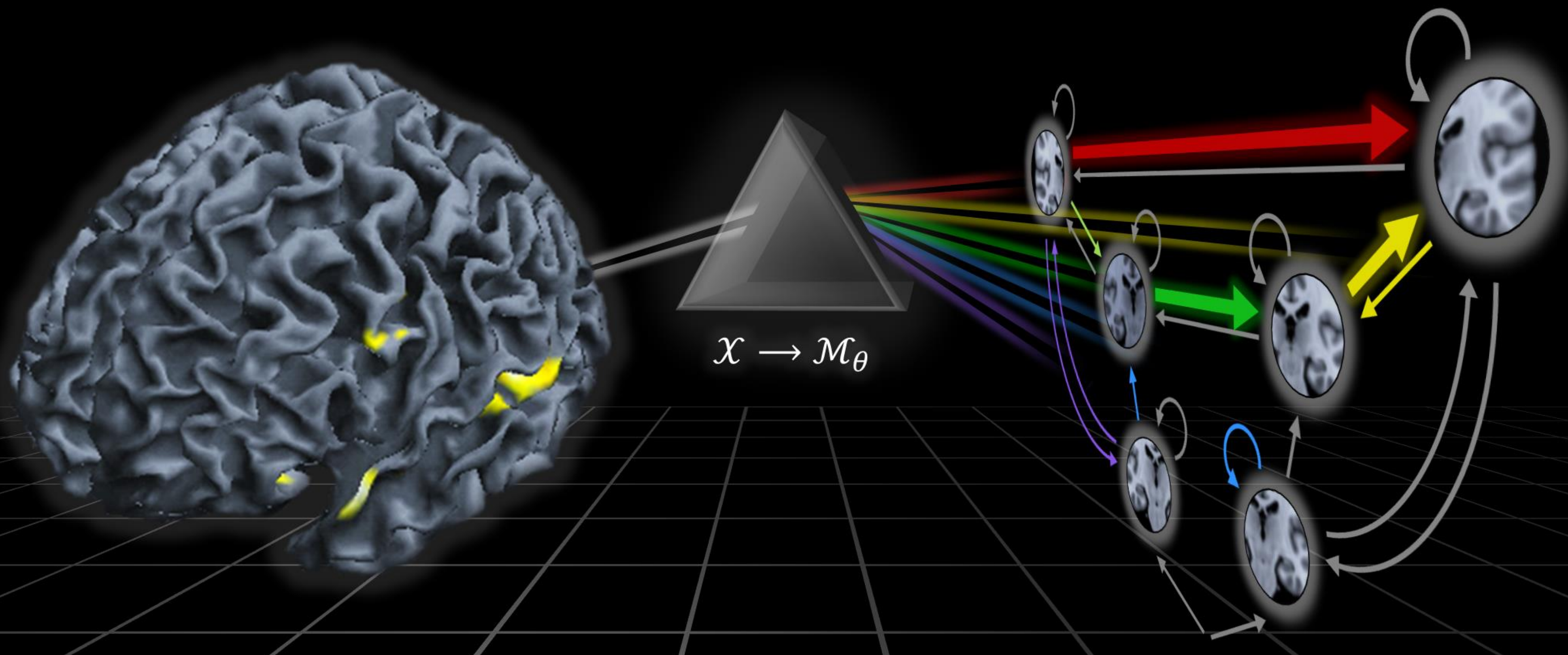
Generative embedding and translational neuromodeling

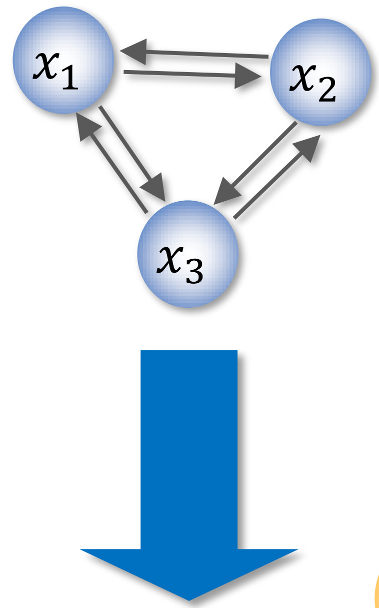
Kay H. Brodersen[1,2]

[1] Translational Neuromodeling Unit (TNU), Institute of Biomedical Engineering, University of Zurich & ETH Zurich
[2] Machine Learning Laboratory, Department of Computer Science, ETH Zurich

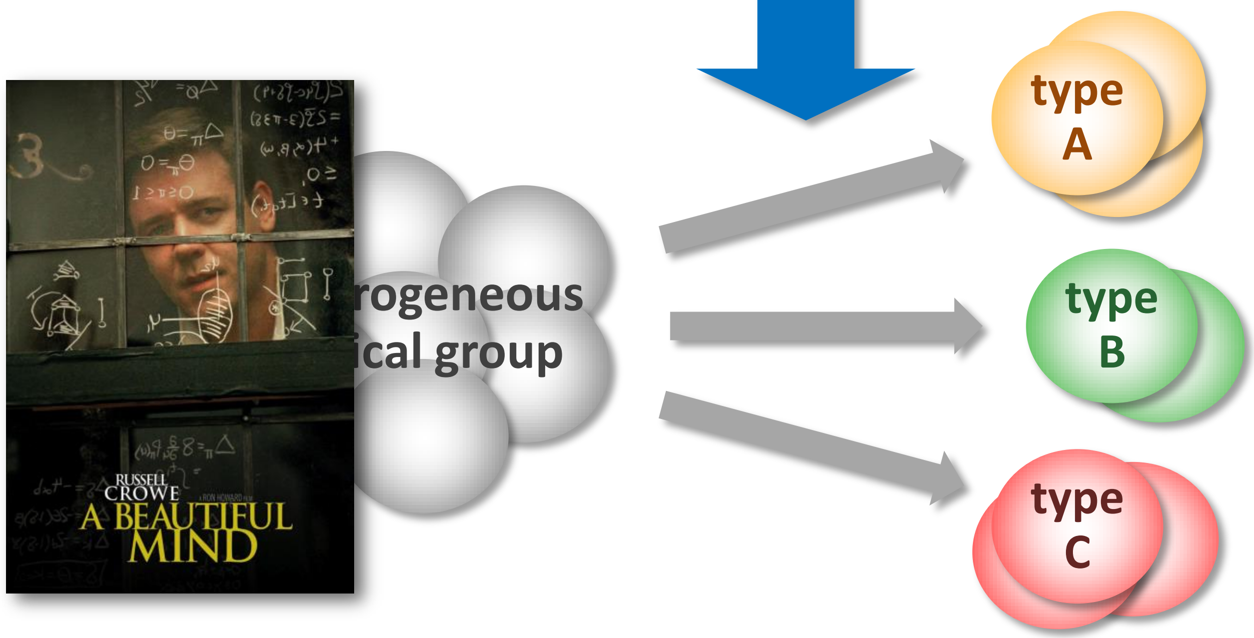


A computational approach to dissecting spectrum disorders

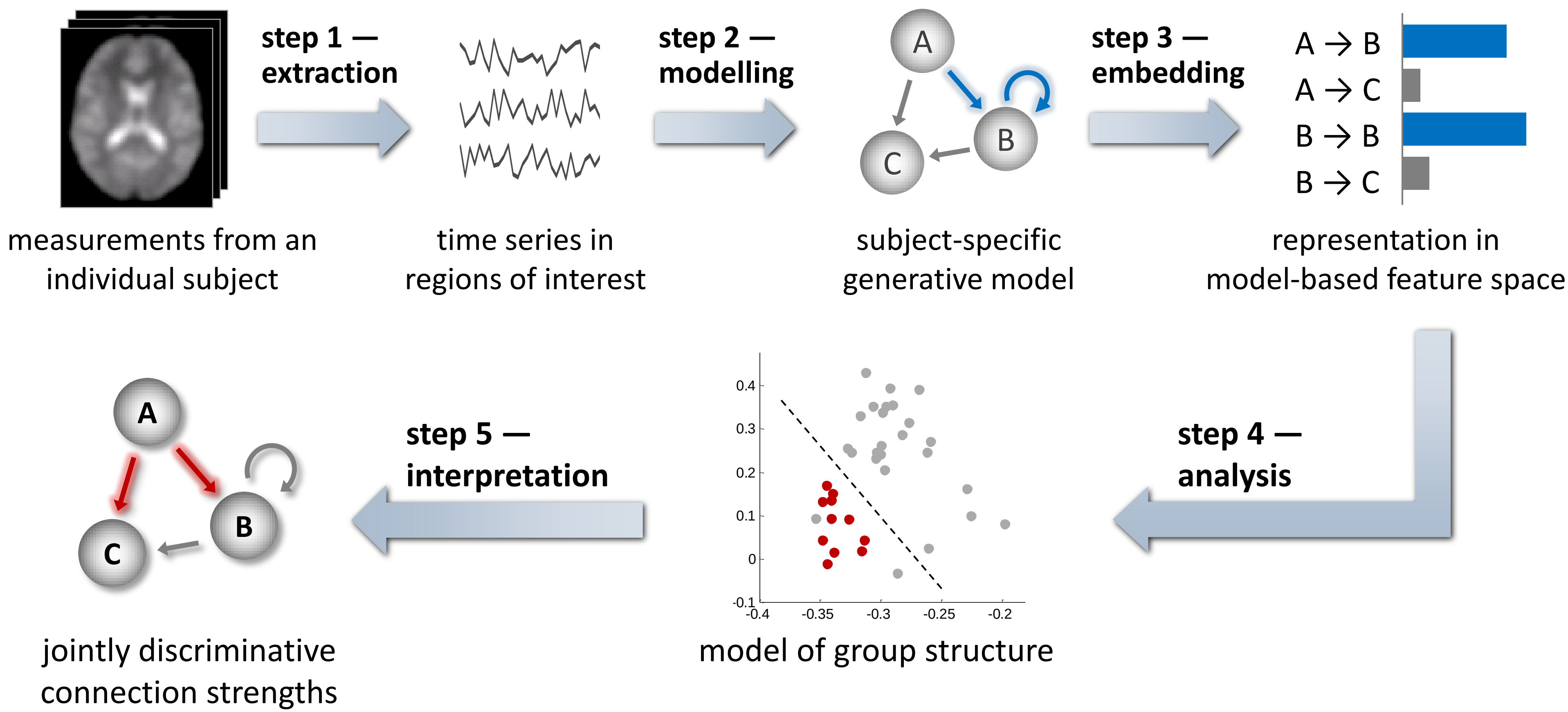
❶ **Psychiatry** lacks pathophysiologically informed diagnostic classifications.



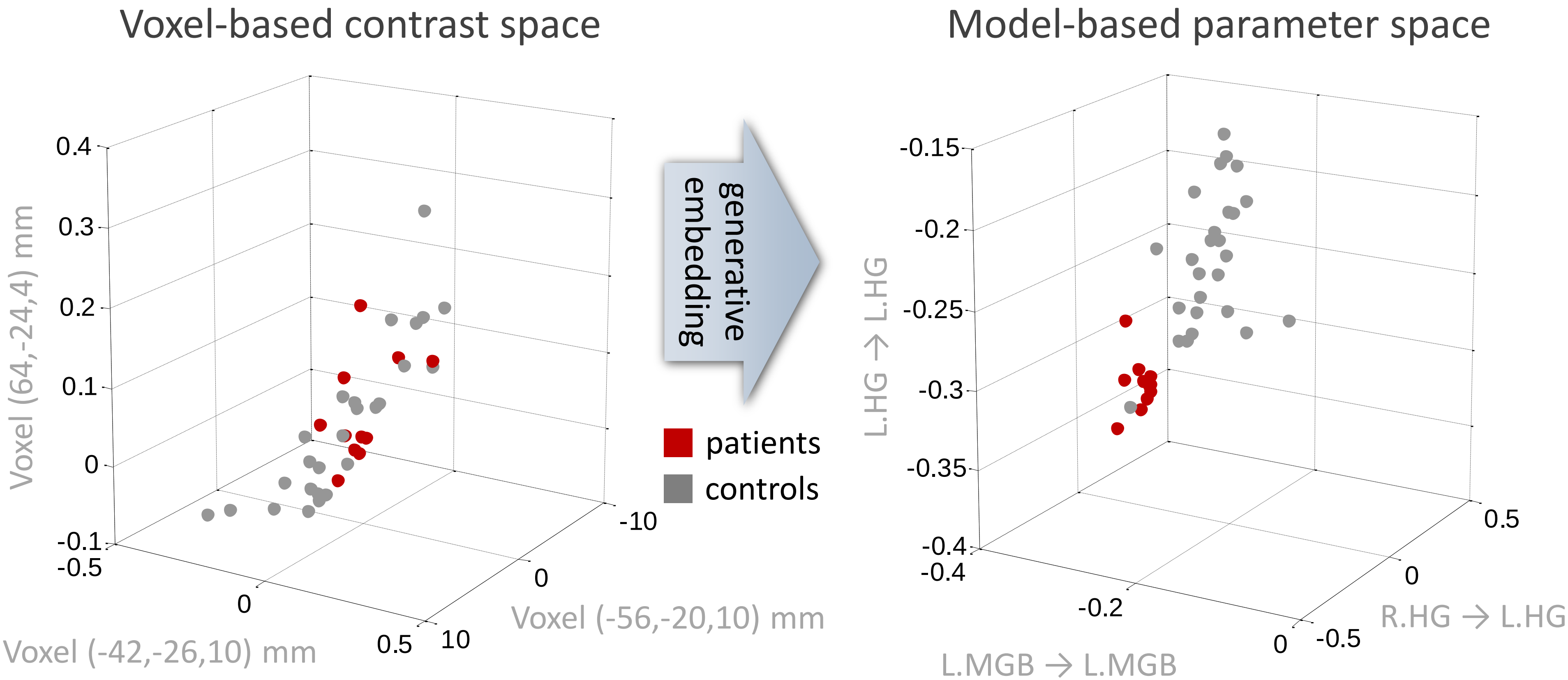
❷ Could **machine learning** help dissect spectrum disorders into mechanistically defined subgroups?



