Chalmers Machine Learning Summer School
Approximate message passing and biomedicine

Part 2: Multivariate fMRI analysis using a sparsifying spatio-temporal prior

Tom Heskes

joint work with Marcel van Gerven and Botond Cseke

Machine Learning Group, Institute for Computing and Information Sciences
Radboud University Nijmegen, The Netherlands

April 15, 2015

Outline

Outline

Large p, small N

Many datasets grow wide, with many more features than samples.

Neuroimaging: $p = 20K$ pixels, $N = 100$ experiments/subjects.

Document classification: $p = 20K$ features (bag of words), $N = 5K$ documents.

Micro-array studies: $p = 40K$ genes measured for $N = 100$ subjects.

In all of these we use linear models, e.g., linear regression, logistic regression, that we have to regularize when $p \gg N$.

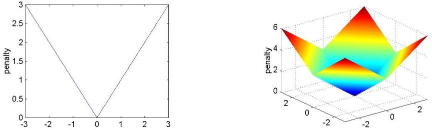
The Lasso

Given observations $\{y_i, x_{i1}, \ldots, x_{ip}\}_{i=1}^{N}$, minimize

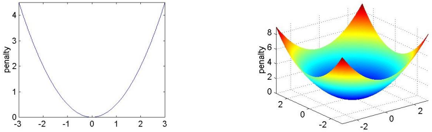
$$E(\boldsymbol{\theta}) = \sum_{i=1}^{N} \left(y_i - \theta_0 - \sum_{j=1}^{p} x_{ij}\theta_j \right)^2 + \gamma \sum_j |\theta_j| \, .$$

- Similar to ridge regression, with penalty $\sum_j \theta_j^2$.
- Lasso leads to sparse solutions (many θ_j's to zero), whereas ridge regression only shrinks.

Lasso



Ridge regression



Probabilistic Interpretation

- Likelihood:

$$p(Y|\boldsymbol{\theta}, X) = \prod_{i=1}^{N} \mathcal{N}(y_i; \boldsymbol{\theta}^T \mathbf{x}_i, \beta^{-1}) \, .$$

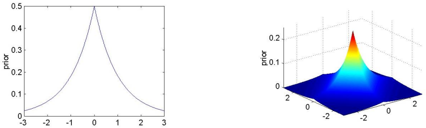
- Prior, for Lasso,

$$p(\boldsymbol{\theta}) = \prod_{j=1}^{p} \mathcal{L}(\theta_j; 0, \lambda) \, ,$$

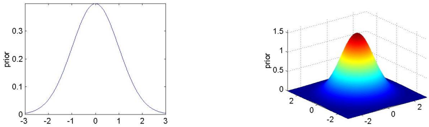
whereas for ridge regression,

$$p(\boldsymbol{\theta}) = \prod_{j=1}^{p} \mathcal{N}(\theta_j; 0, \lambda^2) \, .$$

Laplace prior



Gaussian prior



Posterior Distribution

- Bayes' rule yields the posterior distribution

$$p(\boldsymbol{\theta}|Y, X) \propto p(Y|\boldsymbol{\theta}, X)p(\boldsymbol{\theta}) .$$

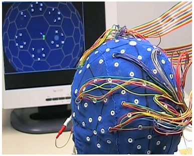
- The solution of Lasso/ridge regression corresponds to the maximum a posteriori solution.
- The Bayesian framework provides a principled approach for computing errorbars, optimizing hyperparameters, incorporating prior knowledge, experiment selection, ...

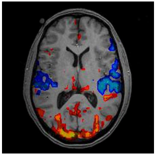
Outline

Brain Reading

Deduce a person's intentions by "reading his brain".

- Brain-computer interfaces based on on-line EEG analysis.
- Classification of brain activity measured through fMRI into prespecified categories.





- Assumption: if a particular voxel/electrode/frequency is relevant, then we expect its neighbors to be relevant as well.
- How do we incorporate such knowledge into our priors?

Scale Mixture Models

- The Laplace distribution (as many others) can be written as a scale mixture distribution:

$$\mathcal{L}(\theta; 0, \lambda) = \int_0^\infty d\sigma^2 \, \mathcal{E}(\sigma^2; 2\lambda) \, \mathcal{N}(\theta; 0, \sigma^2) \, .$$

- We interpret the scales σ_j as the "relevance" or "importance" of feature j: higher σ_j implies more important.
- By coupling the scales, we will couple the relevances.

