interpolating the map

The posterior mean $\mu(x)$ of the model can be evaluated for any x. This makes it possible to extend the map to locations at which no data was recorded. This can be useful for two kinds of applications: first, if the measurement is corrupted in some pixels (e.g. because of a vessel artifact), we can attempt to recover the map in this region by model-based interpolation. In addition, multi-electrode arrays (Nauhaus et al., 2008) can be used to measure neural activity at multiple locations simultaneously. Provided that the electrode spacing is small enough, it should be possible to reconstruct at least a rough estimate of the map from such discrete measurements only (Swindale et al., 1987). We simulated a multi-electrode recording by only using the measured activity at 49 pixel locations which were chosen to be spaced 400 μm apart (similar to the spacing of available multi-electrode arrays (Nauhaus et al., 2008)). Then, we attempted to infer the full map using only these 49 measurements, and our prior knowledge about OPMs encoded in the prior covariance. The reconstruction is shown in Fig. 5C. Both interpolation methods yielded an approximate reconstruction of the map. The GP-map slightly outperformed the smoothing approach in terms of correlation coefficient($c=0.81$ vs. $c=0.79$). Again, the performance measure for

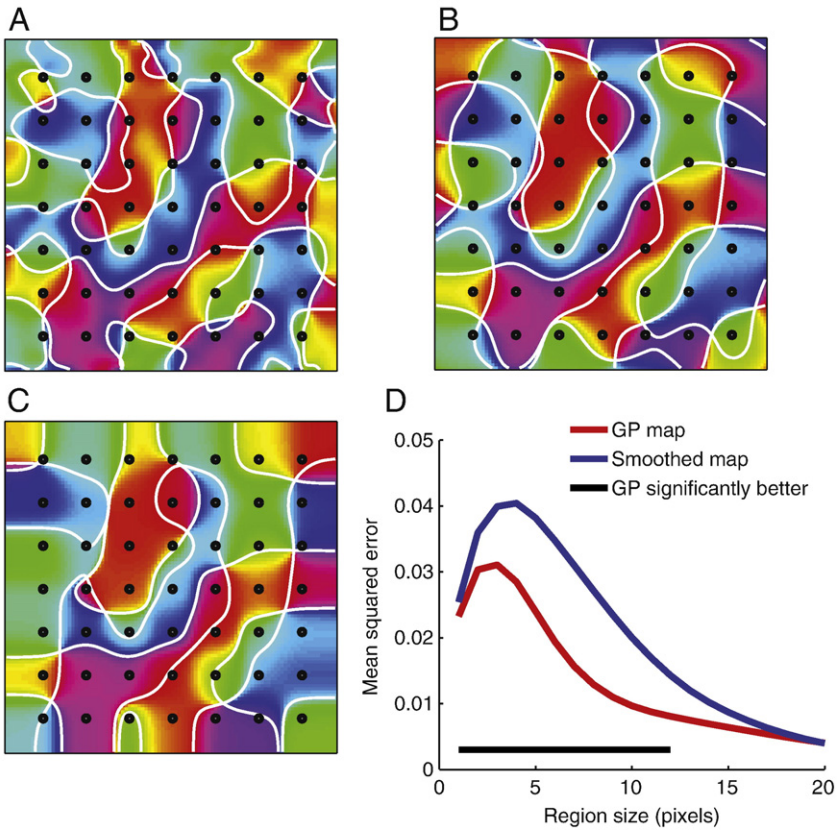

Fig. 5. Interpolations: the GP can be used to interpolate the map from sparse measurements, here from 49 simulated recording locations. A) Orientation preference map estimated by vector averaging data from all 800, and filtering with a Gaussian of width $\sigma = 2$ pixels. Marked in gray are 49 simulated recording locations, which were used for the reconstructions of the map plotted in B and C. B) Gaussian process reconstruction of the map from sparse measurements. Only data from the 49 pixels marked in gray was used for reconstruction. The correlation of the reconstructed map with the map plotted in A is 0.76. C) Map-reconstruction by fixed filtering. For comparison, we reconstructed the map by using a smoothing approach (see text, correlation is 0.74). D) Predicting homogeneity indices. We calculated the predicted homogeneity index at each of the 49 simulated recording locations for each of the three maps in A–C, and show the mean square error of the two reconstructions as a function of window size used. The GP-map has better prediction performance for all window sizes.

