Why MCA? Nonlinear sparse coding with spike-and-slab prior for neurally plausible image encoding

Jacquelyn A. Shelton[1], Philip Sterne[1], Jörg Bornschein[1], Abdul-Saboor Sheikh[1], and Jörg Lücke[1,2]

[1]Frankfurt Institute for Advanced Studies, Goethe-Universität Frankfurt; [2]Dept. of Physics, Goethe Universität Frankfurt

FIAS Frankfurt Institute for Advanced Studies

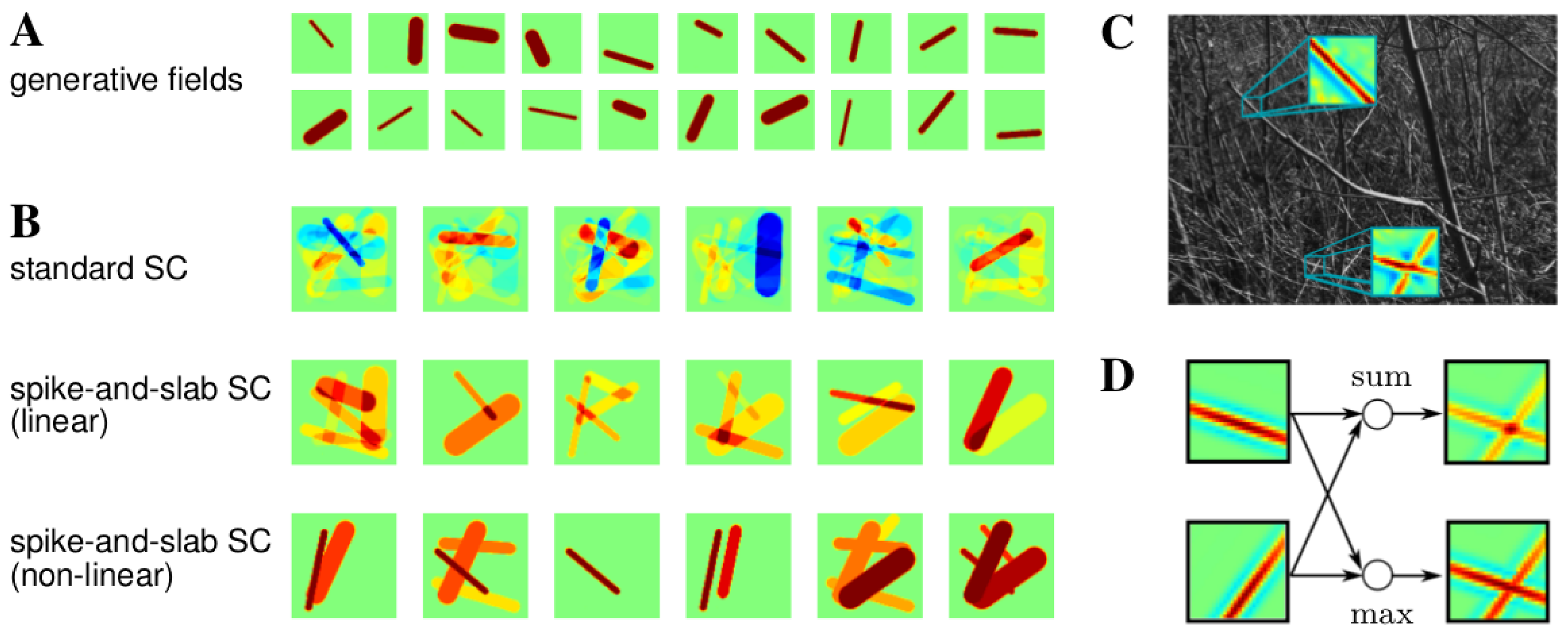
Bernstein Focus: Neurotechnology Frankfurt

DFG

Introduction

- Sparse coding (SC) as realistic model for low-level image statistics/V1 simple cells could be improved.
- Novel model generalizing SC in 2 ways:
  (1) spike-and-slab prior distribution for component absence/intensity,
  (2) nonlinear component combination; maximal causes analysis, MCA.
- Challenge: intractable parameter optimization $\rightarrow$ either (1) or (2) results in strongly multimodal posteriors
- Plan: Tackle intractabilities with an exact piecewise Gibbs sampling method combined with preselection of latent dimensions [1, 2]