Multivariate Extension

- Exponential distribution is equivalent to chi-square distribution in two dimensions:

$$\text{if } \{u, v\} \sim \mathcal{N}(0, \lambda) \text{ then } \sigma^2 = u^2 + v^2 \sim \mathcal{E}(2\lambda) \; ,$$

  and thus

$$\mathcal{L}(\theta; 0, \lambda) = \int du \mathcal{N}(u; 0, \lambda) \int dv \mathcal{N}(v; 0, \lambda) \mathcal{N}(\theta; 0, u^2 + v^2).$$

- We define a multivariate Laplace distribution by coupling the u and the v's:

$$\mathcal{L}(\boldsymbol{\theta}; 0, \boldsymbol{\Lambda}) = \int d\mathbf{u} \mathcal{N}(\mathbf{u}; \mathbf{0}, \boldsymbol{\Lambda}) \int d\mathbf{v} \mathcal{N}(\mathbf{v}; \mathbf{0}, \boldsymbol{\Lambda}) \prod_{j=1}^{p} \mathcal{N}(\theta_j; 0, u_j^2 + v_j^2),$$

  with $\boldsymbol{\Lambda}$ a covariance matrix.

The Joint Posterior

$$p(\boldsymbol{\theta}, \mathbf{u}, \mathbf{v}|X, Y) \propto \underbrace{\prod_{i=1}^{N} \mathcal{N}(y_i; \mathbf{x}_i^T \boldsymbol{\theta}, \beta^{-1})}_{\text{likelihood (Gaussian)}}$$

$$\times \underbrace{\prod_{j=1}^{p} \mathcal{N}(\theta_j; 0, u_j^2 + v_j^2)}_{\text{couplings between coefficients and scales}}$$

$$\times \underbrace{\mathcal{N}(\mathbf{u}; \mathbf{0}, \boldsymbol{\Lambda}) \mathcal{N}(\mathbf{v}; \mathbf{0}, \boldsymbol{\Lambda})}_{\text{multi-variate Gaussian}} .$$

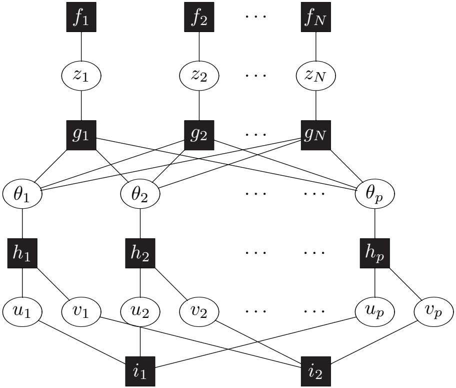
Interested in:

- mean of $\boldsymbol{\theta}$ for mean predictions;
- co-variance of $\boldsymbol{\theta}$ for errorbars;
- variance of $\mathbf{u}$ and $\mathbf{v}$ for relevance.

By symmetry:

- mean of $\mathbf{u}$ and $\mathbf{v}$ are zero;
- co-variance of $\mathbf{u}$ and $\mathbf{v}$ are the same;
- $\mathbf{u}$ and $\mathbf{v}$ are uncorrelated;
- $\{\mathbf{u}, \mathbf{v}\}$ and $\boldsymbol{\theta}$ are uncorrelated.

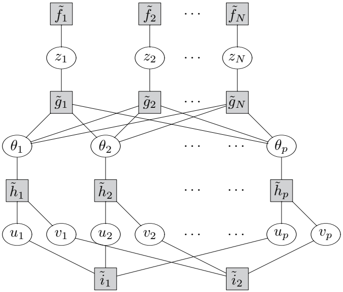
Factor Graph



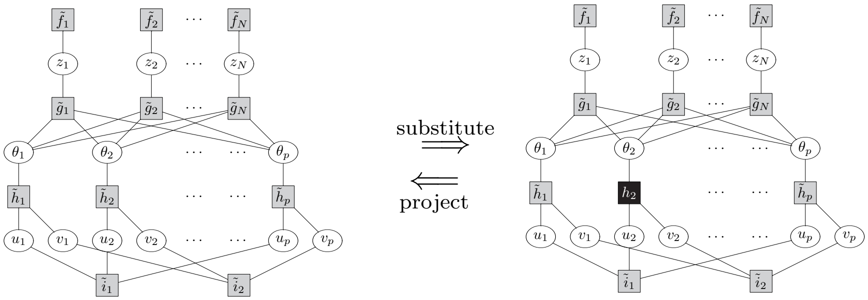
- f_i for the likelihood term corresponding to data point i.
- g_i implements the linear constraint $z_i = \boldsymbol{\theta}^T \mathbf{x}_i$.
- h_j corresponds to the coupling between regression coefficients and scales.
- $i_{1,2}$ represent the multi-variate Gaussians on $\{\mathbf{u}, \mathbf{v}\}$.

