City Research Online

City, University of London Institutional Repository

Kullback-Leibler Divergence
for the Normal-Gamma Distribution

Joram Soch[1,3,•] & Carsten Allefeld[1,2]

[1] Bernstein Center for Computational Neuroscience, Berlin, Germany
[2] Berlin Center for Advanced Neuroimaging, Berlin, Germany
[3] Department of Psychology, Humboldt-Universität zu Berlin, Germany

• Corresponding author: joram.soch@bccn-berlin.de.

Abstract

We derive the Kullback-Leibler divergence for the normal-gamma distribution and show that it is identical to the Bayesian complexity penalty for the univariate general linear model with conjugate priors. Based on this finding, we provide two applications of the KL divergence, one in simulated and one in empirical data.

Contents

1   Introduction

Let x be a $k \times 1$ random vector and $y > 0$ be a random variable. Then, x and y are said to follow a *normal-gamma distribution* (NG distribution), if their joint probability density function is given by

$$p(x, y) = \mathrm{N}(x; \mu, (y\Lambda)^{-1}) \cdot \mathrm{Gam}(y; a, b) \tag{1}$$

where $\mathrm{N}(x; \mu, \Sigma)$ denotes a multivariate normal density with mean μ and covariance Σ and $\mathrm{Gam}(x; a, b)$ denotes a gamma density with shape a and rate b. In full, the density function is given by (Koch, 2007, p. 55)

$$p(x, y) = \sqrt{\frac{|y\Lambda|}{(2\pi)^k}} \, \exp\left[-\frac{y}{2}(x-\mu)^T \Lambda (x-\mu)\right] \cdot \frac{b^a}{\Gamma(a)} \, y^{a-1} \, \exp[-by] \ . \tag{2}$$

The *Kullback-Leibler divergence* (KL divergence) is a non-symmetric distance measure for two probability distributions P and Q and is defined as

$$\mathrm{KL}[P||Q] = \sum_{i \in \Omega} P(i) \ln \frac{P(i)}{Q(i)} \ . \tag{3}$$

For continuous probability distribtions P and Q with probability density functions $p(x)$ and $q(x)$ on the same domain X, it is given by (Bishop, 2006, p. 55)

$$\mathrm{KL}[P||Q] = \int_X p(x) \ln \frac{p(x)}{q(x)} \, \mathrm{d}x \ . \tag{4}$$

The KL divergence becomes important in information theory and statistical inference. Here, we derive the KL divergence for two NG distributions with vector-valued x and real-positive y and provide two examples of its application.

2 Theory

2.1 Multivariate normal KL divergence

First, consider two multivariate normal distributions over the $k \times 1$ vector x specified by

$$
\begin{aligned}
p(x) &= \mathrm{N}(x; \mu_1, \Sigma_1) \\
q(x) &= \mathrm{N}(x; \mu_2, \Sigma_2)
\end{aligned}
\tag{5}
$$

According to equation (4), the KL divergence of P from Q is defined as

$$
\mathrm{KL}[P||Q] = \int_{\mathbb{R}^k} \mathrm{N}(x; \mu_1, \Sigma_1) \ln \frac{\mathrm{N}(x; \mu_1, \Sigma_1)}{\mathrm{N}(x; \mu_2, \Sigma_2)} \, \mathrm{d}x \; .
\tag{6}
$$

Using the multivariate normal density function

$$
\mathrm{N}(x; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \, \exp\left[-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu) \right] \; ,
\tag{7}
$$

it evaluates to (Duchi, 2014)

$$
\mathrm{KL}[P||Q] = \frac{1}{2} \left[(\mu_2 - \mu_1)^T \Sigma_2^{-1} (\mu_2 - \mu_1) + \mathrm{tr}(\Sigma_2^{-1} \Sigma_1) - \ln \frac{|\Sigma_1|}{|\Sigma_2|} - k \right] \; .
\tag{8}
$$

2.2 Univariate gamma KL divergence

Next, consider two univariate gamma distributions over the real-positive y specified by

$$
\begin{aligned}
p(y) &= \mathrm{Gam}(y; a_1, b_1) \\
q(y) &= \mathrm{Gam}(y; a_2, b_2)
\end{aligned}
\tag{9}
$$

According to equation (4), the KL divergence of P from Q is defined as

$$
\mathrm{KL}[P||Q] = \int_0^\infty \mathrm{Gam}(y; a_1, b_1) \ln \frac{\mathrm{Gam}(y; a_1, b_1)}{\mathrm{Gam}(y; a_2, b_2)} \, \mathrm{d}y \; .
\tag{10}
$$

Using the univariate gamma density function

$$
\mathrm{Gam}(y; a, b) = \frac{b^a}{\Gamma(a)} \, y^{a-1} \, \exp[-by] \quad \text{for} \quad y > 0 \; ,
\tag{11}
$$

it evaluates to (Penny, 2001)

$$
\mathrm{KL}[P||Q] = a_2 \ln \frac{b_1}{b_2} - \ln \frac{\Gamma(a_1)}{\Gamma(a_2)} + (a_1 - a_2)\, \psi(a_1) - (b_1 - b_2)\frac{a_1}{b_1}
\tag{12}
$$

where $\psi(x)$ is the digamma function.