the smoothing approach is optimistic. In addition, the simple smoothing approach has a tendency to produce maps with zero-crossings that are aligned with the array, as it places a Gaussian bump at each electrode location. Secondly, as it does not take into account the semi-periodic nature of maps, it cannot extrapolate meaningful maps beyond the array.

The local homogeneity index (Nauhaus et al., 2008) can be used to quantify the homogeneity of orientation preference maps within a local region, and has been linked to the tuning-properties of cells recorded with electrode arrays (Nauhaus et al., 2008). To evaluate as to what extent our reconstruction can predict the local homogeneity of the map, we calculated the local homogeneity index for the real map and each of the two reconstruction methods at each of the 49 electrode locations. Fig. 5D shows the mean squared error of each of the two reconstructions as a function of the neighbourhood size of the homogeneity index, i.e. the size of the region around the electrode used to compute the index. As can be seen, the GP-reconstruction has a significantly better reconstruction performance for all neighbourhood sizes. For example, using a neighbourhood size of 5 pixels (150 μm), the GP-reconstruction provided a more precise prediction of the index for 36 out of 49 electrode positions ($p < 8 \cdot 10^{-4}$). The GP-prediction of the homogeneity index was significantly better than the smoothing-prediction for neighbourhood sizes up to 360 μm. On region 1, the correlations with the ground truth were 0.69 and 0.65, respectively. The GP-prediction was significantly better for all neighbourhood sizes (38 out of 49 for 150 μm).

On both regions, there was a slight, but consistent advantage of the GP-method over the smoothing approach. One further feature of the GP-approach is that the interpolated map is constructed from a probabilistic model of the data. Therefore, we do not only get the posterior mean as the (in this model) optimal reconstruction, but also a full distribution over maps. This can be used to assess how well the interpolated map is constrained by the data, and how much confidence one should place in the properties of the map that are 'guessed' by the interpolation, for example pinwheel counts or locations. Discriminative analysis methods for imaging data (e.g. Gabbay et al., 2000; Yokoo et al., 2001) cannot be used for such interpolations, as they are not based on a generative model of the map.

Decoding of orientation and direction

In this section, we will use our model for decoding both stimulus orientation and direction from the measurements. We split our data-set into 10 training sets and 10 non-overlapping test sets. On each training set, we fitted a GP-model with an uncorrelated noise model as well as a GP-model with noise-correlations of dimensionality 50. In addition, we calculated a model based on a smoothed map by vector averaging and subsequent smoothing with a Gaussian of width 2 pixels, and equipped it with an independent noise model.

For each measurement in the test set, we decoded orientation via maximum-likelihood methods, and evaluated decoding performance using the error rate, i.e. the proportion of stimuli that was not correctly classified. For decoding the correct orientation out of 4 possibilities, the chance level is therefore an error rate of 75%. Using the full GP-model on all 100×100 pixels, the error rate was only $5.4\% \pm 0.8$ and $4.0\% \pm 0.7$ on the two regions, respectively. When using the GP-model with

independent noise or the smoothed map model, error rates increased by factors of 2.6 and 3.0 on region 1, and by 2.6 and 4.2 on region 2. Thus, using a model of the data which also models noise-correlations made it possible to decode with much higher accuracy, achieving error rates that were at least a factor 2 better than independent models.

We wanted to know how decoding performance depends on the number of pixels used for decoding. To this end, we selected random subsets of pixels, and pruned our model by keeping only those elements of the mean vector and covariance that corresponded to these pixels. Error rates were averaged across multiple random subsets. Fig. 6 shows the error rates as a function of the number of pixels used for decoding. On both regions, using an average of 316 pixels, the performance of the full GP-model was superior to the smoothed map model using the full map with 10,000 pixels.