Approximate Inference

- The joint posterior is essentially a (possibly huge) Gaussian random field with some (low-dimensional) nonlinear interaction terms.
- Exact inference is intractable, even with independent scales.
- Method of choice: expectation propagation.
- Approximate exact joint posterior by a multi-variate Gaussian.



Expectation Propagation

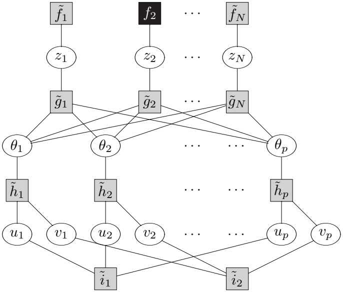


- Iteratively approximate non-Gaussian terms by Gaussian terms.
- Each approximation boils down to (low-dimensional) moment matching.
- Some clever sparse matrix tricks make this computationally doable.
- Main operation: Takahashi procedure for computing the diagonal elements of its inverse.

Cseke and Heskes: Journal of Machine Learning Research, 2011.

Possible Extensions

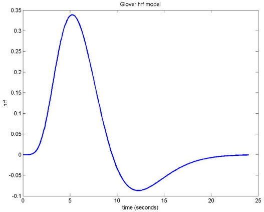
- Logistic regression instead of linear regression:
    - likelihood terms have to be approximated as well;
    - further no essential difference.
- Spike-and-slab prior instead of Laplace prior:
    - scale mixture with two scales;
    - couplings between discrete latent variables or squashed Gaussian variables;
    - similar ideas in Hernández (2×) & Dupont, JMLR 2013.



Outline

fMRI classification

- fMRI activations for different handwritten digits.
- 50 “six” trials and 50 “nine” trials, i.e., $N = 100$.
- Full dataset: 5228 voxels measured over 7 consecutive 2500 ms time steps, i.e., $p \approx 35000$.

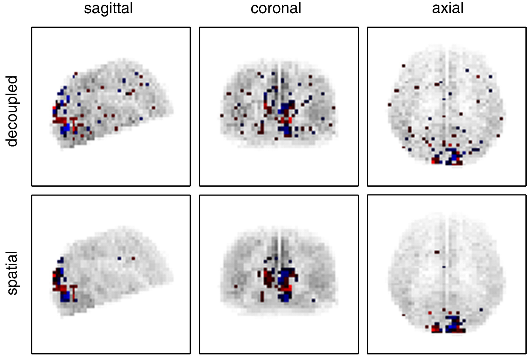


Hemodynamic (**BOLD**) response

Typical data samples

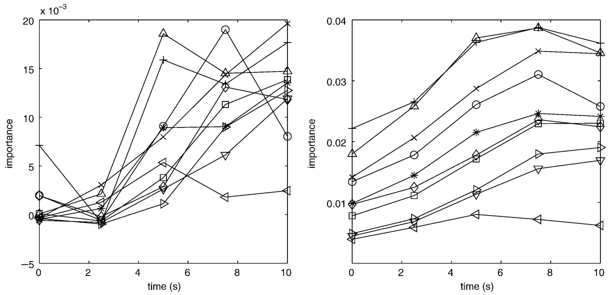
Van Gerven, Cseke, de Lange, and Heskes: Neuroimage, 2010.

Spatial Importance Maps



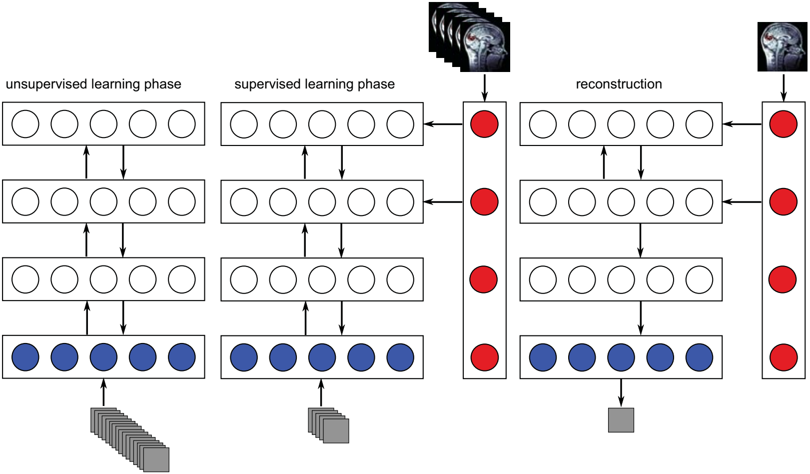
- Data averaged 10 to 15 s after stimulus onset.
- Decoupled (top) versus spatial coupling of scale variables (bottom).
- Most relevant voxels in the occipital lobe (Brodmann Areas 17 and 18).