Model: Nonlinear Spike-and-slab Sparse Coding



- Generative model for sensory data $\vec{y} = (y_1, \ldots, y_D)$ with hidden causes/objects $\vec{s} = (s_1, \ldots, s_H)$ and parameters Θ:

$$p(y_d \,|\, \vec{s}, \Theta) = \mathcal{N}(y_d;\, \max_h\{s_h W_{dh}\},\, \sigma^2)$$

MCA's $\max_h$ [3, 4] considers all H latents, takes h with max $s_h W_{dh}$

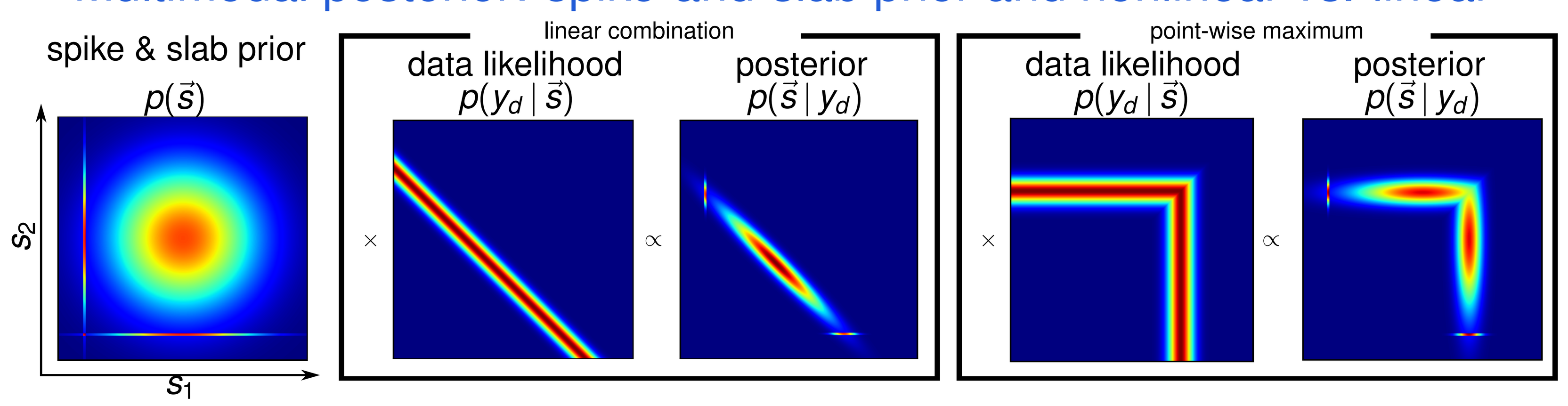
s_h distributed as spike-and-slab, $s_h = b_h z_h$:

$$p(b_h \,|\, \Theta) = \mathcal{B}(b_h; \pi) = \pi^{b_h}(1-\pi)^{1-b_h}$$
$$p(z_h \,|\, \Theta) = \mathcal{N}(z_h;\, \mu_{\mathrm{pr}}, \sigma^2_{\mathrm{pr}})$$

- Expectation values to maximize for h and average over n and d:

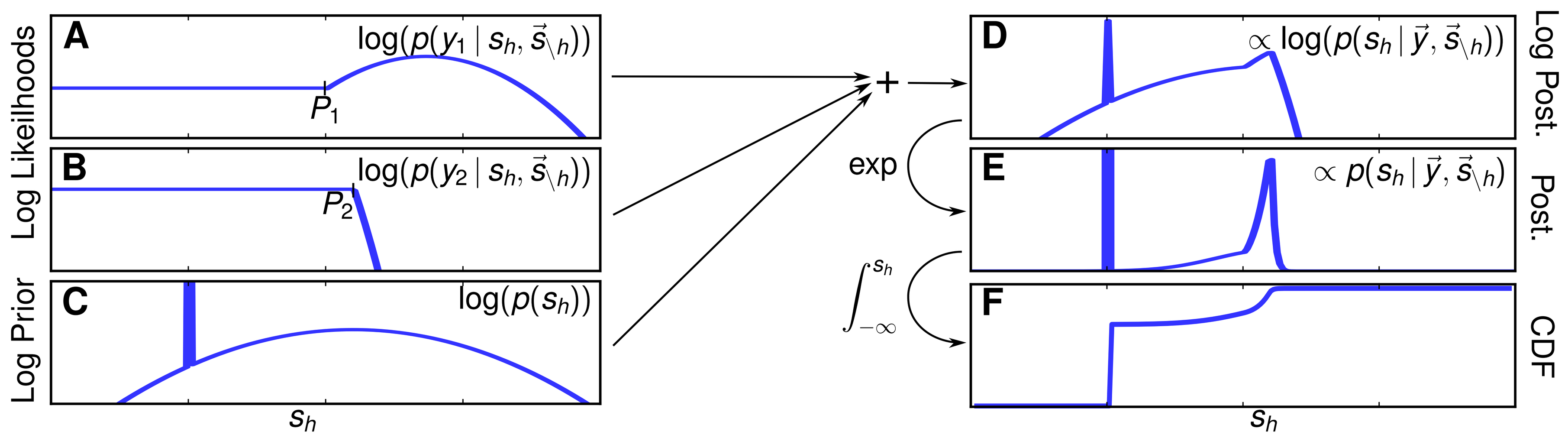
$$\langle f(s)\rangle^* = \sum_n \frac{\int_s p(\vec{s}|\vec{y}^{(n)}, \Theta)\, \delta(\text{h is max})\, f(s)}{\int_s p(\vec{s}|\vec{y}^{(n)}, \Theta)\, \delta(\text{h is max})}$$