Each of the 4 orientations of the gratings was presented in two directions of motion. For each orientation, we want to decode direction of motion. Given that orientation preference maps in ferret V1 are more robust than direction preference maps, this constitutes a harder decoding task. Nevertheless, using the full GP-model, the error rate was $9.1\% \pm 1.0$ for region 1 and $3.6\% \pm 0.67$ for region 2. In contrast, the uncorrelated GP-model and the smoothing map model only went down to $35.9\% \pm 1.7$ and $39.1\% \pm 1.7$ ($27.4\% \pm 1.6$ and $35.0\% \pm 1.7$ on region 2). The chance level for this task is 50%. Thus, whereas the uncorrelated models barely achieved decoding performance above chance level, the GP-model with noise-correlations was able to achieve very low error rates. The large difference in performance was also evident when we looked at the dependence of decoding performance on the number of pixels used for decoding: using only 100 pixels, the correlated GP-model already beat the uncorrelated model with all 10,000 pixels.

Discussion

In this paper, we described Gaussian process methods for modeling functional imaging data and estimating functional maps. Specifically, we focussed on the estimation of orientation preference maps and direction preference maps from optical imaging data. Our method can be regarded as a generalization of the conventional approach of first averaging and smoothing: if the noise is uncorrelated and isotropic, and we assume cosine-tuning curves, the posterior mean of the GP will be exactly what is recovered using the smoothing approach with an appropriately chosen filter. However, we provide a principled, probabilistic interpretation of the smoothing-step, and allow for both noise-correlations and in-homogeneous noise-variances. Our approach yielded better estimates of maps on both synthetic and real imaging data. In addition, we showed how the approach can be used for model-based interpolations and for obtaining error bars on the map. Modeling of noise-correlations turned out to be of critical importance for decoding stimuli with high accuracy from the imaging measurements.

In contrast to previously used analysis methods for intrinsic signal imaging, ours is based on a generative model of the data. Having an explicit probabilistic model of the data allowed us to investigate the coding properties of the map. In particular, we used our statistical model for decoding stimulus identity from cortical responses. Our