Model-based analysis by generative embedding



The generative projection



classification accuracy
(using all voxels in the regions of interest)
75%

classification accuracy
(using all 23 model parameters)
98%

Colleagues and collaborators

Joachim M. Buhmann

Lorenz Deserno

Ajita Gupta

Alex Leff

Zhihao Lin

Cheng Soon Ong

Will Penny

Florian Schlagenhauf

Tom Schofield

Klaas E. Stephan

Overview

Approximate Bayesian inference

Bayesian inference formalizes *model inversion*, the process of passing from a prior to a posterior in light of data.

$$\underbrace{p(\theta|y)}_{\text{posterior}} = \frac{\overbrace{p(y|\theta)}^{\text{likelihood}}\ \overbrace{p(\theta)}^{\text{prior}}}{\underbrace{\int p(y,\theta)\mathrm{d}\theta}_{\text{marginal likelihood } p(y) \text{ (model evidence)}}}$$

In practice, evaluating the posterior is usually difficult because we cannot easily evaluate $p(y)$, especially when:

- analytical solutions are not available
- numerical integration is too expensive

Approximate Bayesian inference

There are two approaches to approximate inference. They have complementary strengths and weaknesses.

Stochastic approximate inference

in particular sampling

❶ design an algorithm that draws samples $\theta^{(1)}, \dots, \theta^{(m)}$ from $p(\theta|y)$

❷ inspect sample statistics (e.g., histogram, sample quantiles, …)

☑ asymptotically exact
☒ computationally expensive
☒ tricky engineering concerns

Structural approximate inference

in particular variational Bayes

❶ find an analytical proxy $q(\theta)$ that is maximally similar to $p(\theta|y)$

❷ inspect distribution statistics of $q(\theta)$ (e.g., mean, quantiles, intervals, …)

☑ often insightful – and lightning-fast!
☒ often hard work to derive
☒ requires validation via sampling

The Laplace approximation

The Laplace approximation provides a way of approximating a density whose normalization constant we cannot evaluate, by fitting a Gaussian distribution to its mode.

Pierre-Simon Laplace
(1749 – 1827)
French mathematician and astronomer

$$p(z) = \boxed{\frac{1}{Z}} \times \boxed{f(z)}$$

normalization constant (unknown)

main part of the density (easy to evaluate)

This is exactly the situation we face in Bayesian inference:

$$p(\theta|y) = \boxed{\frac{1}{p(y)}} \times \boxed{p(y,\theta)}$$

model evidence (unknown)

joint density (easy to evaluate)

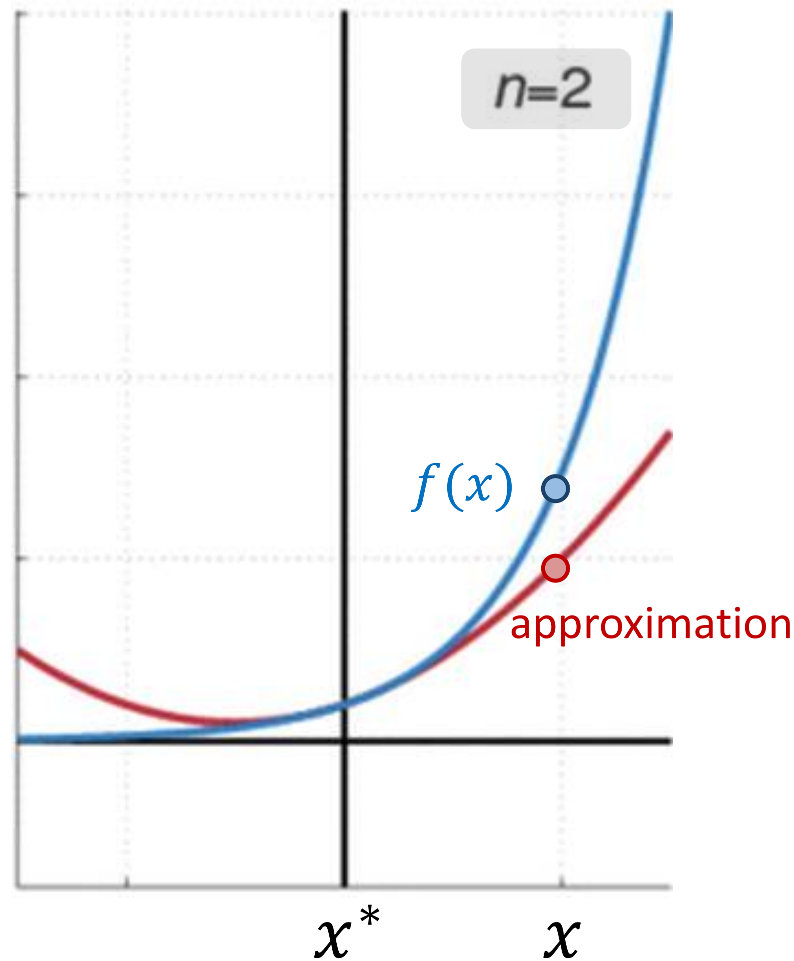
The Taylor approximation

The evaluation of any function $f(x)$ can be approximated by a series:

$$
\begin{aligned}
f(x) \approx\ & f(x^*) \\
& + f'(x^*)(x - x^*) \\
& + \frac{1}{2!} f''(x^*)(x - x^*)^2 \\
& + \frac{1}{3!} f'''(x^*)(x - x^*)^3 \\
& + \cdots
\end{aligned}
$$

Brook Taylor
(1685 – 1731)

English mathematician,
introduced Taylor series



Deriving the Laplace approximation

We begin by expressing the log-joint density $\mathcal{L}(\theta) \equiv \ln p(y, \theta)$ in terms of a second-order Taylor approximation around the mode θ^*:

$$\mathcal{L}(\theta) \approx \mathcal{L}(\theta^*) + \underbrace{\mathcal{L}'(\theta^*)}_{0}(\theta - \theta^*) + \frac{1}{2}\mathcal{L}''(\theta^*)(\theta - \theta^*)^2$$

$$= \mathcal{L}(\theta^*) + \frac{1}{2}\,\mathcal{L}''(\theta^*)\,(\theta - \theta^*)^2$$

This already has the same form as a Gaussian density:

$$\ln \mathcal{N}(\theta|\mu, \eta^{-1}) = \frac{1}{2}\ln \eta - \frac{1}{2}\ln 2\pi - \frac{\eta}{2}(\theta - \mu)^2$$

$$= \frac{1}{2}\ln \frac{\eta}{2\pi} + \frac{1}{2}\,(-\eta)\,(\theta - \mu)^2$$

And so we have an approximate posterior:

$q(\theta) = \mathcal{N}(\theta|\mu, \eta^{-1})$ with $\mu = \theta^*$ (mode of the log-posterior)

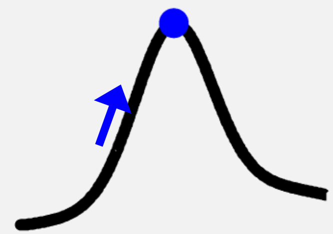
$\eta = -\mathcal{L}''(\theta^*)$ (negative curvature at the mode)

Applying the Laplace approximation

Given a model with parameters $\theta = (\theta_1, \dots, \theta_p)$, the Laplace approximation reduces to a simple three-step procedure:

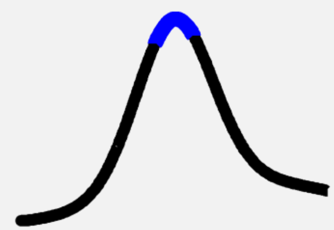
1. Find the mode of the log-joint:

   $$\theta^* = \arg\max_\theta \ \ln p(y, \theta)$$



2. Evaluate the curvature of the log-joint at the mode:

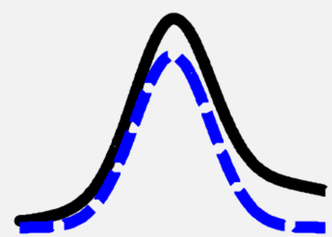
   $$\nabla\nabla \ln p(y, \theta^*)$$



3. We obtain a Gaussian approximation:

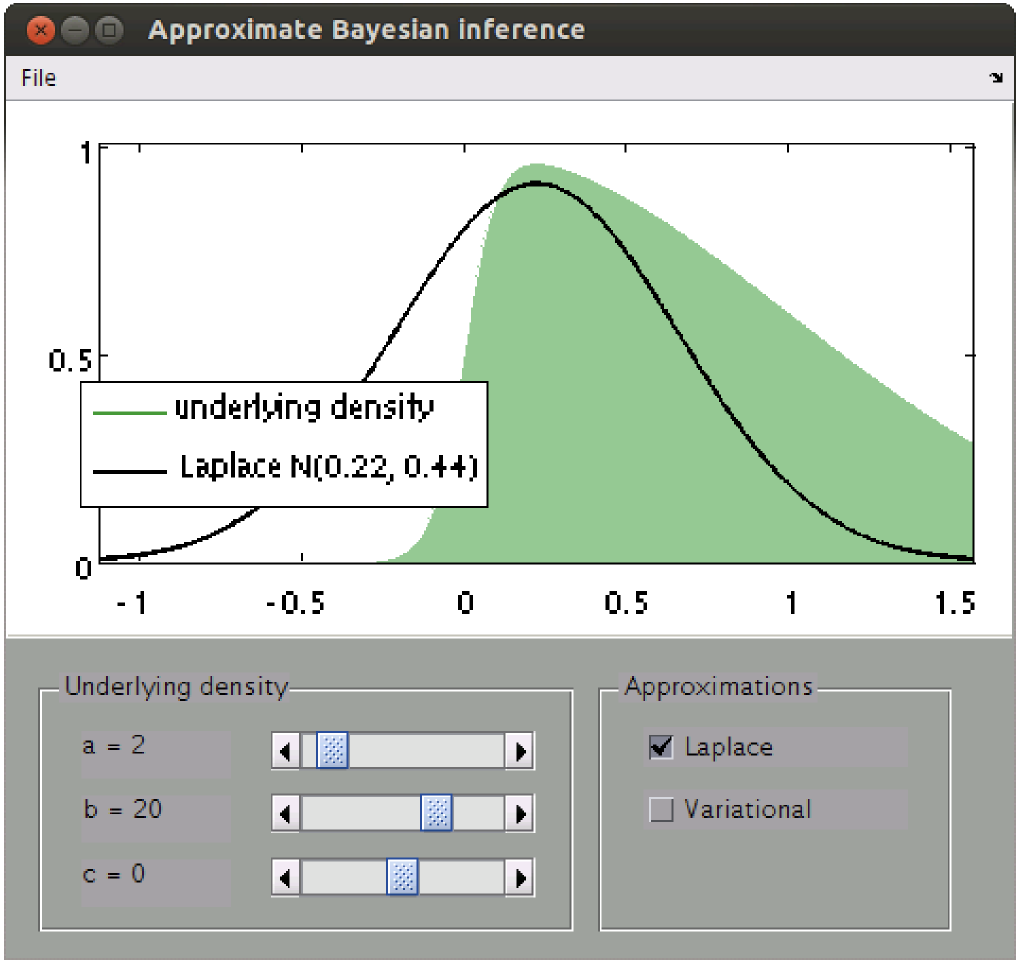
   $\mathcal{N}(\theta | \mu, \Lambda^{-1})$ with $\mu = \theta^*$