Multimodal posterior: spike-and-slab prior and nonlinear vs. linear



Inference: Exact Gibbs Sampling with Latent Preselection

Gibbs Sampling for multimodal posteriors



- Construct a Markov chain with target density given by conditional posterior:

$$p(s_h|\vec{s}_{H\backslash h}, \vec{y}, \theta) \;\propto\; p(s_h|\theta) \prod_{d=1} p(y_d|s_h, \vec{s}_{H\backslash h}, \theta)$$

where distribution factorizes into $D+1$ factors: *1 : prior* and *D : likelihoods*

- MCA likelihood of a single data dimension y_d is a piecewise function (**A** & **B**):

$$p(y_d|s_h, \vec{s}_{H\backslash h}, \theta) = \mathcal{N}(y_d;\, \max_{h'}\{W_{dh'}s_{h'}\}, \sigma^2)$$
$$= \begin{cases} \underbrace{\mathcal{N}(y_d;\, \max_{h'\backslash h}\{W_{dh'}s_{h'}\}, \sigma^2)}_{\text{constant}} = \exp(l_d(s_h)) & \text{if } s_h < P_d \\ \mathcal{N}(y_d;\, W_{dh}s_h, \sigma^2) \quad\;\; = \exp(r_d(s_h)) & \text{if } s_h \geq P_d \end{cases}$$

- Transition points define where $s_h W_{dh}$ becomes the maximal cause of y_d:

$$P_d = \max_{h'\backslash h}\{W_{dh'}s_{h'}\} \;/\; W_{dh}$$

- Log of $p(\vec{y}|s_h, \vec{s}_{H\backslash h}, \theta)$ results in several piecewise functions – left-piece constant and right-piece quadratic – that are easily summed:

$$m(s_h) = \textstyle\sum_d^D \log p(y_d|s_h, \vec{s}_{H\backslash h}, \theta)$$

- Prior slab $\rightarrow$ add its 2nd degree polynomial to all pieces $m_i(s_h)$ (**C**)
- All function segments $m_i(s_h)$ are 2nd degree polynomials $\rightarrow$ expressed by computing 3 coefficients for each segment $m_i(s_h)$ of $p(y_d|s_h, \vec{s}_{H\backslash h}, \theta)$ (**D**)
- Construct piecewise cumulative distribution function (CDF): relate each segment $m_i(s_h)$ to the Gaussian $\propto \exp(m_i(s_h))$ it defines (**E**)
- Prior spike $\rightarrow$ introduce a step into the CDF corresponding to $s_h = 0$ (**F**)
- Sample $s_h \sim p(s_h|\vec{s}_{\backslash h}, \vec{y}, \theta)$ by inverse transform sampling from CDF

Preselection

- Variational approximation to posterior with support reduced to $\mathcal{K}_n$ [1, 2]:

$$p(\vec{s}\,|\,\vec{y}^{(n)}, \Theta) \approx q_n(\vec{s}; \Theta) = \frac{p(\vec{s}\,|\,\vec{y}^{(n)}, \Theta)}{\sum_{\vec{s}'\in\mathcal{K}_n} p(\vec{s}'\,|\,\vec{y}^{(n)}, \Theta)}\, \delta(\vec{s} \in \mathcal{K}_n)$$