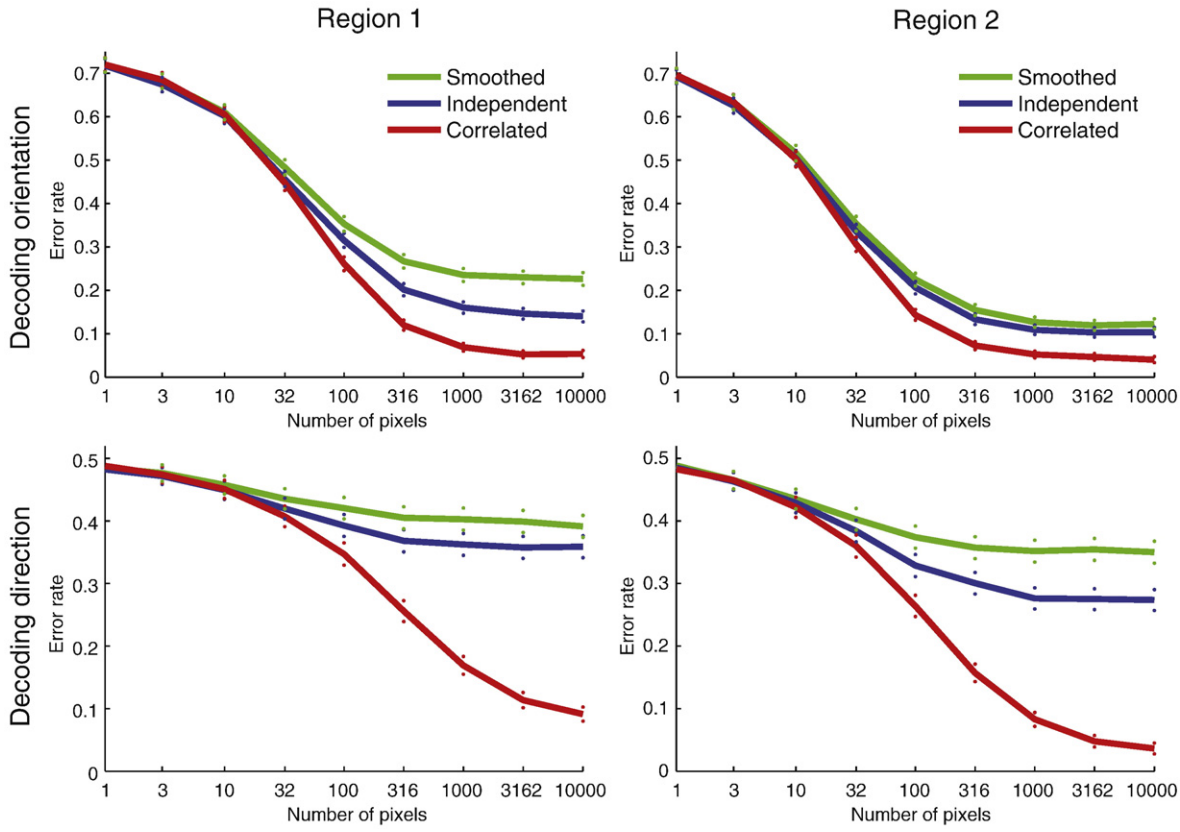

Fig. 6. Decoding performance of the Gaussian process models. The GP-model can be used for decoding stimulus identity (here: stimulus direction) from the experimental data using maximum-likelihood methods. The top row shows the decoding performance of the GP (with correlated noise), the GP (with independent noise) and the smoothing map (with independent noise) as a function of the number of pixels used for decoding. Pixels were randomly sampled from the map, and performance was averaged across samples. The top row shows decoding results for orientation (chance level is 75% error rate). The bottom row shows results for decoding of direction (up/down) when the orientation is known (chance level is 50%). The GP with correlations outperforms both the independent GP and the smoothed map. For decoding of direction, only the GP with correlations yields low error rates.

analysis showed that noise-correlations have to be taken into account to decode stimulus direction with high accuracy.

A number of studies have investigated the decoding of stimuli identity and other covariates from functional magnetic resonance imaging data (Haynes and Rees, 2006). The predominant approach is to use binary classification algorithms such as support vector machines or Fisher's linear discriminant. Our model differs from these approaches in that we assume a specific model for relating stimuli to imaging measurements which incorporates prior assumptions about the structure of cortical maps. This model is fit to the entire data-set including all stimulus presentations. Thus, our integrated model has the potential to provide better decoding performance if the number of presentations of each individual stimulus is low. However, it could also result in worse decoding performance if the model is applied to data for which the model assumptions are not justified.

With our model, we can estimate not only the posterior mean of the model as an optimally smoothed (or interpolated) map, but also the full posterior covariance over maps. Therefore, by sampling maps from the posterior distribution, we can generate different maps which are relatively likely under our model. With this approach, we can get point-wise error bars on the map-parameters. In addition, we can also put error bars on global statistical properties, such as the number or location of pinwheels (zero-crossings) of orientation preference maps (Macke et al., 2009). During an experiment, uncertainty estimates can be useful for allowing the experimenter to determine whether the quantity of interest with sufficient certainty, or whether more stimulus presentations are necessary.

In the context of analyzing functional imaging data, the theory of Gaussian random fields also has been used to perform statistical significance testing for smoothed-data with spatial correlations (Chumbley et al., 2009). These tests are usually based on the properties of thresholded random fields. In our framework, we would generate multiple samples from our posterior distribution to get insights into how well map properties are constrained by the experimental data.