   $$\Lambda = -\nabla\nabla \ln p(y, \theta^*)$$



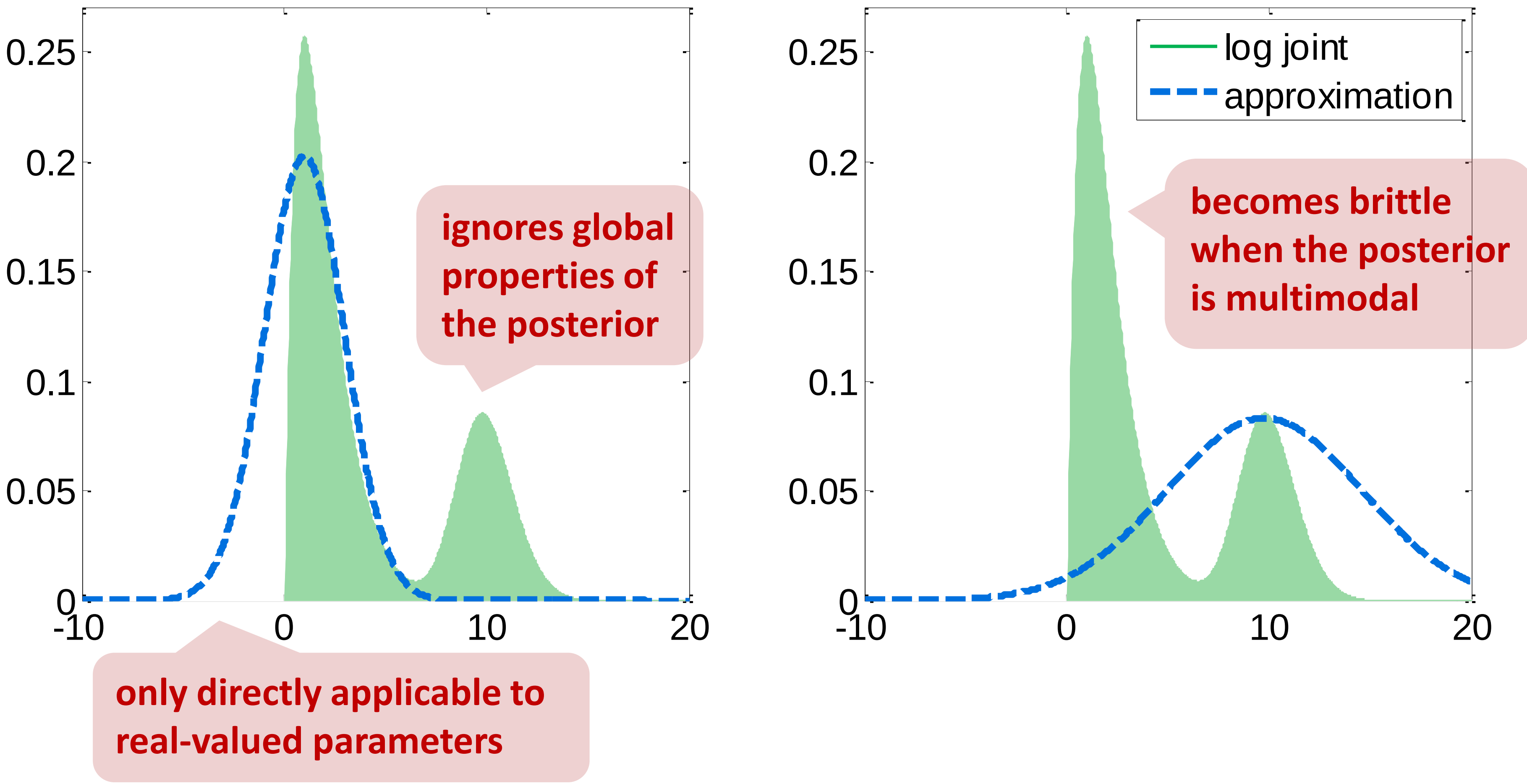
The Laplace approximation: demo

~kbroders/teaching/vb_gui.m



Limitations of the Laplace approximation

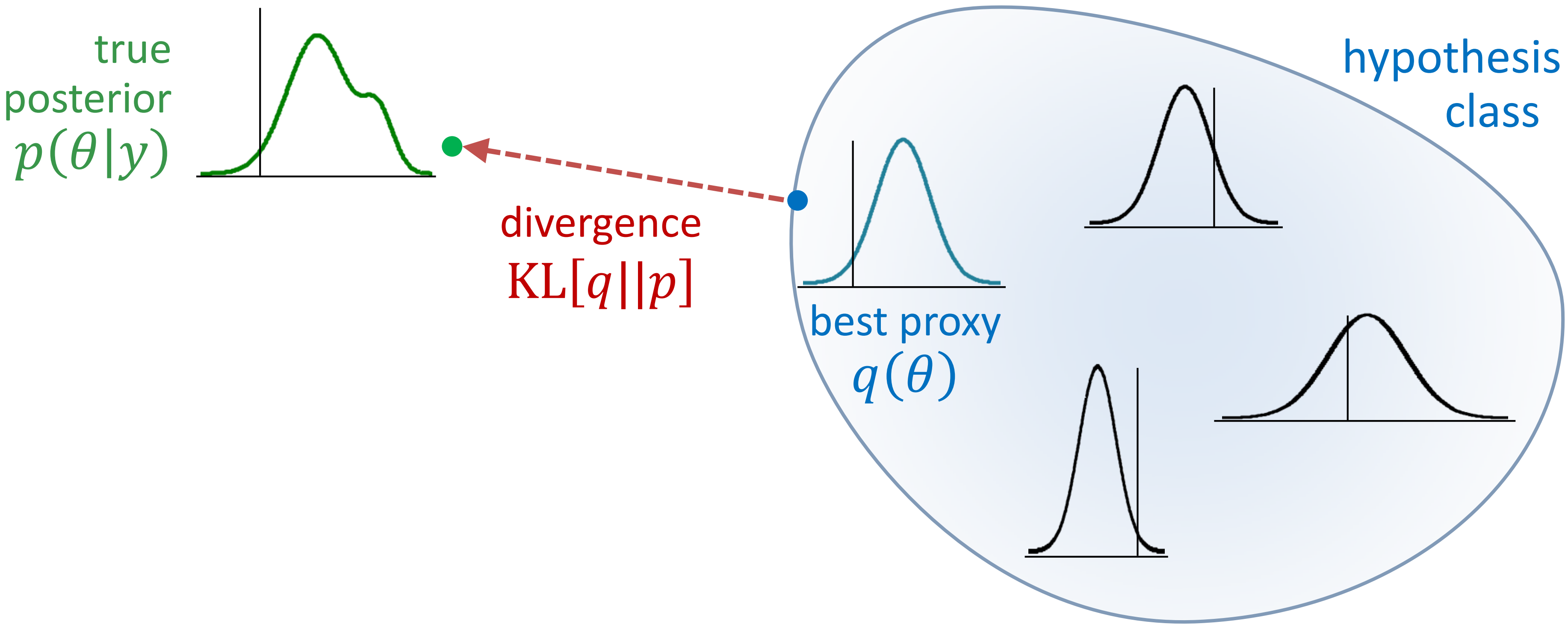
The Laplace approximation is often too strong a simplification.



Overview

Variational Bayesian inference

Variational Bayesian (VB) inference generalizes the idea behind the Laplace approximation. In VB, we wish to find an approximate density that is maximally similar to the true posterior.



Variational calculus

Variational Bayesian inference is based on variational calculus.

Standard calculus

Newton, Leibniz, and others

- functions $f: x \mapsto f(x)$
- derivatives $\frac{\mathrm{d}f}{\mathrm{d}x}$

Example: maximize the likelihood expression $p(y|\theta)$ w.r.t. θ

Variational calculus

Euler, Lagrange, and others

- functionals $F: f \mapsto F(f)$
- derivatives $\frac{\mathrm{d}F}{\mathrm{d}f}$

Example: maximize the entropy $H[p]$ w.r.t. a probability distribution $p(x)$

Leonhard Euler
(1707 – 1783)

Swiss mathematician, ‘Elementa Calculi Variationum’

Variational calculus and the free energy

Variational calculus lends itself nicely to approximate Bayesian inference.

$$
\begin{aligned}
\ln p(y) &= \ln \frac{p(y,\theta)}{p(\theta|y)} \\
&= \int q(\theta) \ln \frac{p(y,\theta)}{p(\theta|y)} \, \mathrm{d}\theta \\
&= \int q(\theta) \ln \frac{p(y,\theta)}{p(\theta|y)} \frac{q(\theta)}{q(\theta)} \, \mathrm{d}\theta \\
&= \int q(\theta) \left(\ln \frac{q(\theta)}{p(\theta|y)} + \ln \frac{p(y,\theta)}{q(\theta)} \right) \mathrm{d}\theta \\
&= \underbrace{\int q(\theta) \ln \frac{q(\theta)}{p(\theta|y)} \, \mathrm{d}\theta}_{\substack{\mathrm{KL}[q||p] \\ \text{divergence between} \\ q(\theta) \text{ and } p(\theta|y)}} + \underbrace{\int q(\theta) \ln \frac{p(y,\theta)}{q(\theta)} \, \mathrm{d}\theta}_{\substack{F(q,y) \\ \text{free energy}}}
\end{aligned}
$$

Variational calculus and the free energy

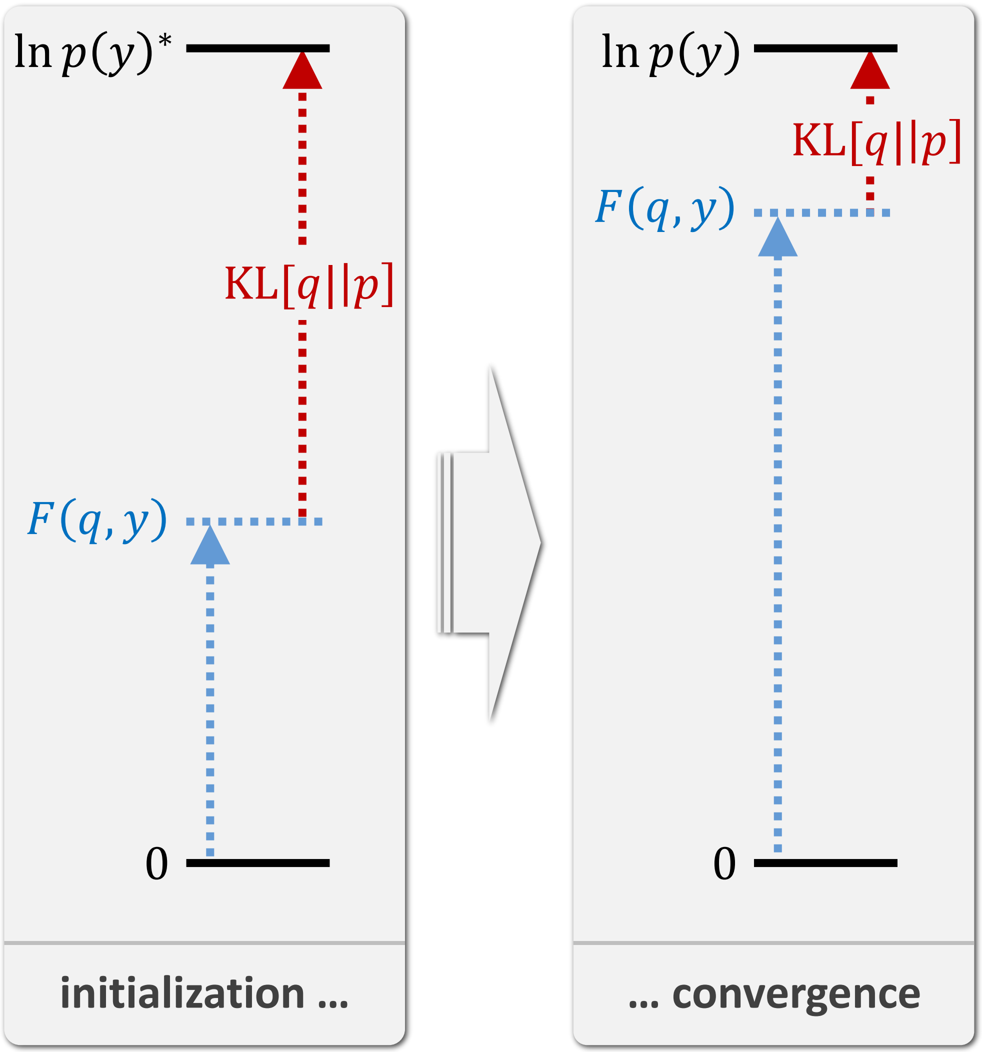
In summary, the log model evidence can be expressed as:

$$\ln p(y) = \underbrace{\mathrm{KL}[q||p]}_{\substack{\textbf{divergence} \\ \geq\ 0 \\ \text{(unknown)}}} + \underbrace{F(q,y)}_{\substack{\textbf{free energy} \\ \text{(easy to evaluate} \\ \text{for a given } q)}}$$

Maximizing $F(q,y)$ is equivalent to:

- minimizing $\mathrm{KL}[q||p]$
- tightening $F(q,y)$ as a lower bound to the log model evidence

* In this illustrative example, the log model evidence and the free energy are positive; but the above equivalences hold just as well when the log model evidence is negative.



Computing the free energy

We can decompose the free energy $F(q, y)$ as follows:

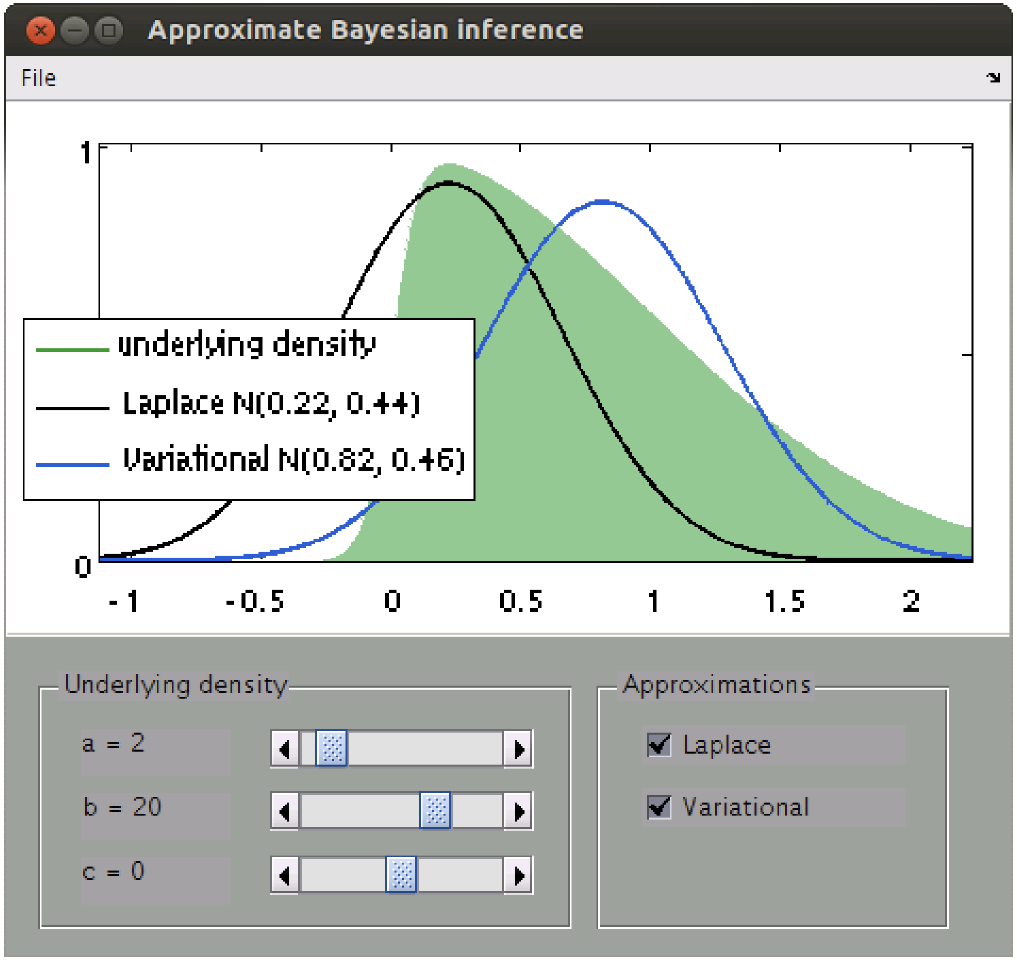
$$F(q, y) = \int q(\theta) \ln \frac{p(y, \theta)}{q(\theta)} \mathrm{d}\theta$$

$$= \int q(\theta) \ln p(y, \theta) \; d\theta - \int q(\theta) \ln q(\theta) \; d\theta$$

$$= \underbrace{\langle \ln p(y, \theta) \rangle_q}_{\textbf{expected log-joint}} + \underbrace{H[q]}_{\textbf{Shannon entropy}}$$