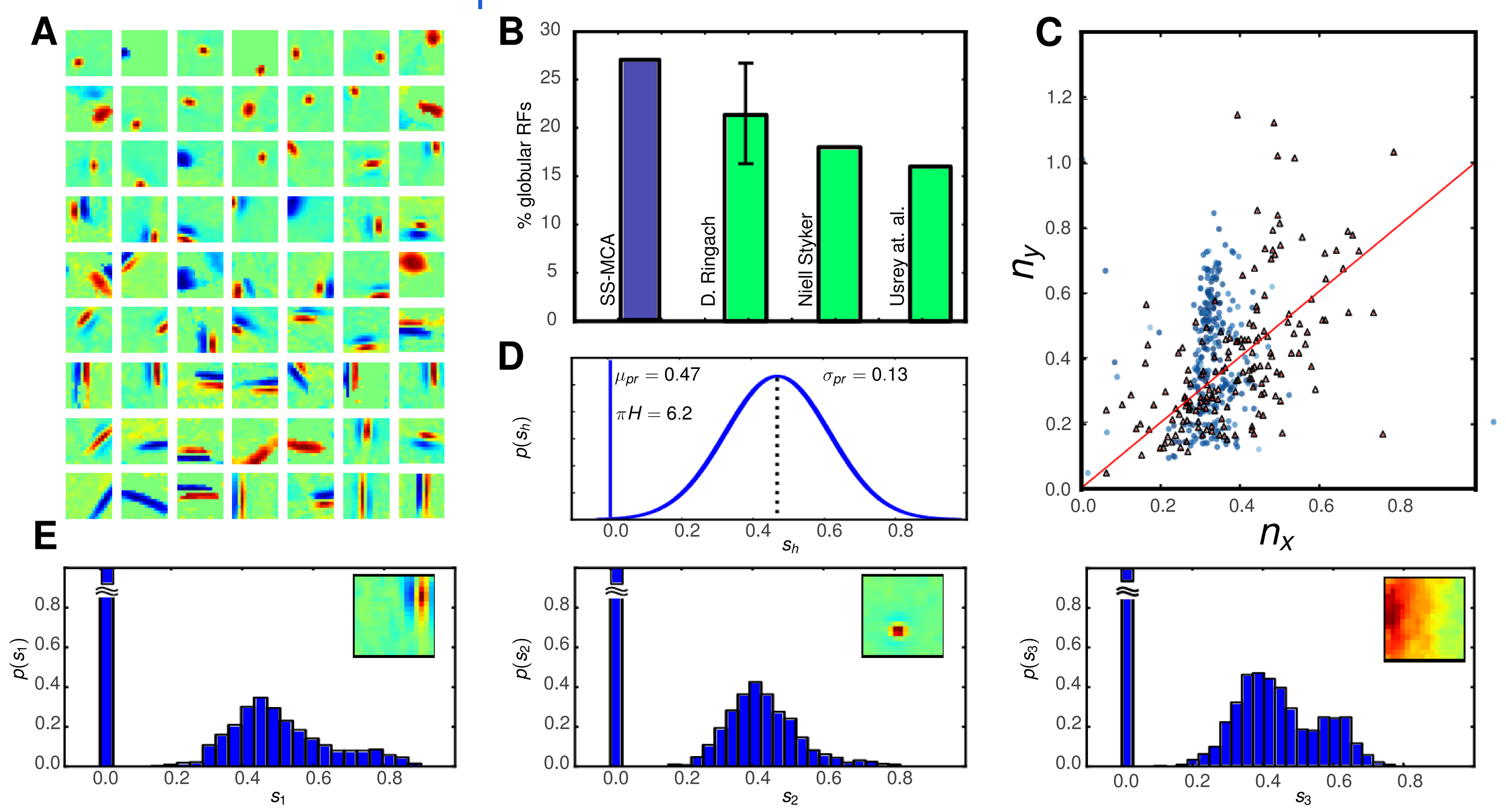
- Preselection of latent subset $\mathcal{K}_n = \{\vec{s}\,|\,\forall\, h \notin \mathcal{I}_n : s_h = 0\}$ with data-driven deterministic *selection function* to find most likely causes s_h of data for $\mathcal{I}_n$:

$$\mathcal{S}_h(\vec{y}^{(n)}) \;=\; |\vec{W}_h - \vec{y}^{(n)}|_2^2 \;/\; |\vec{W}_h|_2$$