Contrast-agent enhanced fMRI methods can be used to study orientation preference maps in some animal models (Fukuda et al., 2006). While it is possible to decode stimulus orientation from fMRI data of the human brain (Kamitani and Tong, 2005), the voxel-sizes of current human fMRI methods are not small enough to make our GP-approach for model-based smoothing applicable. However, our methods could still be used to perform a model-based interpolation of the measurements taken at each voxel. In addition, further developments in high resolution fMRI methodology are likely to provide the resolution necessary for estimating OPMs, and would make the methods presented here also applicable to human fMRI data. Furthermore, the GP-methods presented here provide a principled approach for using prior knowledge about the structure of cortical maps with imaging data, and are not necessarily restricted to the estimation of OPMs. Therefore, they can also be used for the estimation of other cortical maps, and in particular ones for which fMRI imaging has sufficient spatial resolution.

Our construction of a prior covariance for the GP can be generalized in a number of ways. First, rather than assuming the same covariance function for each map component, we could allow the prior covariances for different components to have different parameters. Alternatively, one could not restrict the radial component to be of Difference-of-Gaussian form, but estimate it from data in a non-parametric fashion. Similarly, the assumption of uncorrelated map components can be dropped, for example when modeling maps which are known to be related in particular ways. For example, if $m_k(x)$ and $m_l(x)$ are known to be negatively correlated (Hübener et al., 1997), one could define $K_{\text{prior}}(x, x', k, k) = K_{\text{prior}}(x, x', l, l) = (\beta_1^2 + \beta_2^2)K_c(x, x')$, and $K_{\text{prior}}(x, x', a, b) = -\beta_1^2 K_c(x, x')$. In this case, the prior correlation of the two map components at the same pixel would be given by $-\beta_1^2/(\beta_1^2 + \beta_2)^2$. As derivatives and gradients are linear functions of the map, the prior covariance can also be used to encode assumptions about the gradients of the map. This can be useful, e.g., for modeling orthogonality relationships between maps (Hübener et al., 1997; Kara and Boyd, 2009), or boundary effects (Shmuel and Grinvald, 2000).

We assumed a GP-prior over maps, i.e. the higher-order correlations of the maps to be minimal. It is known that the statistical structure of OPMs shows systematic deviations from Gaussian random fields (Erwin et al., 1995; Kaschube et al., 2008; Wolf and Geisel, 1998). This implies that there is room for improvement in the definition of the prior. For example, priors which are sparse have been shown to lead to superior reconstruction ability in a number of applications (Nickisch and Seeger, 2009). In addition, use of a non-Gaussian prior could facilitate reconstructions which go beyond the auto-correlation length of the GP-prior (Wolf et al., 1994). Finally, our assumption of a Gaussian noise model could be relaxed by using generalized linear models rather than Gaussian noise (Nickisch and Seeger, 2009; Rahnama Rad and Paninski, 2010). This would be especially useful for estimating maps for multi-electrode recordings of spike-activity, for which a Gaussian noise model is not appropriate. One future direction would be to investigate how general noise-correlation structures can be integrated in these models in a flexible manner. In addition, it remains to be seen whether the additional complexity of using a more involved noise model would lead to a substantial increase in performance. Our focus here was on imaging data obtained using trial based paradigms in which stimuli are presented sequentially. Paradigms using periodically changing stimuli (Kalatsky and Stryker, 2003; Sornborger et al., 2003; Sornborger et al., 2005) are based on frequency-based approaches to separate signal and noise components by looking for components which are time-locked to the periodicity of the stimulus. The development of Bayesian data-analysis methods for frequency-based approaches is an interesting direction for future research.

Acknowledgments

We would like to thank Audrey Sederberg, Michael Schnabel and Stefanie Liebe for useful discussions. This work is supported by the German Ministry of Education, Science, Research and Technology through the Bernstein award to MB (BMBF; FKZ: 01GQ0601), the Werner-Reichardt Centre for Integrative Neuroscience Tübingen, and the Max Planck Society. M.K. acknowledges support through NIH/NIGMS P50 GM071508.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2010.04.272.