Variational Bayes: demo

~kbroders/teaching/vb_gui.m

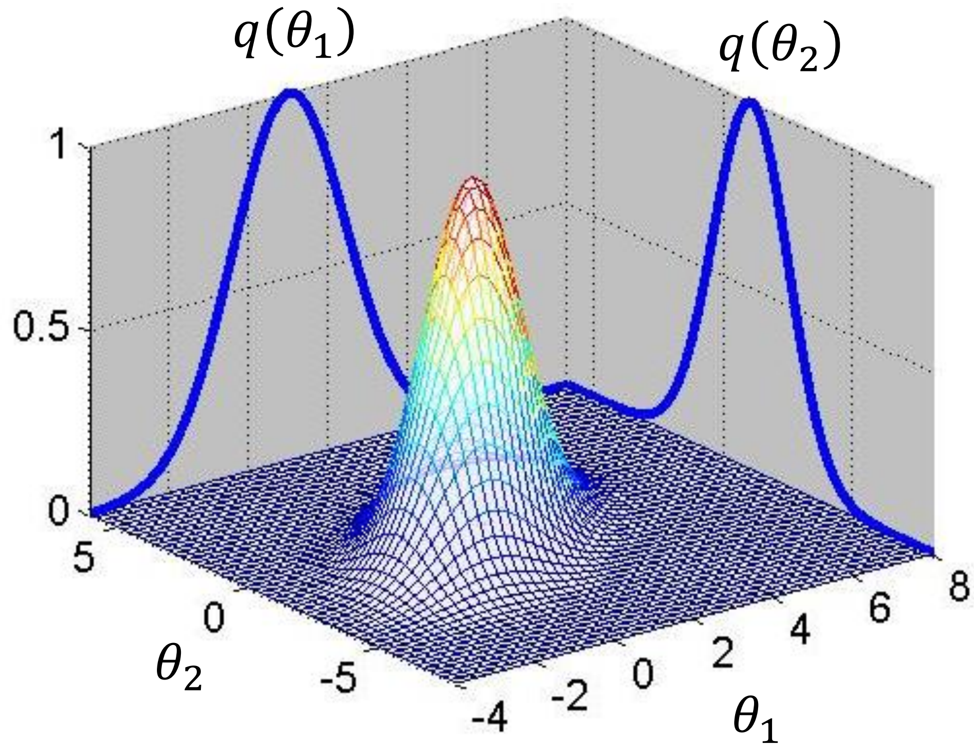


The mean-field assumption

When inverting models with several parameters, a common way of restricting the class of approximate posteriors $q(\theta)$ is to consider those posteriors that factorize into independent partitions,

$$q(\theta) = \prod_i q_i(\theta_i)\,,$$

where $q_i(\theta_i)$ is the approximate posterior for the i^{th} subset of parameters.



Jean Daunizeau, www.fil.ion.ucl.ac.uk/~jdaunize/presentations/Bayes2.pdf

Variational inference under the mean-field assumption

$$
\begin{aligned}
F(q,y) &= \int q(\theta)\ln\frac{p(y,\theta)}{q(\theta)}\,\mathrm{d}\theta \\
&= \int \prod_i q_i \times \left(\ln p(y,\theta) - \sum_i \ln q_i\right)\mathrm{d}\theta \\
&= \int q_j \prod_{\setminus j} q_i \left(\ln p(y,\theta) - \ln q_j\right)\mathrm{d}\theta - \int q_j \prod_{\setminus j} q_i \sum_{\setminus j} \ln q_i\,\mathrm{d}\theta \\
&= \int q_j \left(\underbrace{\int \prod_{\setminus j} q_i \ln p(y,\theta)\,\mathrm{d}\theta_{\setminus j}}_{\langle \ln p(y,\theta)\rangle_{q_{\setminus j}}} - \ln q_j\right)\mathrm{d}\theta_j - \int q_j \int \prod_{\setminus j} q_i \ln \prod_{\setminus j} q_i\,\mathrm{d}\theta_{\setminus j}\,\mathrm{d}\theta_j \\
&= \int q_j \ln\frac{\exp\left(\langle \ln p(y,\theta)\rangle_{q_{\setminus j}}\right)}{q_j}\,\mathrm{d}\theta_j \quad + \quad c \\
&= -\mathrm{KL}\left[q_j \,||\, \exp\left(\langle \ln p(y,\theta)\rangle_{q_{\setminus j}}\right)\right] \quad + \quad c
\end{aligned}
$$

mean-field assumption:
$q(\theta) = \prod_i q_i(\theta_i)$

Typical strategies in variational inference

	no parametric assumptions	parametric assumptions $q(\theta) = F(\theta\|\delta)$
no mean-field assumption	(variational inference = exact inference)	fixed-form optimization of moments
mean-field assumption $q(\theta) = \prod q(\theta_i)$	iterative free-form variational optimization	iterative fixed-form variational optimization

Variational algorithm under the mean-field assumption

We can rewrite the free energy as:

$$F(q,y) = -\mathrm{KL}\left[q_j || \exp\left(\langle \ln p(y,\theta)\rangle_{q_{\setminus j}}\right)\right] + c$$

Suppose the densities $q_{\setminus j} \equiv q\left(\theta_{\setminus j}\right)$ are kept fixed. Then the approximate posterior $q\left(\theta_j\right)$ that maximizes $F(q,y)$ is given by:

$$\begin{aligned} q_j^* &= \arg\max_{q_j} F(q,y) \\ &= \frac{1}{Z}\exp\left(\langle \ln p(y,\theta)\rangle_{q_{\setminus j}}\right) \end{aligned}$$

Therefore:

$$\ln q_j^* = \underbrace{\langle \ln p(y,\theta)\rangle_{q_{\setminus j}}}_{=:I\left(\theta_j\right)} - \ln Z$$

This implies a straightforward algorithm for variational inference:

❶ Initialize all approximate posteriors $q(\theta_i)$, e.g., by setting them to their priors.

❷ Cycle over the parameters, revising each given the current estimates of the others.

❸ Loop until convergence.

Application: variational linear regression

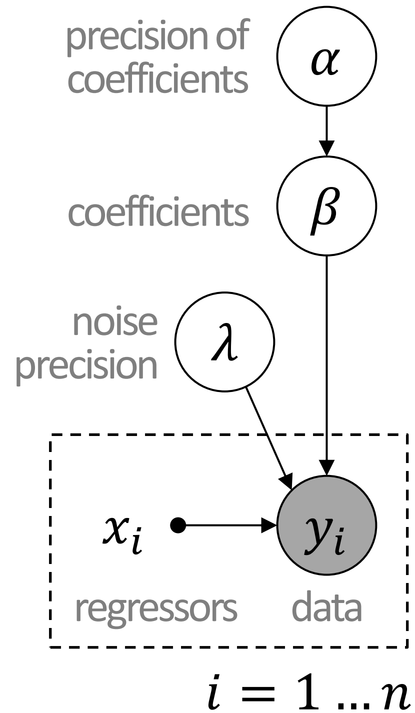
We consider a multiple linear regression model with a shrinkage prior on the regression coefficients.

We wish to infer on the coefficients β, their precision α, and the noise precision λ. There is no analytical posterior

$$p(\beta, \alpha, \lambda | y).$$

We therefore seek a variational approximation:

$$q(\beta, \alpha, \lambda) = q_\beta(\beta)\, q_\alpha(\alpha)\, q_\lambda(\lambda).$$



$$p(\alpha) = \mathrm{Ga}(\alpha | a_0, b_0)$$

$$p(\beta | \alpha) = \mathcal{N}_d(\beta | 0, \alpha^{-1} I_{d \times d})$$

$$p(\lambda) = \mathrm{Ga}(\lambda | c_0, d_0)$$

$$p(y_i | \beta) = \mathcal{N}(y_i | \beta^T x_i, \lambda^{-1})$$

Variational linear regression: coefficients precision α

$$
\begin{aligned}
\ln q^*(\alpha) &= \langle \ln p(y,\beta,\alpha,\lambda)\rangle_{q(\beta,\lambda)} + c \\
&= \underbrace{\langle \ln \textstyle\prod \mathcal{N}(y_i|\beta^T x_i, \lambda^{-1})\rangle_{q(\beta)q(\lambda)}}_{c} + \langle \ln \mathcal{N}_d(\beta|0,\alpha^{-1}I)\rangle_{q(\beta)q(\lambda)} + \langle \ln \mathrm{Ga}(\alpha|a_0,b_0)\rangle_{q(\beta)q(\lambda)} + c \\
&= \left\langle -\frac{1}{2}\ln \underbrace{|\alpha^{-1}I|}_{\alpha^{-d}} \underbrace{-\frac{d}{2}\ln 2\pi}_{c} - \frac{1}{2}(\beta-0)^T \alpha I(\beta-0)\right\rangle_{q(\beta)} \\
&\qquad + \langle a_0 \ln b_0 - \ln\Gamma(a_0) + (a_0-1)\ln\alpha - b_0\alpha\rangle_{q(\beta)} + c \\
&= \frac{d}{2}\ln\alpha - \frac{\alpha}{2}\langle\beta^T\beta\rangle_{q(\beta)} + (a_0-1)\ln\alpha - b_0\alpha + c \\
&= \left(\frac{d}{2} + a_0 - 1\right)\ln\alpha \;-\; \left(\frac{1}{2}\langle\beta^T\beta\rangle_{q(\beta)} + b_0\right)\alpha \;+ c
\end{aligned}
$$

$\Longrightarrow q^*(\alpha) = \mathrm{Ga}(\alpha|a_n, b_n)$ with

$$a_n = a_0 + \frac{d}{2}$$

$$b_n = b_0 + \frac{1}{2}\langle\beta^T\beta\rangle_{q(\beta)}$$

Variational linear regression: cofficients β

$$
\begin{aligned}
\ln q^*(\beta) &= \langle \ln p(y,\beta,\alpha,\lambda)\rangle_{q(\alpha,\lambda)} + c \\
&= \langle \ln \prod \mathcal{N}(y_i|\beta^T x_i, \lambda^{-1})\rangle_{q(\alpha)q(\lambda)} + \langle \ln \mathcal{N}_d(\beta|0,\alpha^{-1} I)\rangle_{q(\alpha)q(\lambda)} + \underbrace{\langle \ln \mathrm{Ga}(\alpha|a_0,b_0)\rangle_{q(\alpha)q(\lambda)}}_{c} + c \\
&= \sum_i^n \left\langle \underbrace{\frac{1}{2}\ln\lambda}_{c} - \underbrace{\frac{1}{2}\ln 2\pi}_{c} - \frac{\lambda}{2}(y_i - \beta^T x_i)^2 \right\rangle_{q(\alpha)q(\lambda)} + \left\langle \underbrace{-\frac{1}{2}\ln|\alpha^{-1} I|}_{c} - \underbrace{\frac{d}{2}\ln 2\pi}_{c} - \frac{1}{2}\beta^T \alpha I \beta \right\rangle_{q(\alpha)} + c \\
&= -\frac{\langle\lambda\rangle_{q(\lambda)}}{2} \sum_i^n (y_i - \beta^T x_i)^2 - \frac{1}{2}\langle\alpha\rangle_{q(\alpha)} \beta^T \beta + c \\
&= \underbrace{-\frac{\langle\lambda\rangle_{q(\lambda)}}{2} y^T y}_{c} + \langle\lambda\rangle_{q(\lambda)} \beta^T X^T y - \frac{\langle\lambda\rangle_{q(\lambda)}}{2} \beta^T X^T X \beta - \frac{1}{2}\beta^T \langle\alpha\rangle_{q(\alpha)} I \beta + c \\
&= -\frac{1}{2}\beta^T \boxed{\left\{\langle\lambda\rangle_{q(\lambda)} X^T X + \langle\alpha\rangle_{q(\alpha)} I\right\}} \beta + \beta^T \boxed{\langle\lambda\rangle_{q(\lambda)} X^T y} + c
\end{aligned}
$$

$\Longrightarrow q^*(\beta) = \mathcal{N}_d(\beta|\mu_n, \Lambda_n^{-1})$ with $\Lambda_n = \langle\alpha\rangle_{q(\alpha)} I + \langle\lambda\rangle_{q(\lambda)} X^T X,$ $\mu_n = \Lambda_n^{-1}\langle\lambda\rangle_{q(\lambda)} X^T y$