2.3 Normal-gamma KL divergence

Now, consider two normal-gamma distributions over x and y specified by

$$p(x, y) = \mathrm{N}(x; \mu_1, (y\Lambda_1)^{-1})\, \mathrm{Gam}(y; a_1, b_1)$$
$$q(x, y) = \mathrm{N}(x; \mu_2, (y\Lambda_2)^{-1})\, \mathrm{Gam}(y; a_2, b_2)$$

(13)

According to equation (4), the KL divergence of P from Q is defined as

$$\mathrm{KL}[P||Q] = \int_0^\infty \int_{\mathbb{R}^k} p(x, y) \ln \frac{p(x, y)}{q(x, y)}\, \mathrm{d}x\, \mathrm{d}y \;.$$

(14)

Using the law of conditional probability, it can be evaluated as follows:

$$
\begin{aligned}
\mathrm{KL}[P||Q] &= \int_0^\infty \int_{\mathbb{R}^k} p(x|y)p(y) \ln \frac{p(x|y)p(y)}{q(x|y)q(y)}\, \mathrm{d}x\, \mathrm{d}y \\
&= \int_0^\infty p(y) \int_{\mathbb{R}^k} p(x|y) \ln \frac{p(x|y)}{q(x|y)}\, \mathrm{d}x\, \mathrm{d}y \\
&\quad + \int_0^\infty p(y) \ln \frac{p(y)}{q(y)} \int_{\mathbb{R}^k} p(x|y)\, \mathrm{d}x\, \mathrm{d}y \\
&= \langle \mathrm{KL}[p(x|y)||q(x|y)] \rangle_{p(y)} + \mathrm{KL}[p(y)||q(y)]
\end{aligned}
$$

(15)

In other words, the KL divergence for two normal-gamma distributions over x and y is equal to the sum of a multivariate normal KL divergence regarding x conditional on y, expected over y, and a univariate gamma KL divergence regarding y. Together with equation (8), the first term becomes

$$
\begin{aligned}
&\langle \mathrm{KL}[p(x|y)||q(x|y)] \rangle_{p(y)} \\
&= \left\langle \frac{1}{2} \left[(\mu_2 - \mu_1)^T (y\Lambda_2)(\mu_2 - \mu_1) + \mathrm{tr}\left((y\Lambda_2)(y\Lambda_1)^{-1}\right) - \ln \frac{|(y\Lambda_1)^{-1}|}{|(y\Lambda_2)^{-1}|} - k \right] \right\rangle_{p(y)} \\
&= \left\langle \frac{y}{2} (\mu_2 - \mu_1)^T \Lambda_2 (\mu_2 - \mu_1) + \frac{1}{2} \mathrm{tr}(\Lambda_2 \Lambda_1^{-1}) - \frac{1}{2} \ln \frac{|\Lambda_2|}{|\Lambda_1|} - \frac{k}{2} \right\rangle_{p(y)} .
\end{aligned}
$$

(16)

Using the relation $y \sim \mathrm{Gam}(a, b) \Rightarrow \langle y \rangle = a/b$, we have

$$\langle \mathrm{KL}[p(x|y)||q(x|y)] \rangle_{p(y)} = \frac{1}{2} \frac{a_1}{b_1} (\mu_2 - \mu_1)^T \Lambda_2 (\mu_2 - \mu_1) + \frac{1}{2} \mathrm{tr}(\Lambda_2 \Lambda_1^{-1}) - \frac{1}{2} \ln \frac{|\Lambda_2|}{|\Lambda_1|} - \frac{k}{2} \;.$$

(17)

Thus, from (17) and (12), the KL divergence in (15) becomes

$$
\begin{aligned}
\mathrm{KL}[P||Q] &= \frac{1}{2} \frac{a_1}{b_1} \left[(\mu_2 - \mu_1)^T \Lambda_2 (\mu_2 - \mu_1) \right] + \frac{1}{2} \mathrm{tr}(\Lambda_2 \Lambda_1^{-1}) - \frac{1}{2} \ln \frac{|\Lambda_2|}{|\Lambda_1|} - \frac{k}{2} \\
&\quad + a_2 \ln \frac{b_1}{b_2} - \ln \frac{\Gamma(a_1)}{\Gamma(a_2)} + (a_1 - a_2)\, \psi(a_1) - (b_1 - b_2) \frac{a_1}{b_1} \;.
\end{aligned}
$$