Experiments

Natural image patches

Comparison with *in vivo* neural data



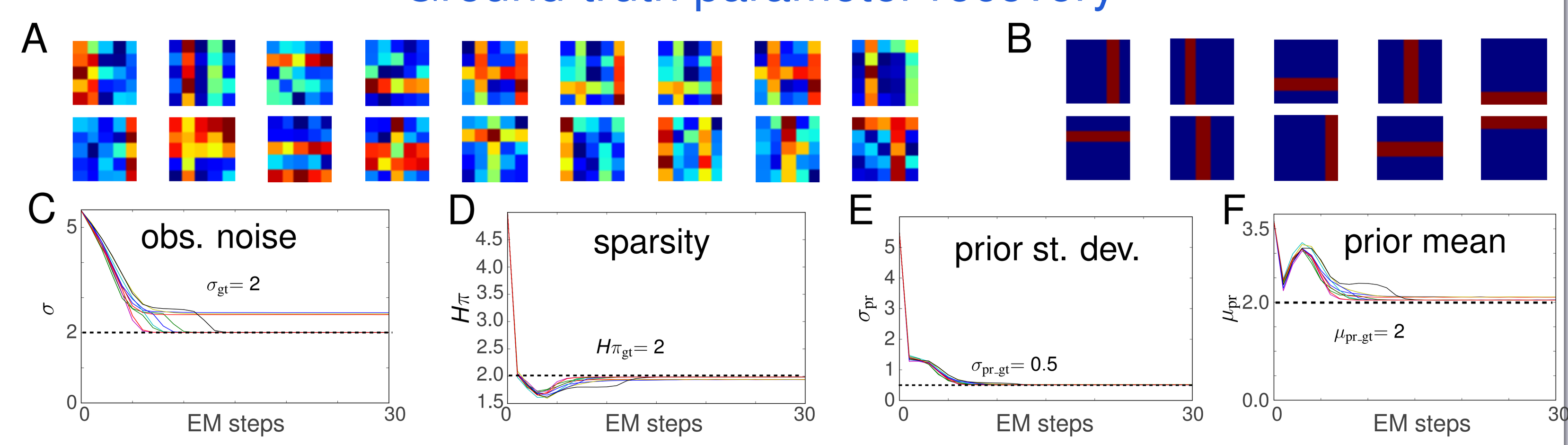
- Efficient large-scale application: $N = 50,000$ image patches with $D = 16 \times 16$ pixels, $H = 500$ latents, preselected to subset $H' = 20$
- Model consistency: satisfies necessary condition for true model [5]:

$$\lim_{N\to\infty} \tfrac{1}{N} \textstyle\sum_n p(\vec{s}\,|\,\vec{y}^{(n)}, \Theta) = p(\vec{s}\,|\,\Theta),$$

a test standard sparse coding fails (see [6] for a discussion).

Artificial data

Ground-truth parameter recovery



Discussion

- First time a model combining modifications (1) and (2) can be trained efficiently while retaining the rich structure of the posteriors.
- Derived algorithm enables efficient inference of all model parameters.
- Optimal prior shows asymmetric and bimodal activity of simple cells.
- Model is consistent; average posterior is approximately equal to prior.
- Model predicts a high percentage of globular receptive fields alongside Gabor-like fields; similar to proportions observed *in vivo*.

References & Acknowledgements

[1] J. Lücke and J. Eggert. (2010). Expectation Truncation And the Benefits of Preselection in Training Generative Models. Journal of Machine Learning Research (JMLR).
[2] J. Shelton, J. Bornschein, A.-S. Sheikh, P. Berkes, and J. Lücke. (2011). Select and sample - a model of efficient neural inference and learning. Advances in Neural Information Processing Systems (NIPS), 24.
[3] J. Lücke and M. Sahani. (2008). Maximal causes for non-linear component extraction. Journal of Machine Learning Research (JMLR).
[4] G. Puertas, J. Bornschein, and J. Lücke. (2010). The maximal causes of natural scenes are edge filters. Advances in Neural Information Processing Systems (NIPS), 23.
[5] P. Berkes, G. Orban, M. Lengyel, and J. Fiser. (2011). Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. Science, 331(6013):83–87.
[6] B. Olshausen and K. Millman. (2000). Learning sparse codes with a mixture-of-Gaussians prior. Advances in Neural Information Processing Systems (NIPS), 13.

We acknowledge support by the German Research Foundation (DFG) in the project LU 1196/4-2, by the German Federal Ministry of Education and Research (BMBF), project 01GQ0840, and by the LOEWE Neuronale Koordination Forschungsschwerpunkt Frankfurt (NeFF). Finally, we acknowledge support by the Frankfurt Center for Scientific Computing (CSC).