Variational linear regression: noise precision λ

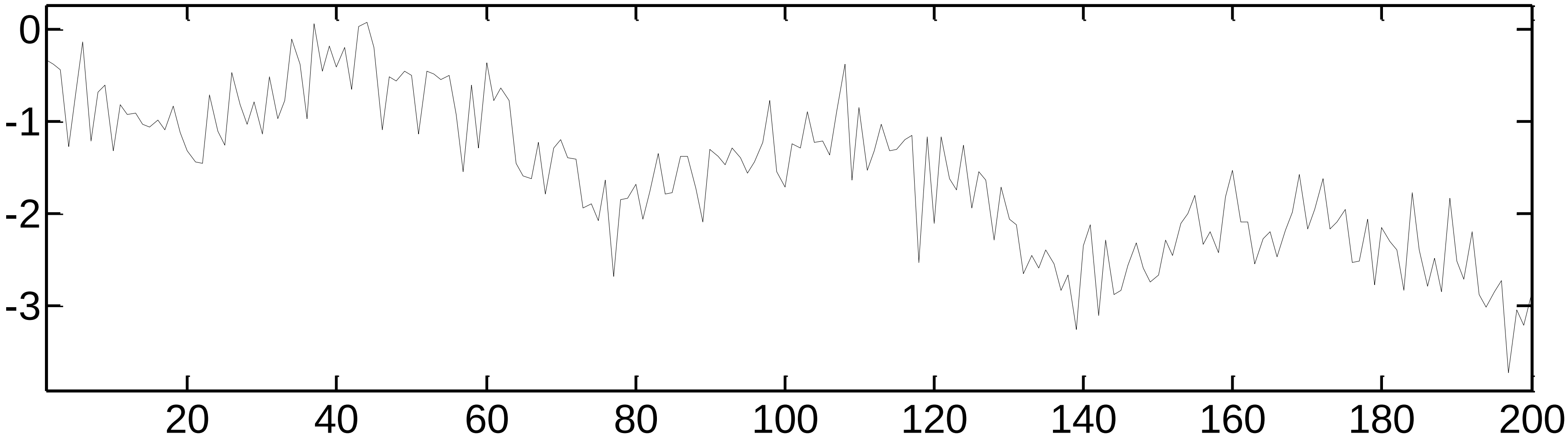
$$
\begin{aligned}
\ln q^*(\lambda) &= \langle \ln p(y,\beta,\alpha,\lambda)\rangle_{q(\beta,\alpha)} + c \\
&= \left\langle \sum_i^n \frac{1}{2}\ln\lambda - \underbrace{\frac{1}{2}\ln 2\pi}_{c} - \frac{\lambda}{2}(y_i - \beta^T x_i)^2 \right\rangle_{q(\beta)q(\alpha)} \\
&\qquad + \left\langle \underbrace{c_0 \ln d_0}_{c} - \underbrace{\ln\Gamma(c_0)}_{c} + (c_0 - 1)\ln\lambda - d_0\lambda \right\rangle_{q(\beta)q(\alpha)} + c \\
&= \frac{n}{2}\ln\lambda - \frac{\lambda}{2} y^T y + \lambda\langle\beta\rangle_{q(\beta)}^T X^T y - \frac{\lambda}{2}\langle\beta\rangle_{q(\beta)}^T X^T X \langle\beta\rangle_{q(\beta)} + (c_0 - 1)\ln\lambda - d_0\lambda + c \\
&= \boxed{\left\{c_0 + \frac{n}{2} - 1\right\}} \ln\lambda - \boxed{\left\{\frac{1}{2}y^T y - \langle\beta\rangle_{q(\beta)}^T X^T y + \frac{1}{2}\langle\beta\rangle_{q(\beta)}^T X^T X\langle\beta\rangle_{q(\beta)} + d_0\right\}} \lambda + c
\end{aligned}
$$

$$
\Longrightarrow q^*(\lambda) = \mathrm{Ga}(\lambda | c_n, d_n), \qquad c_n = c_0 + \frac{n}{2}
$$

$$
d_n = d_0 + \frac{1}{2}y^T y - \langle\beta\rangle_{q(\beta)}^T X^T y + \frac{1}{2}\langle\beta\rangle_{q(\beta)}^T X^T X\langle\beta\rangle_{q(\beta)}
$$