(18)

2.4 The Bayesian model evidence

Consider Bayesian inference on data y using model m with parameters θ. In this case, Bayes' theorem is a statement about the posterior density:

$$p(\theta|y, m) = \frac{p(y|\theta, m)\, p(\theta|m)}{p(y|m)} \ . \tag{19}$$

The denominator $p(y|m)$ acts as a normalization constant on the posterior density $p(\theta|y, m)$ and according to the law of marginal probability is given by

$$p(y|m) = \int p(y|\theta, m)\, p(\theta|m)\, \mathrm{d}\theta \ . \tag{20}$$

This is the probability of the data given only the model, regardless of any particular parameter values. It is also called "marginal likelihood" or "model evidence" and can act as a model quality criterion in Bayesian inference, because parameters are integrated out of the likelihood.

For computational reasons, only the logarithmized or log model evidence (LME) $L(m) = \ln p(y|m)$ is of interest in most cases. By rearranging equation (19), the model evidence can be represented as

$$p(y|m) = \frac{p(y|\theta, m)\, p(\theta|m)}{p(\theta|y, m)} \ . \tag{21}$$

Logarithmizing both sides of the equation and taking the expectation with respect to the posterior density over model parameters θ gives the LME

$$L(m) = \int p(\theta|y, m) \ln p(y|\theta, m)\, \mathrm{d}\theta - \int p(\theta|y, m) \ln \frac{p(\theta|y, m)}{p(\theta|m)}\, \mathrm{d}\theta \ . \tag{22}$$

Using this reformulation, the LME as a model quality measure can be naturally decomposed into an accuracy term, the posterior expected likelihood, and a complexity term, the KL divergence between the posterior and the prior distribution:

$$\begin{aligned} L(m) &= \mathrm{Acc}(m) - \mathrm{Com}(m) \\ \mathrm{Acc}(m) &= \langle \log p(y|\theta, m) \rangle_{p(\theta|y, m)} \\ \mathrm{Com}(m) &= \mathrm{KL}\left[p(\theta|y, m) || p(\theta|m) \right] \end{aligned} \tag{23}$$

Intuitively, the accuracy acts increasing and the complexity acts decreasing on the log model evidence. This reflects the capability of the LME to select models that achieve the best balance between accuracy and complexity, i.e. models that explain the observations sufficiently well (high accuracy) without employing too many principles (low complexity). The fact that the complexity term is a KL divergence between posterior and prior means that models with prior assumptions that are close to the posterior evidence receive a low complexity penalty, because one is not surprised very much when accepting such a model which renders the Bayesian complexity a measure of surprise.

2.5   The general linear model

Consider multiple linear regression using the univariate general linear model (GLM)

$$y = X\beta + \varepsilon, \ \varepsilon \sim N(0, \sigma^2 V) \tag{24}$$

where y is an $n \times 1$ vector of measured data, X is an $n \times p$ matrix called the design matrix, β is a $p \times 1$ vector of weight parameters called regression coefficients and ε is an $n \times 1$ vector of errors or noise. These residuals are assumed to follow a multivariate normal distribution whose covariance matrix is the product of a variance factor σ^2 and an $n \times n$ correlation matrix V. Usually, X and V are known while β and τ are unknown parameters to be inferred via model estimation.
For mathematical convenience, we rewrite $\sigma^2 = 1/\tau$ and $V = P^{-1}$ so that equation (24) implies the following likelihood function:

$$p(y|\beta, \tau) = \mathrm{N}(y; X\beta, (\tau P)^{-1}) . \tag{25}$$

The conjugate prior relative to this likelihood function is a normal-gamma distribution on the model parameters β and τ (Koch, 2007, ch. 2.6.3):

$$p(\beta|\tau) = \mathrm{N}(\beta; \mu_0, (\tau \Lambda_0)^{-1})$$
$$p(\tau) = \mathrm{Gam}(\tau; a_0, b_0) \tag{26}$$