Importance over Time



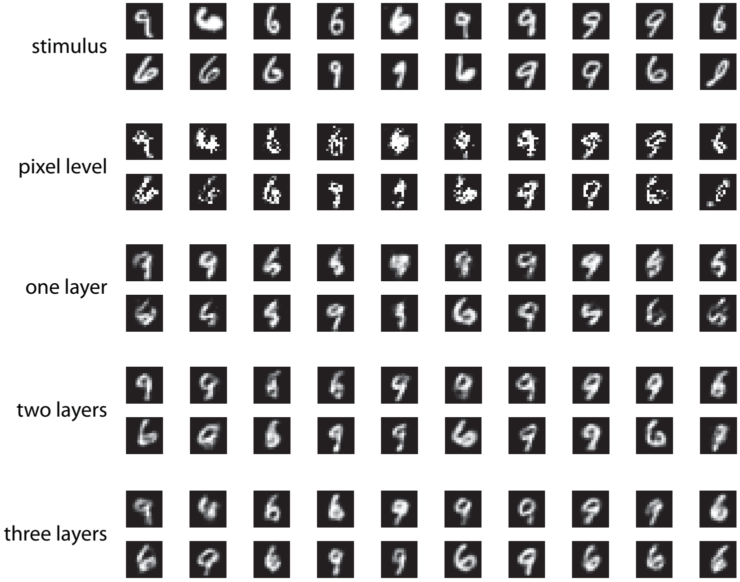
- Importance shown for 10 most relevant voxels.
- Decoupled (left) versus temporal coupling of scale variables (right).
- Increasing importance corresponds with lagged BOLD response.

Intermezzo: Decoding fMRI using DBM's



Van Gerven, de Lange, and Heskes: Neural Computation, 2010.

Reconstructions

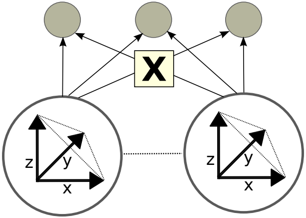


Source Localization

- Sensor readings $\mathbf{Y}$ are related to source currents $\mathbf{S}$ through

$$\mathbf{Y} = \mathbf{X}\mathbf{S} + \text{noise}$$

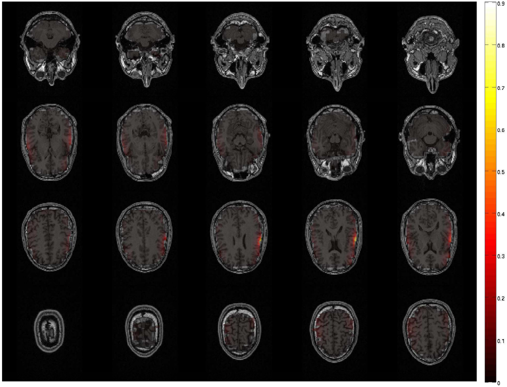
with $\mathbf{X}$ a known lead field matrix, corresponding to the forward model derived from a structural MRI.



- Essentially an (ill-posed) linear regression problem in which $\mathbf{S}$ plays the role of the regression coefficients $\boldsymbol{\theta}$.
- Different regression problems for different time points. $N = 275$, the number of sensors. $p = 1K$ (a bit depending on the discretization), the number of sources.

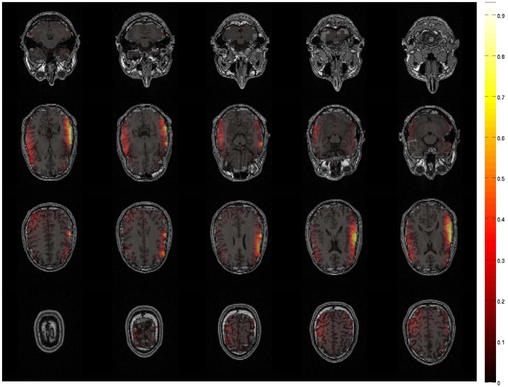
Van Gerven, Cseke, Oostenveld, and Heskes: NIPS, 2009.

Without Constraints



Not so clear

With Spatial Constraints



Sources where you'd expect them to see

Outline

Conclusions and Discussion

Take-home-message:

- A novel way to specify multi-variate sparsifying priors through scale-mixture representations.
- Posterior estimates of these scales can be used for relevance determination.
- Efficient techniques for approximate inference.
- Increased interpretability when analyzing neuroimaging data.

Future directions:

- Extensions to (correlated) spike-and-slab priors.
- Improved stability of expectation propagation.