Variational linear regression: example

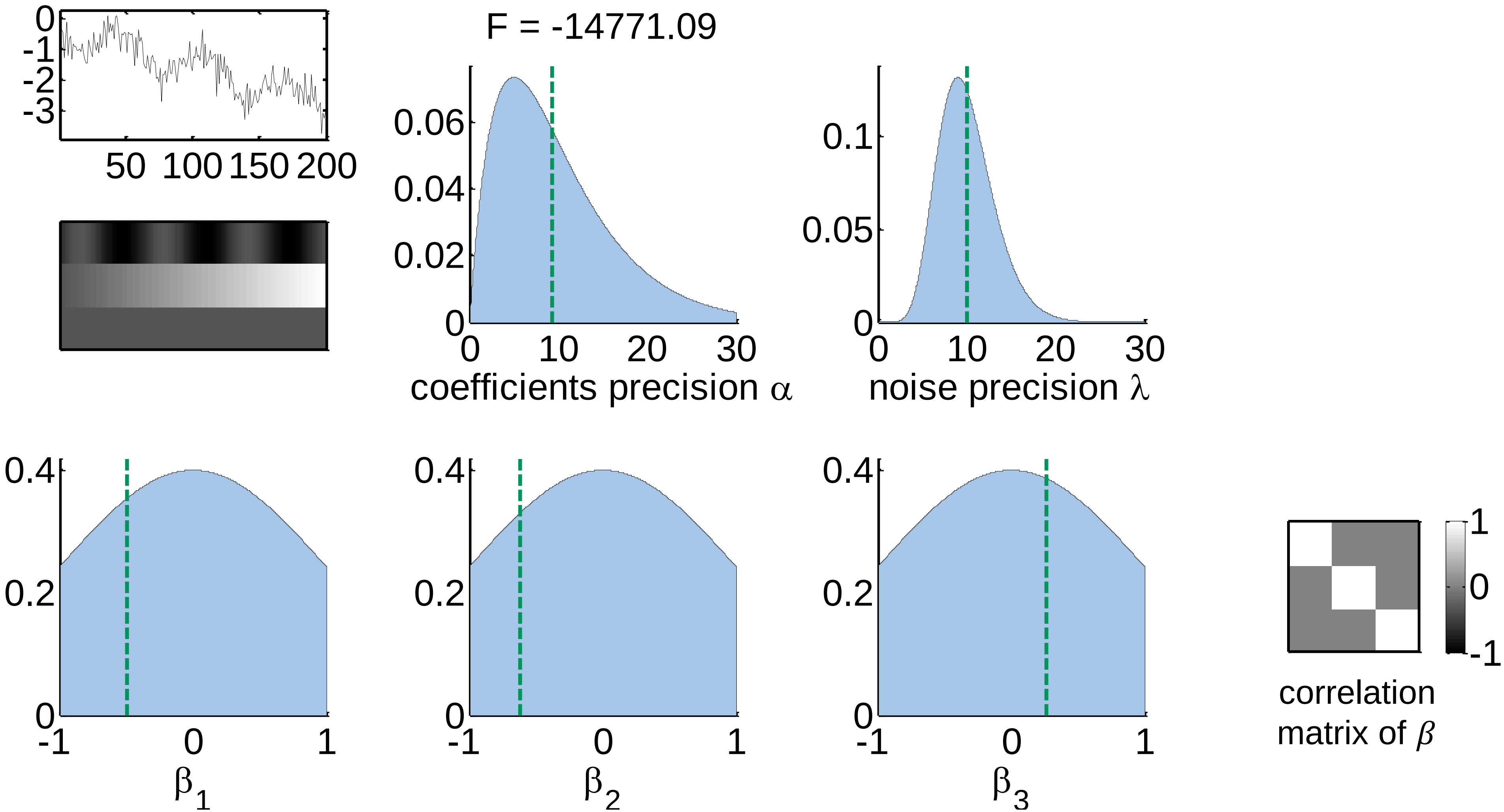
Data y^T



Design matrix X^T

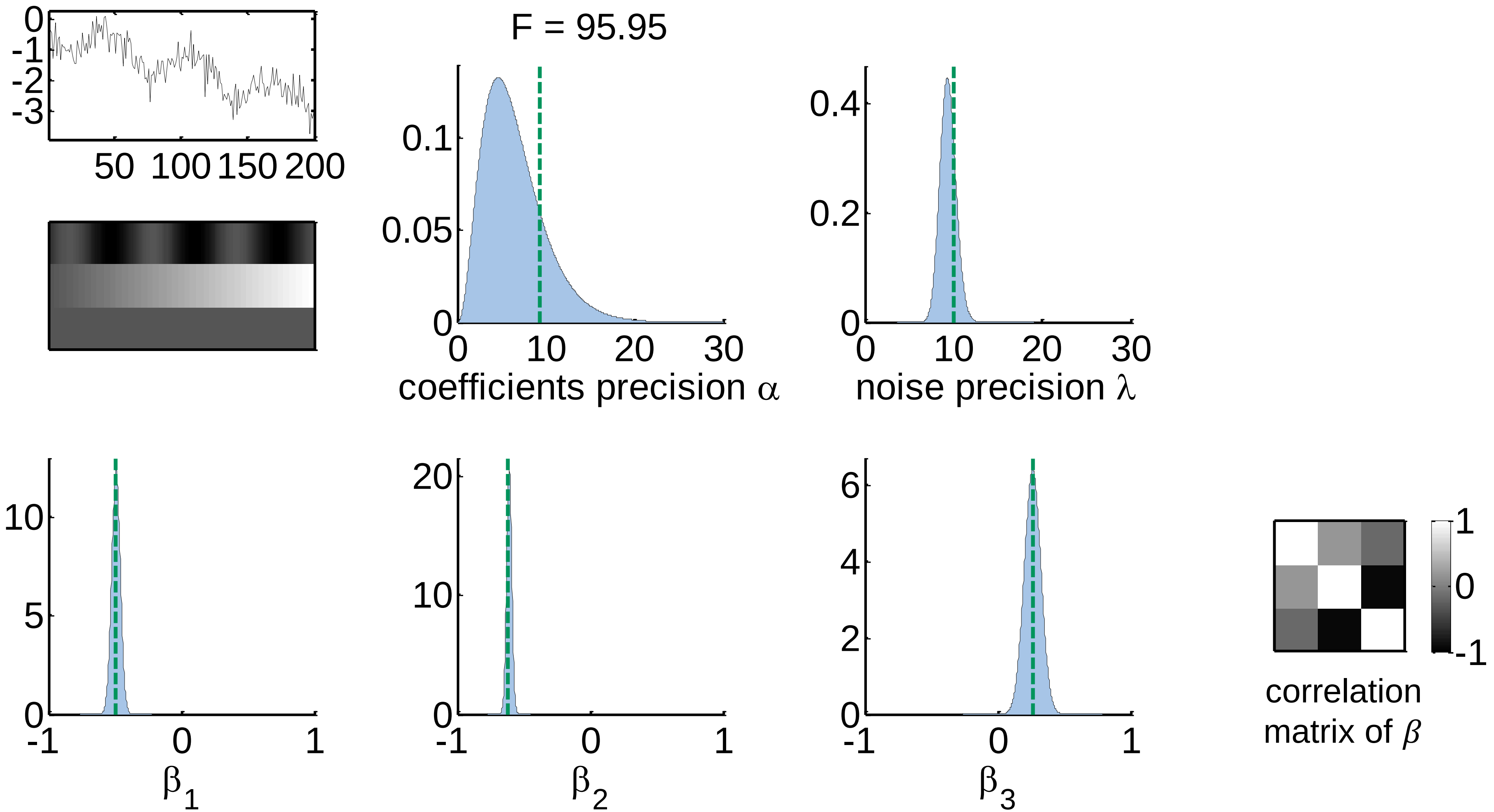
Variational linear regression: example

Iteration 0 0:00:00'000



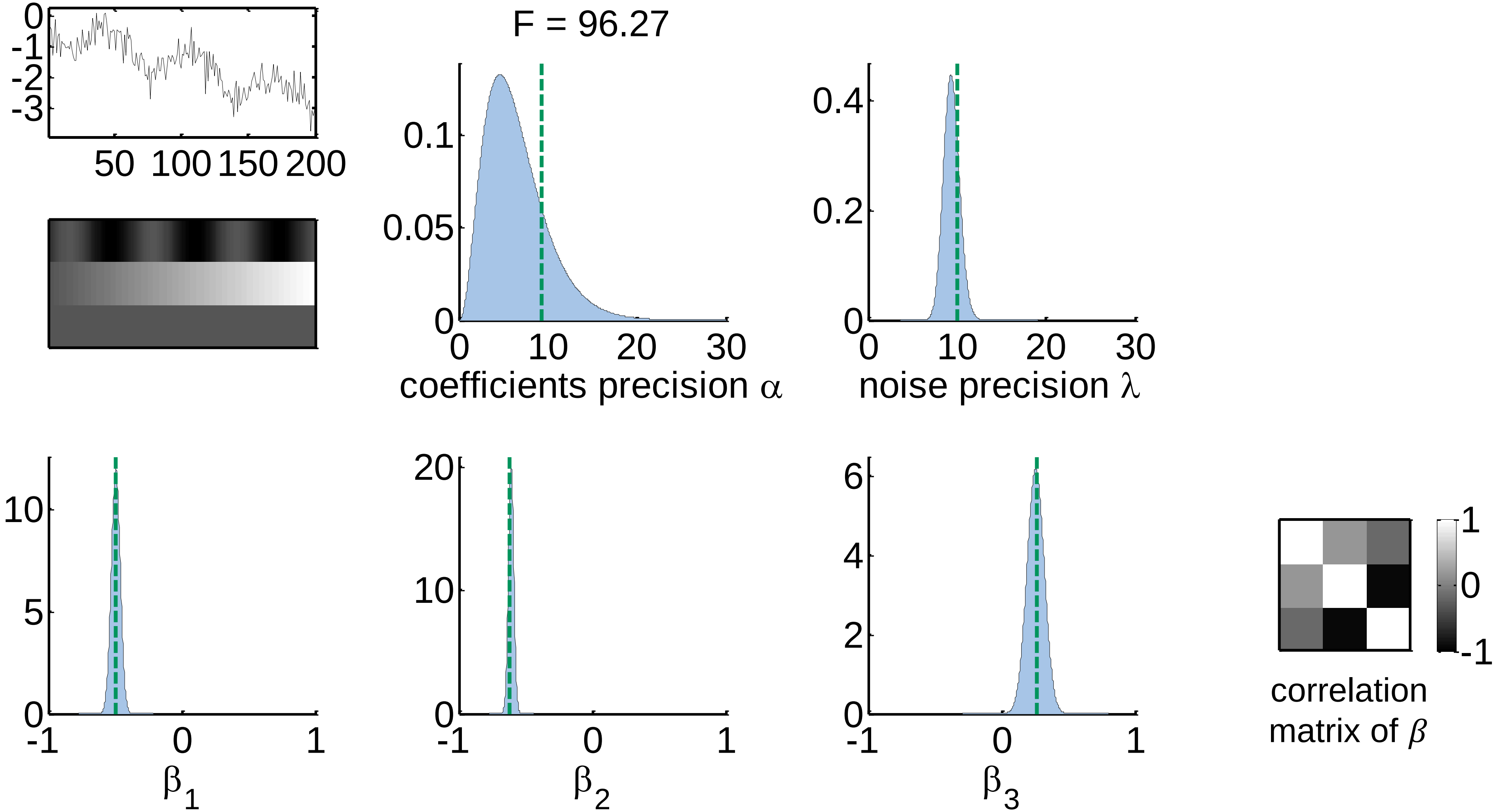
Variational linear regression: example

Iteration 1 **0:00:00’002**



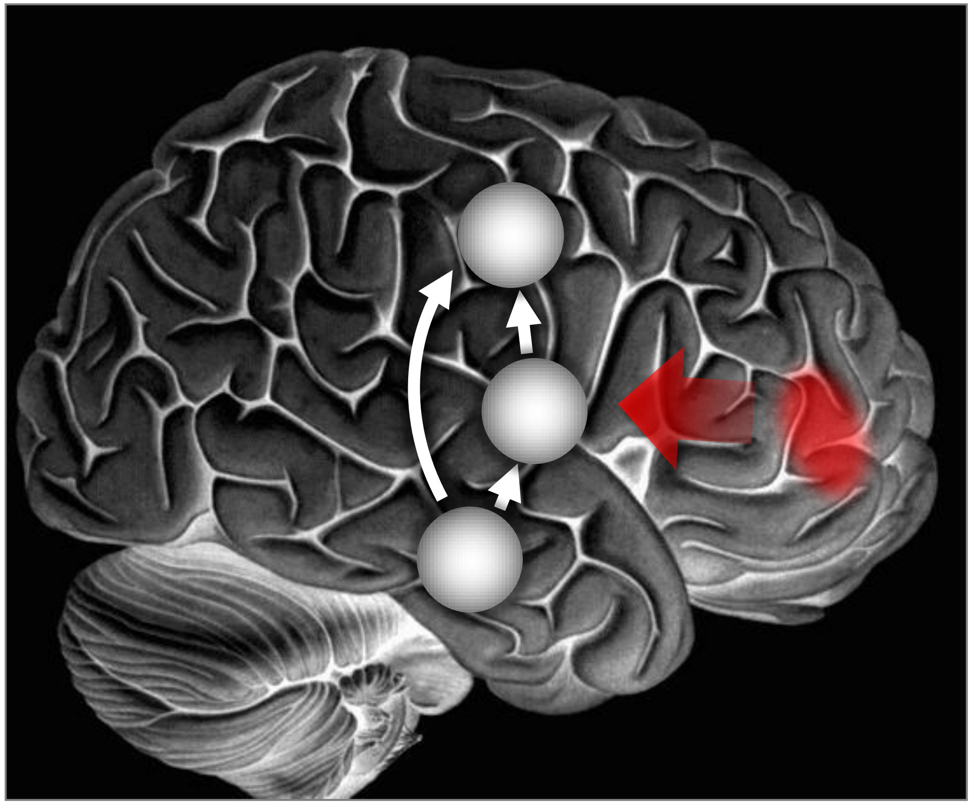
Variational linear regression: example

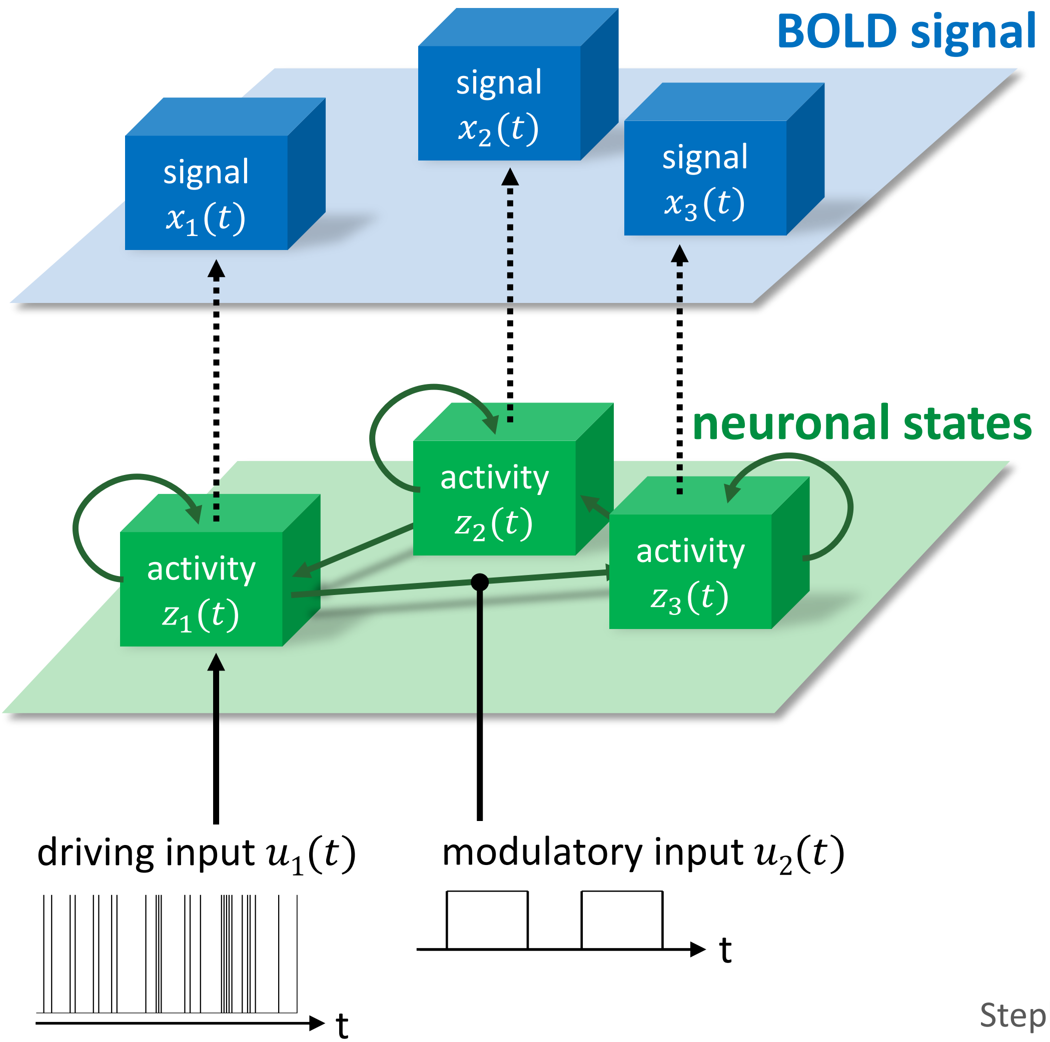
Iteration 2 (convergence) **0:00:00'003**



Overview

Example: diagnosing stroke patients





haemodynamic forward model

$$x = g(z, \theta_h)$$

neural state equation

$$\dot{z} = \left(A + \sum u_j B^{(j)}\right)z + Cu$$

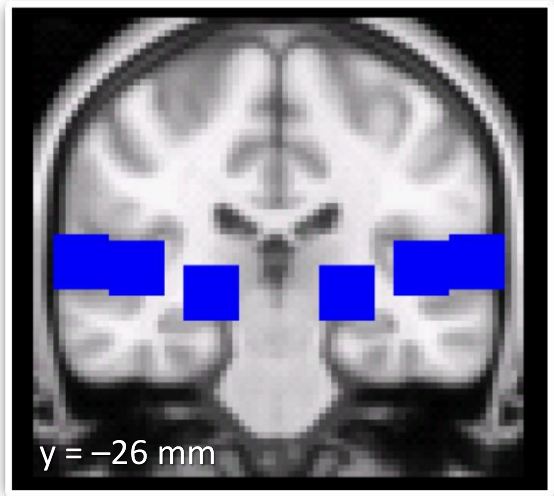
intrinsic connectivity

modulation of connectivity

direct inputs

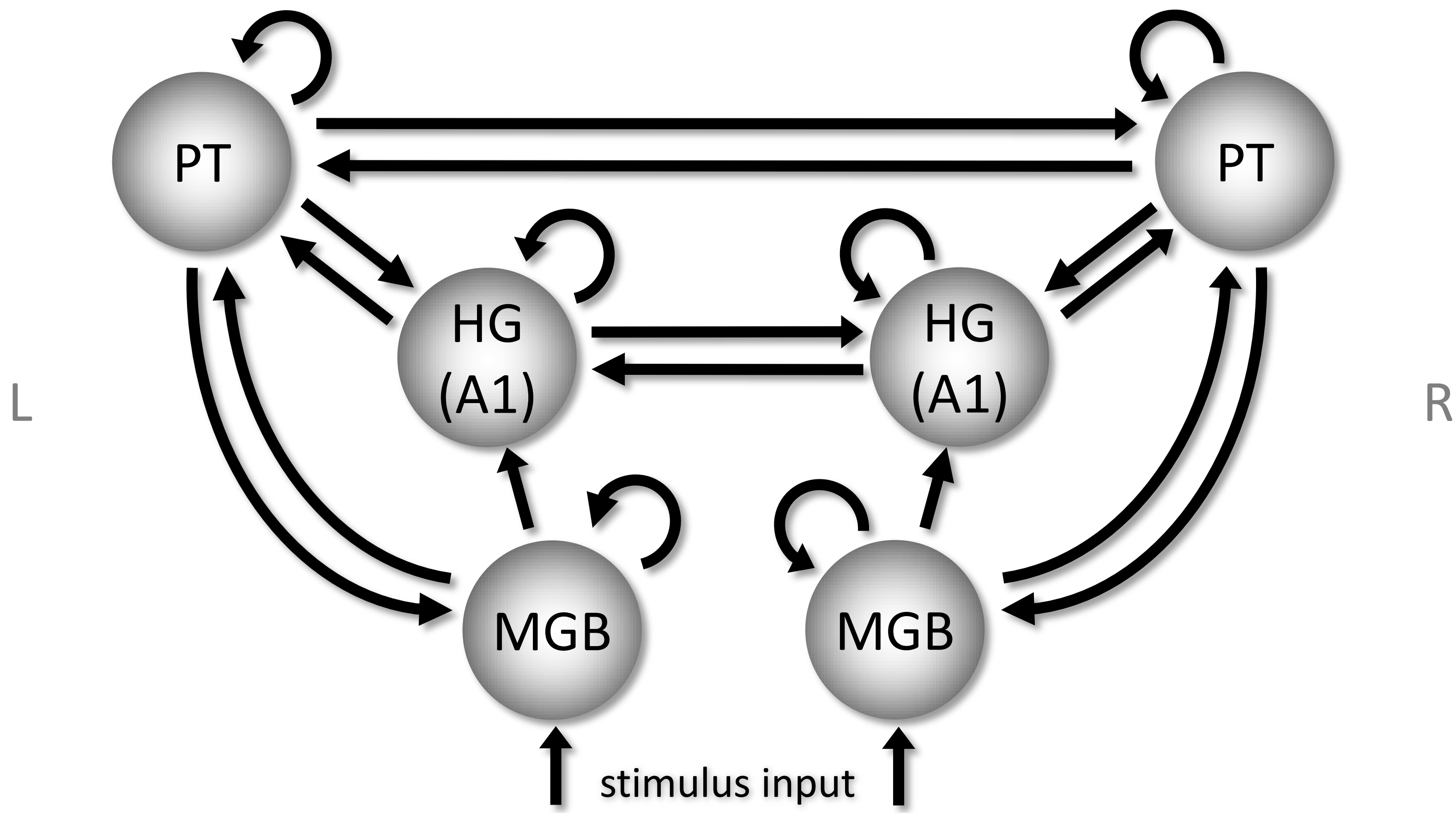
Friston, Harrison & Penny (2003) *NeuroImage*
Stephan & Friston (2007) *Handbook of Brain Connectivity*