References

Bach, F.R., Jordan, M.I., 2002. Kernel independent component analysis. J. Machine Learn. Res. 3 (1), 48.
Blasdel, G.G., 1992. Orientation selectivity, preference, and continuity in monkey striate cortex. J. Neurosci. 12 (8), 3139–3161.
Blasdel, G.G., Salama, G., 1986. Voltage-sensitive dyes reveal a modular organization in monkey striate cortex. Nature 321 (6070), 579–585.
Bonhoeffer, T., Grinvald, A., 1991. Iso-orientation domains in cat visual cortex are arranged in pinwheel-like patterns. Nature 353 (6343), 429–431.
Carreira-Perpinan, M.A., Goodhill, G.J., 2004. Influence of lateral connections on the structure of cortical maps. J. Neurophysiol. 92 (5), 2947–2959.
Chklovskii, D.B., Koulakov, A.A., 2004. Maps in the brain: what can we learn from them? Annu. Rev. Neurosci. 27, 369–392.
Chumbley, J., Worsley, K., Flandin, G., Friston, K., 2009. Topological FDR for neuroimaging. Neuroimage 49 (4), 3057–3064.
Coppola, D.M., White, L.E., Fitzpatrick, D., Purves, D., 1998. Unequal representation of cardinal and oblique contours in ferret visual cortex. Proc. Natl Acad. Sci. USA 95 (5), 2621–2623.
Cornford, D., Csato, L., Evans, D., Opper, M., 2004. Bayesian analysis of the scatterometer wind retrieval inverse problem: some new approaches. J. R. Stat. Soc. B Stat. Methodol. 609–652.