Example: diagnosing stroke patients

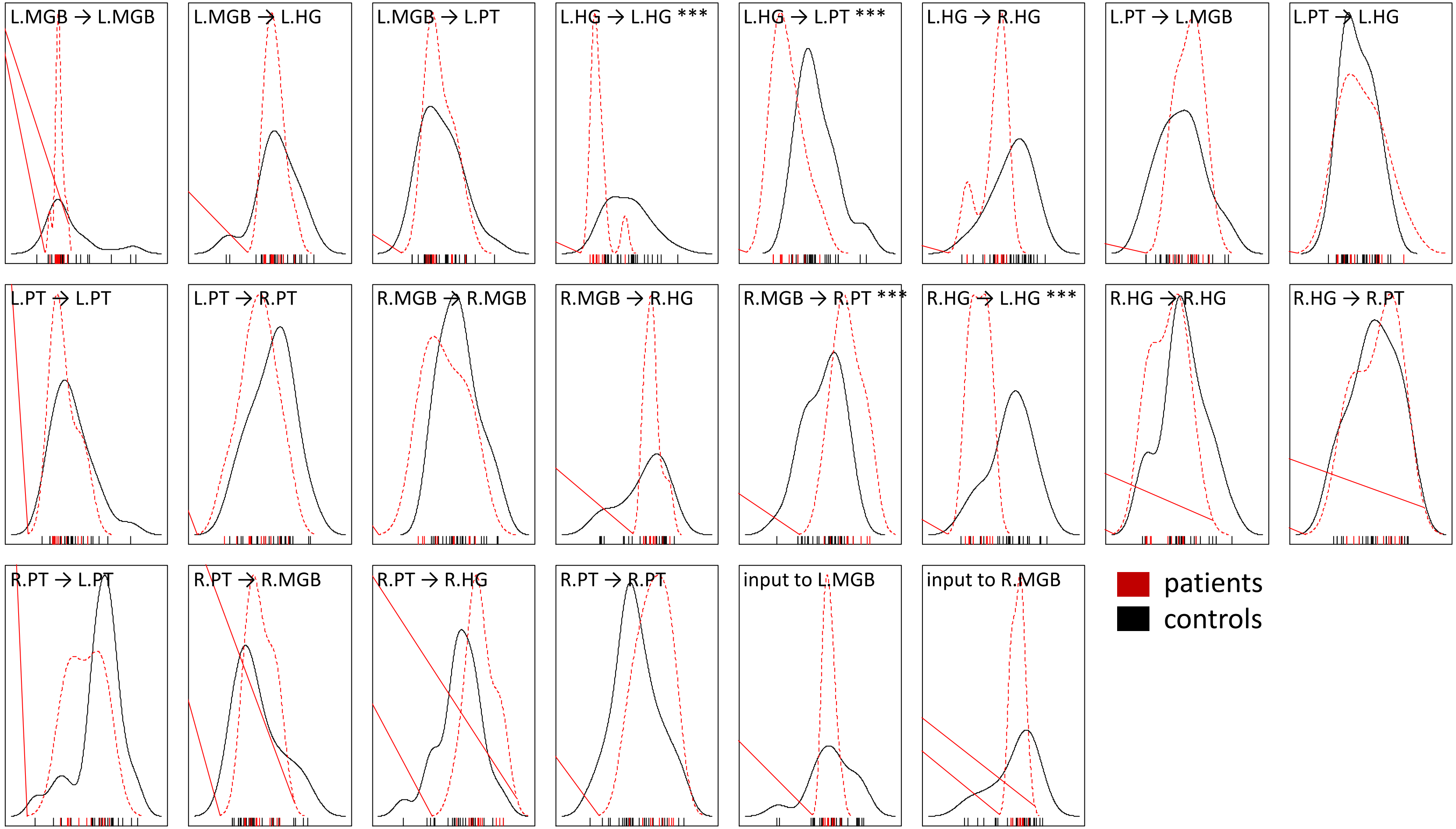


anatomical regions of interest

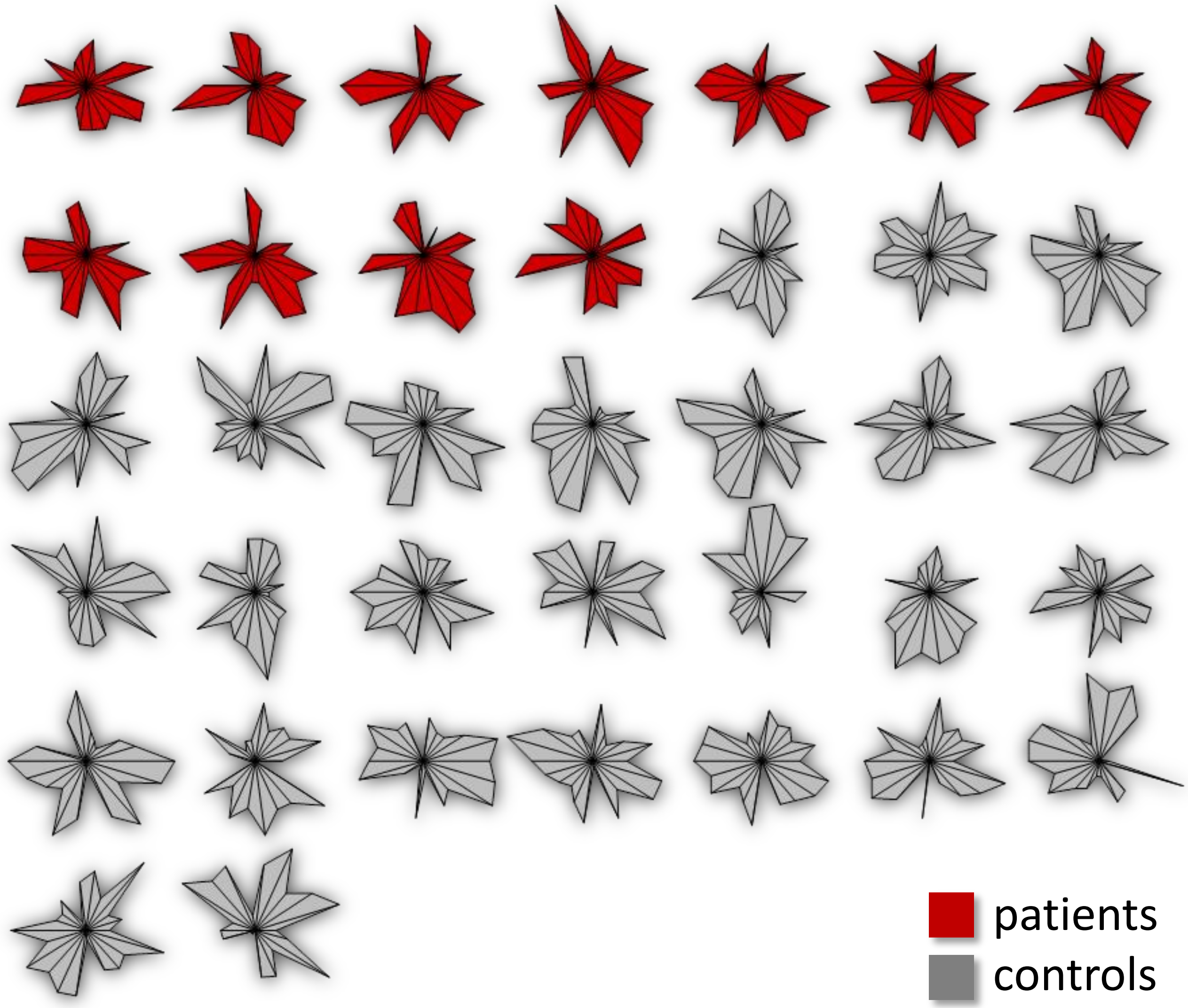
Example: diagnosing stroke patients



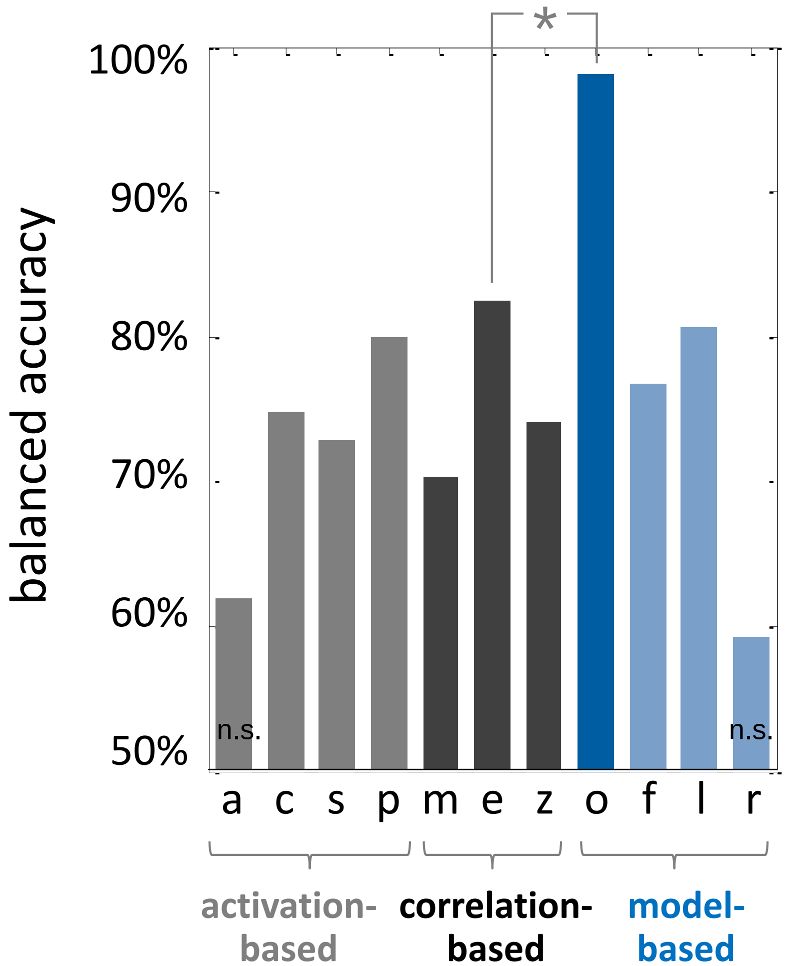
Univariate analysis: parameter densities



Multivariate analysis: connectional fingerprints



Classification performance



Activation-based analyses

a anatomical feature selection

c mass-univariate contrast feature selection

s locally univariate searchlight feature selection

p PCA-based dimensionality reduction

Correlation-based analyses

m correlations of regional means

e correlations of regional eigenvariates

z Fisher-transformed eigenvariates correlations

Model-based analyses

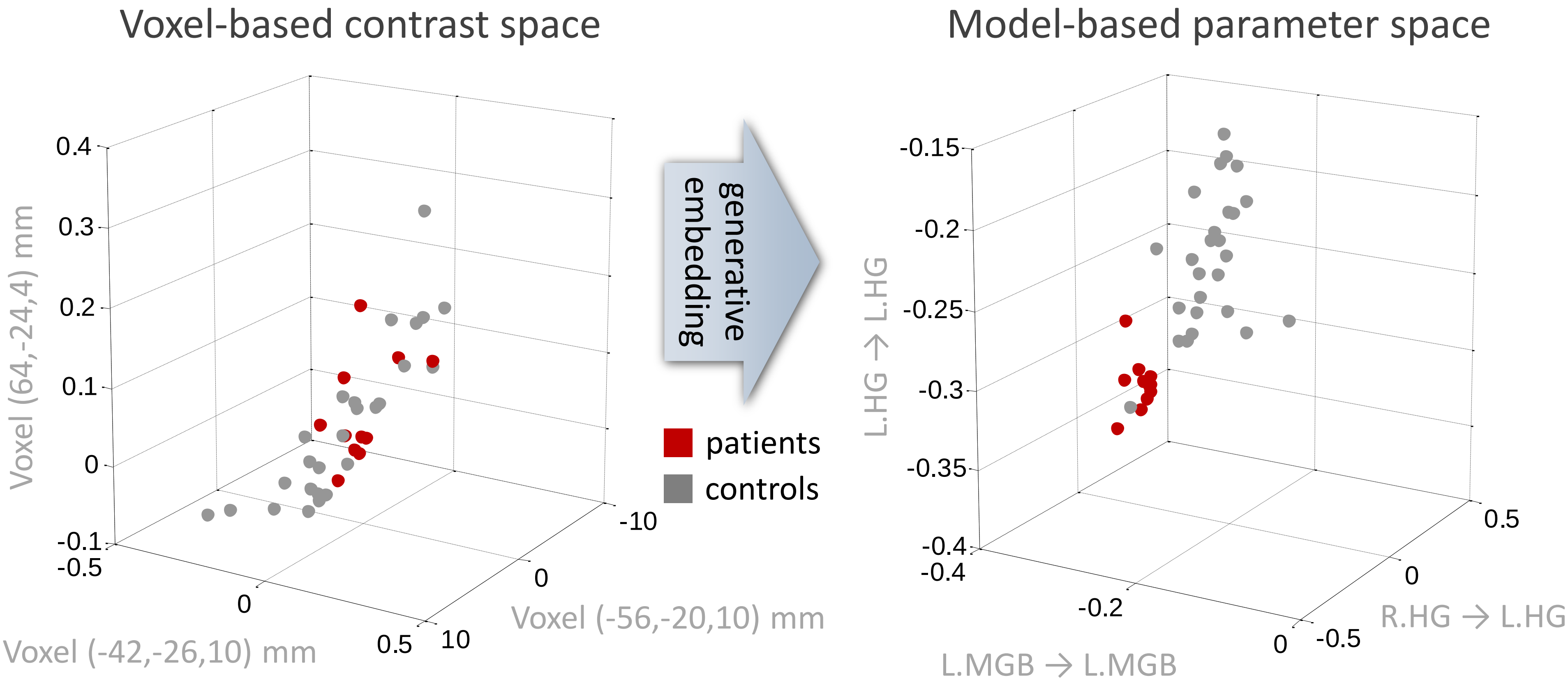
o gen.embed., original full model

f gen.embed., less plausible feedforward model

l gen.embed., left hemisphere only

r gen.embed., right hemisphere only

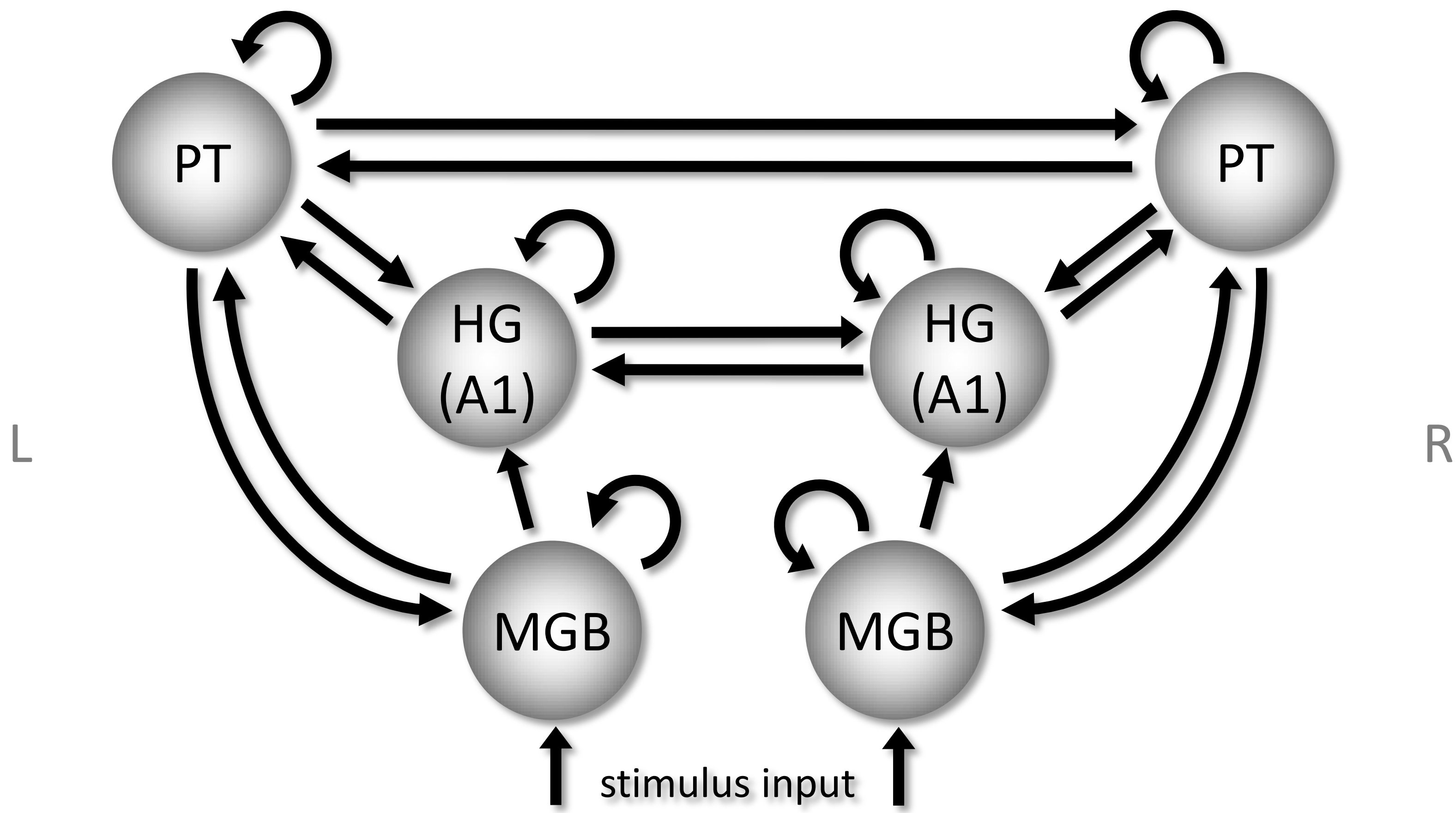
The generative projection



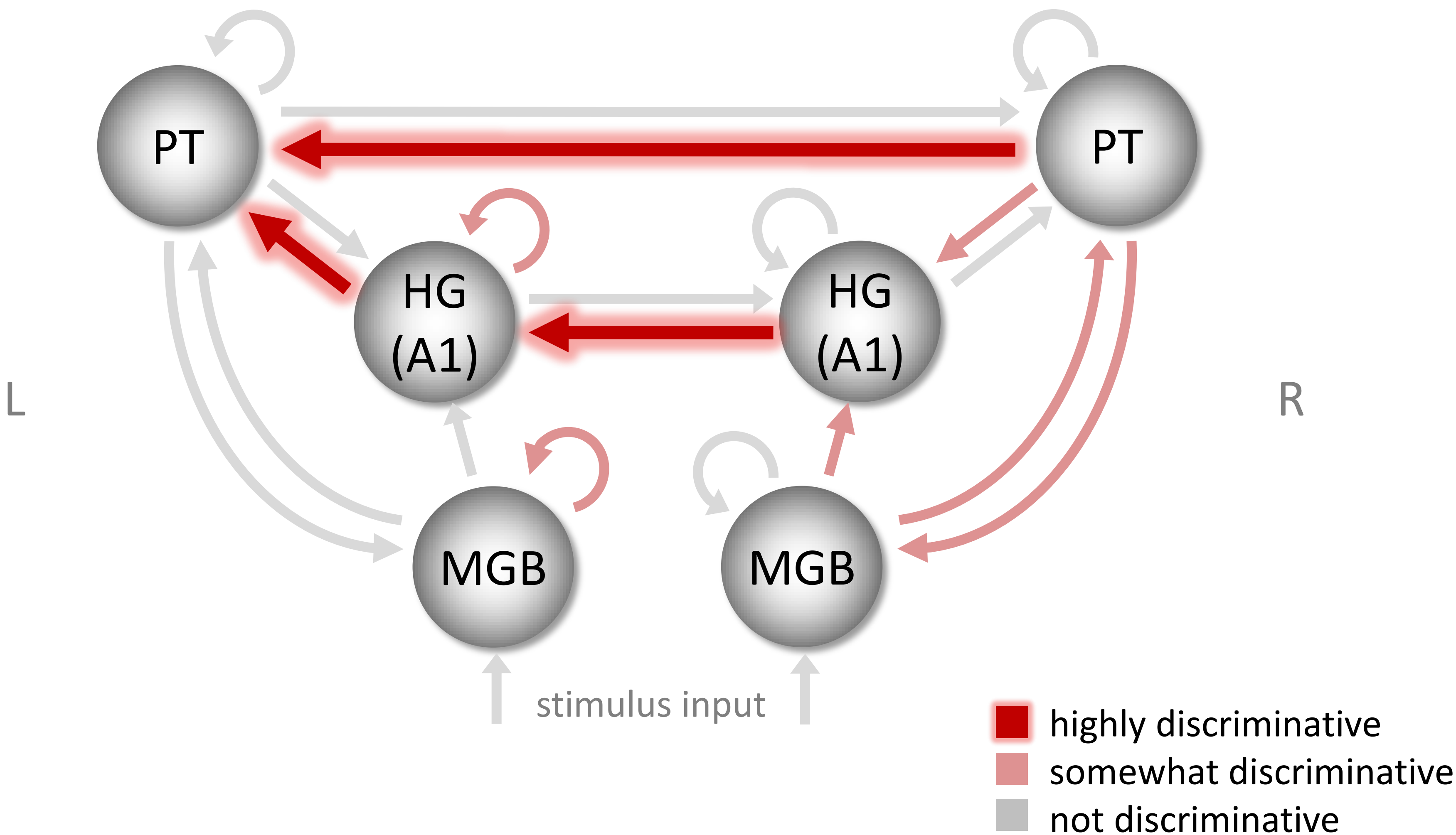
classification accuracy
(using all voxels in the regions of interest)
75%

classification accuracy
(using all 23 model parameters)
98%

Discriminative features in model space

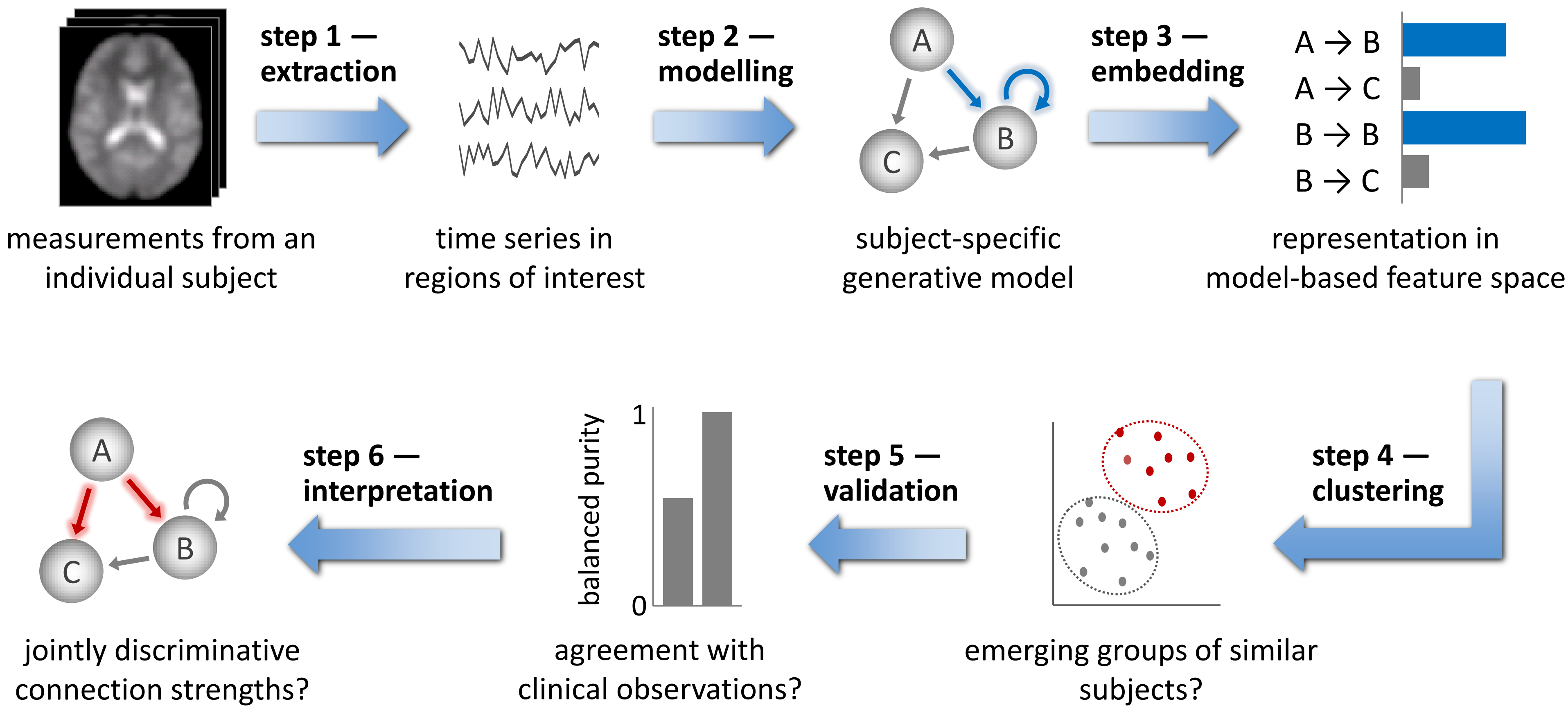


Discriminative features in model space



Overview

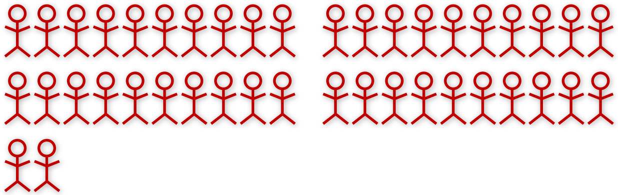
Generative embedding and clustering



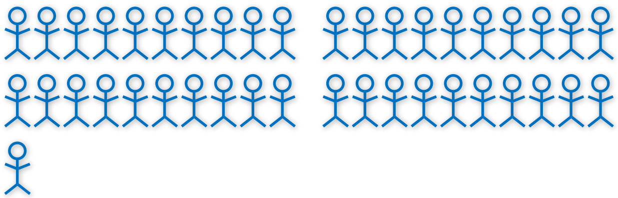
Brodersen et al. *NeuroImage* 2014

Dissecting schizophrenia into subtypes

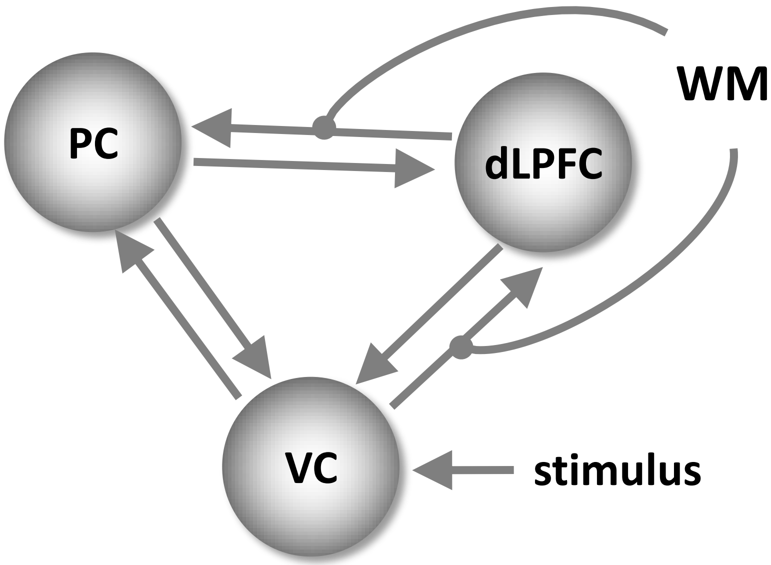
42 patients diagnosed with schizophrenia



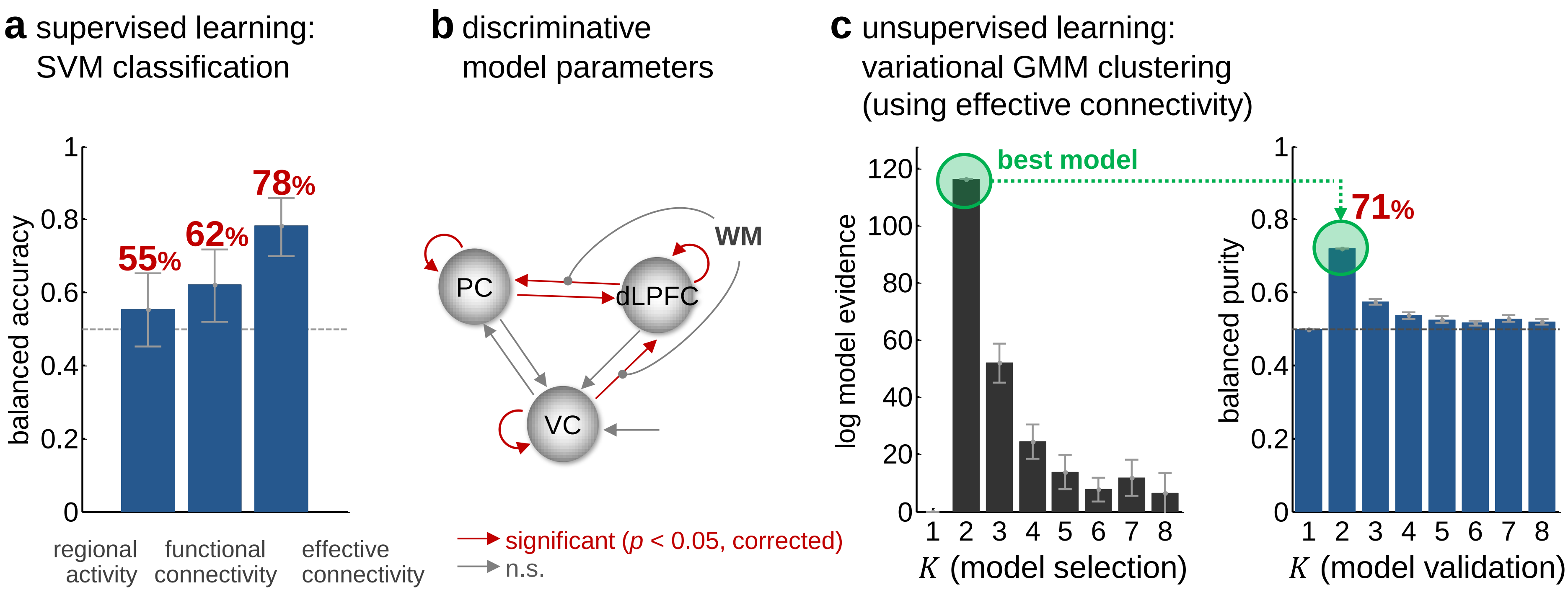
41 healthy controls



fMRI data acquired during working-memory task & modelled using a three-region DCM



Distinguishing between schizophrenia and healthy controls



Discovering new clinical subtypes