Cressie, N., 1992. Statistics for spatial data. Terra Nova 4 (5), 613–617.
Cressie, N., Johannesson, G., 2008. Fixed rank kriging for very large spatial data sets. J. R. Stat. Soc. Stat. Methodol. 70 (1), 209–226.
Csató, L., Cornford, D., Opper, M., 2001. Online approximations for wind-field models. Lect. Notes Comput. Sci. 300–307.
Erwin, E., Obermayer, K., Schulten, K., 1995. Models of orientation and ocular dominance columns in the visual cortex: a critical comparison. Neural Comput. 7 (3), 425–468.
Everitt, B., 1984. An introduction to latent variable models. Chapman and Hall, New York.
Everson, R., Knight, B.W., Sirovich, L., 1997. Separating spatially distributed response to stimulation from background. i. optical imaging. Biol. Cybern. 77 (6), 407–417.
Friston, K.J., 2002. Bayesian estimation of dynamical systems: an application to fMRI. Neuroimage 16 (2), 513–530.
Fukuda, M., Moon, C.-H., Wang, P., Kim, S.-G., 2006. Mapping iso-orientation columns by contrast agent-enhanced functional magnetic resonance imaging: reproducibility, specificity, and evaluation by optical imaging of intrinsic signal. J. Neurosci. 26 (46), 11821–11832.
Gabbay, M., Brennan, C., Kaplan, E., Sirovich, L., 2000. A principal components-based method for the detection of neuronal activity maps: application to optical imaging. Neuroimage 11 (4), 313–325.
Goodhill, G.J., 2007. Contributions of theoretical modeling to the understanding of neural map development. Neuron 56 (2), 301–311.
Haynes, J.-D., Rees, G., 2006. Decoding mental states from brain activity in humans. Nat. Rev. Neurosci. 7 (7), 523–534.
Hübener, M., Shoham, D., Grinvald, A., Bonhoeffer, T., 1997. Spatial relationships among three columnar systems in cat area 17. J. Neurosci. 17 (23), 9270–9284.
Hunt, J.J., Giacomantonio, C.E., Tang, H., Mortimer, D., Jaffer, S., Vorobyov, V., Ericksson, G., Sengpiel, F., Goodhill, G.J., 2009. Natural scene statistics and the structure of orientation maps in the visual cortex. Neuroimage 47 (1), 157–172.
Issa, N.P., Trepel, C., Stryker, M.P., 2000. Spatial frequency maps in cat visual cortex. J. Neurosci. 20 (22), 8504–8514.
Issa, N.P., Rosenberg, A., Husson, T.R., 2008. Models and measurements of functional maps in v1. J. Neurophysiol. 99 (6), 2745–2754.
Kalatsky, V.A., Stryker, M.P., 2003. New paradigm for optical imaging: temporally encoded maps of intrinsic signal. Neuron 38 (4), 529–545.
Kamitani, Y., Tong, F., 2005. Decoding the visual and subjective contents of the human brain. Nat. Neurosci. 8 (5), 679–685.
Kara, P., Boyd, J.D., 2009. A micro-architecture for binocular disparity and ocular dominance in visual cortex. Nature 458 (7238), 627–631.
Kaschube, M., Schnabel, M., Wolf, F., 2008. Self-organization and the selection of pinwheel density in visual cortical development. New J. Phys. 10 (1), 015009.
Kersting, K., Plagemann, C., Pfaff, P., Burgard, W., 2007. Most likely heteroscedastic Gaussian process regression. Proceedings of the 24th international conference on Machine learning. ACM New York, NY, USA, pp. 393–400.
Lawrence, N., Platt, J., 2004. Learning to learn with the informative vector machine. Proceedings of the twenty-first international conference on Machine learning. ACM, p. 65.
LeVay, S., Stryker, M.P., Shatz, C.J., 1978. Ocular dominance columns and their development in layer iv of the cat's visual cortex: a quantitative study. J. Comp. Neurol. 179 (1), 223–244.
Li, Y., Hooser, S.D.V., Mazurek, M., White, L.E., Fitzpatrick, D., 2008. Experience with moving visual stimuli drives the early development of cortical direction selectivity. Nature 456 (7224), 952–956.
Macke, J.H., Gerwinn, S., White, L.E., Kaschube, M., Bethge, M., 2009. Bayesian estimation of orientation preference maps. In: Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C., Culotta, A. (Eds.), Advances in Neural Information Processing Systems, volume 22. MIT Press, pp. 1195–1203.
Nauhaus, I., Benucci, A., Carandini, M., Ringach, D.L., 2008. Neuronal selectivity and local map structure in visual cortex. Neuron 57 (5), 673–679.
Nickisch, H., Seeger, M., 2009. Convex variational Bayesian inference for large scale generalized linear models. In: Danyluk, A., Bottou, L., Littman, M. (Eds.), International Conference on Machine Learning, volume 26. ACM Press, New York, NY, USA, pp. 761–768.
Ohki, K., Chung, S., Ch'ng, Y.H., Kara, P., Reid, R.C., 2005. Functional imaging with cellular resolution reveals precise micro-architecture in visual cortex. Nature 433 (7026), 597–603.
Penny, W.D., Trujillo-Barreto, N.J., Friston, K.J., 2005. Bayesian fMRI time series analysis with spatial priors. Neuroimage 24 (2), 350–362.
Polimeni, J.R., Granquist-Fraser, D., Wood, R.J., Schwartz, E.L., 2005. Physical limits to spatial resolution of optical recording: clarifying the spatial structure of cortical hypercolumns. Proc. Natl. Acad. Sci. U. S. A. 102 (11), 4158–4163.
Rahnama Rad, K., Paninski, L., 2010. Efficient estimation of two-dimensional firing rate surfaces via Gaussian process methods. Network: Computation in Neural Systems 21, 142–168.
Rasmussen, C., Williams, C., 2006. Gaussian processes for machine learning. MIT Press Cambridge, MA, USA.
Robertson, D., Symons, J., 2007. Maximum likelihood factor analysis with rank-deficient sample covariance matrices. J. Multivar. Anal. 98 (4), 813–828.
Rojer, A.S., Schwartz, E.L., 1990. Cat and monkey cortical columnar patterns modeled by bandpass-filtered 2d white noise. Biol. Cybern. 62 (5), 381–391.
Shmuel, A., Grinvald, A., 2000. Coexistence of linear zones and pinwheels within orientation maps in cat visual cortex. Proc. Natl. Acad. Sci. U. S. A. 97 (10), 5568–5573.
Sirovich, L., Everson, R., 1992. Management and analysis of large scientific datasets. Int. J. Supercomp. Applic. 6 (1), 50–68.
Sirovich, L., Uglesich, R., 2004. The organization of orientation and spatial frequency in primary visual cortex. Proc. Natl. Acad. Sci. USA 101 (48), 16941–16946.
Sollich, P., Williams, C., 2005. Using the equivalent kernel to understand Gaussian process regression. Adv. Neural Inf. Process. Syst. 17.
Sornborger, A., Sailstad, C., Kaplan, E., Sirovich, L., 2003. Spatiotemporal analysis of optical imaging data. Neuroimage 18 (3), 610–621.
Sornborger, A., Yokoo, T., Delorme, A., Sailstad, C., Sirovich, L., 2005. Extraction of the average and differential dynamical response in stimulus-locked experimental data. J. Neurosci. Meth. 141 (2), 223–229.
Stetter, M., Schiessl, I., Otto, T., Sengpiel, F., Hübener, M., Bonhoeffer, T., Obermayer, K., 2000. Principal component analysis and blind separation of sources for optical imaging of intrinsic signals. Neuroimage 11 (5 Pt 1), 482–490.
Swindale, N.V., 1998. Orientation tuning curves: empirical description and estimation of parameters. Biol. Cybern. 78 (1), 45–56.
Swindale, N.V., Matsubara, J.A., Cynader, M.S., 1987. Surface organization of orientation and direction selectivity in cat area 18. J. Neurosci. 7 (5), 1414–1427.
Teh, Y.W., Seeger, M., Jordan, M.I., 2005. Semiparametric latent factor models. *Proceedings of the International Workshop on Artificial Intelligence and Statistics*, volume 10.
Weliky, M., Bosking, W.H., Fitzpatrick, D., 1996. A systematic map of direction preference in primary visual cortex. Nature 379 (6567), 725–728.
White, L.E., Bosking, W.H., Williams, S.M., Fitzpatrick, D., 1999. Maps of central visual space in ferret v1 and v2 lack matching inputs from the two eyes. J. Neurosci. 19 (16), 7089–7099.
White, L.E., Bosking, W.H., Fitzpatrick, D., 2001. Consistent mapping of orientation preference across irregular functional domains in ferret visual cortex. Vis. Neurosci. 18 (1), 65–76 (Jan–Feb).
Williams, C., Seeger, M., 2001. Using the Nystrom method to speed up kernel machines. In: Leen, T.K., Dietterich, T.G., Tresp, V. (Eds.), *Neural Information Processing Systems 13*, volume 13, pp. 682–688.
Wolf, F., Geisel, T., 1998. Spontaneous pinwheel annihilation during visual development. Nature 395 (6697), 73–78.
Wolf, F., Pawelzik, K., Geisel, T., Kim, D., Bonhoeffer, T., 1994. Optimal smoothness of orientation preference maps. Network: Computation in Neural Systems Computation in Neurons and Neural Systems, pp. 97–101.
Yacoub, E., Harel, N., Ugurbil, K., 2008. High-field fMRI unveils orientation columns in humans. Proc. Natl Acad. Sci. USA 105 (30), 10607–10612.
Yokoo, T., Knight, B., Sirovich, L., 2001. An optimization approach to signal extraction from noisy multivariate data. Neuroimage 14 (6), 1309–1326.
Yu, B.M., Cunningham, J.P., Santhanam, G., Ryu, S.I., Shenoy, K.V., Sahani, M., 2009. Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. J. Neurophysiol. 102 (1), 614–635.