Due to the conjugacy of (26) to (25), the posterior is also a normal-gamma distribution

$$p(\beta|\tau, y) = \mathrm{N}(\beta; \mu_n, (\tau \Lambda_n)^{-1})$$
$$p(\tau|y) = \mathrm{Gam}(\tau; a_n, b_n) \tag{27}$$

where the posterior parameters in (27) are given by (Koch, 2007, ch. 4.3.2)

$$\begin{aligned}
\mu_n &= \Lambda_n^{-1}(X^T P y + \Lambda_0 \mu_0) \\
\Lambda_n &= X^T P X + \Lambda_0 \\
a_n &= a_0 + \frac{n}{2} \\
b_n &= b_0 + \frac{1}{2}(y^T P y + \mu_0^T \Lambda_0 \mu_0 - \mu_n^T \Lambda_n \mu_n)
\end{aligned} \tag{28}$$

From (23), the complexity for the model defined by (25) and (26) is given by

$$\mathrm{Com}(m) = \mathrm{KL}\left[p(\beta, \tau|y)||p(\beta, \tau)\right] . \tag{29}$$

In other words, the complexity penalty for a general linear model with normal-gamma priors (GLM-NG) is identical to a KL divergence between two NG distributions and using (18) can be written in terms of the prior and posterior parameters as

$$\begin{aligned}
\mathrm{Com}(m) = {}&\frac{1}{2}\frac{a_n}{b_n}\left[(\mu_n - \mu_0)^T \Lambda_0 (\mu_n - \mu_0)\right] + \frac{1}{2}\mathrm{tr}(\Lambda_0 \Lambda_n^{-1}) - \frac{1}{2}\ln\frac{|\Lambda_0|}{|\Lambda_n|} - \frac{p}{2} \\
&+ a_0 \ln\frac{b_n}{b_0} - \ln\frac{\Gamma(a_n)}{\Gamma(a_0)} + (a_n - a_0)\,\psi(a_n) - (b_n - b_0)\frac{a_n}{b_n} .
\end{aligned} \tag{30}$$

3 Application

3.1 Polynomial basis functions

Consider a linear model with polynomial basis functions (Bishop, 2006, p. 5) given by

$$y = \sum_{i=0}^{p} c_i \, x^i \quad \text{with} \quad -1 \leq x \leq +1 \, . \tag{31}$$

Essentially, this model assumes that y is an additive mixture of polynomial terms x^i weighted with the coefficients c_i with $i = 1, \ldots, p$ where the natural number p is called the model order. This means that $p = 0$ corresponds to a constant value (plus noise ε); $p = 1$ corresponds to a linear function; $p = 2$ corresponds to a quadratic pattern; $p = 3$ corresponds to a 3rd degree polynomial etc.

Given that x is an $n \times 1$ vector of real numbers between -1 and $+1$, this model can be rewritten as a GLM given in equation (24) with

$$X = \begin{bmatrix} x_1^0 & x_1^1 & \cdots & x_1^p \\ x_2^0 & x_2^1 & \cdots & x_2^p \\ \vdots & \vdots & \ddots & \vdots \\ x_n^0 & x_n^1 & \cdots & x_n^p \end{bmatrix} \quad \text{and} \quad \beta = \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_p \end{bmatrix} \, . \tag{32}$$

Based on this reformulation, we simulate polynomial data. We perform $N = 100$ simulations with $n = 100$ data points in each simulation. We generate simulated data based on a true model order $p_{\text{true}} = 5$ and analyze these data using a set of models ranging from $p_{\text{min}} = 0$ to $p_{\text{max}} = 20$. The predictor x is equally spaced between -1 and $+1$ and design matrices X_p are created according to equation (32).

In each simulation, six regression coefficients β_{true} are drawn independently from the standard normal distribution $N(0, 1)$. Then, Gaussian noise ε is sampled from the multivariate normal distribution $N(0, \sigma_\varepsilon^2 I_n)$ with a residual variance of $\sigma_\varepsilon^2 = 1$. Finally, simulated data is generated as as $y = X_5 \beta_{\text{true}} + \varepsilon$.

Then, for each $p \in \{0, \ldots, 20\}$, Bayesian model estimation is performed using the design matrix X_p, a correlation matrix $V = I_n$ and the prior distributions (26) with the prior parameters $\mu_0 = 0_p$, $\Lambda_0 = I_p$ invoking a standard multivariate normal distribution and $a_0 = 1$, $b_0 = 1$ invoking a relatively flat gamma prior. Posterior parameters are calculated using equation (28) and give rise to the model complexity via (30) as well as model accuracy and the log model evidence via (23).

Average LMEs, accuracies and complexities are shown in Figure 1. One can see that the true model order is correctly identified by the maximal log model evidence. This is achieved by an increasing complexity penalty which outweighs the saturating accuracy gain for models with $p > 5$. This demonstrates that the KL divergence for the NG distribution can be used to select polynomial basis functions when basis sets cannot be separated based on model accuracy alone.
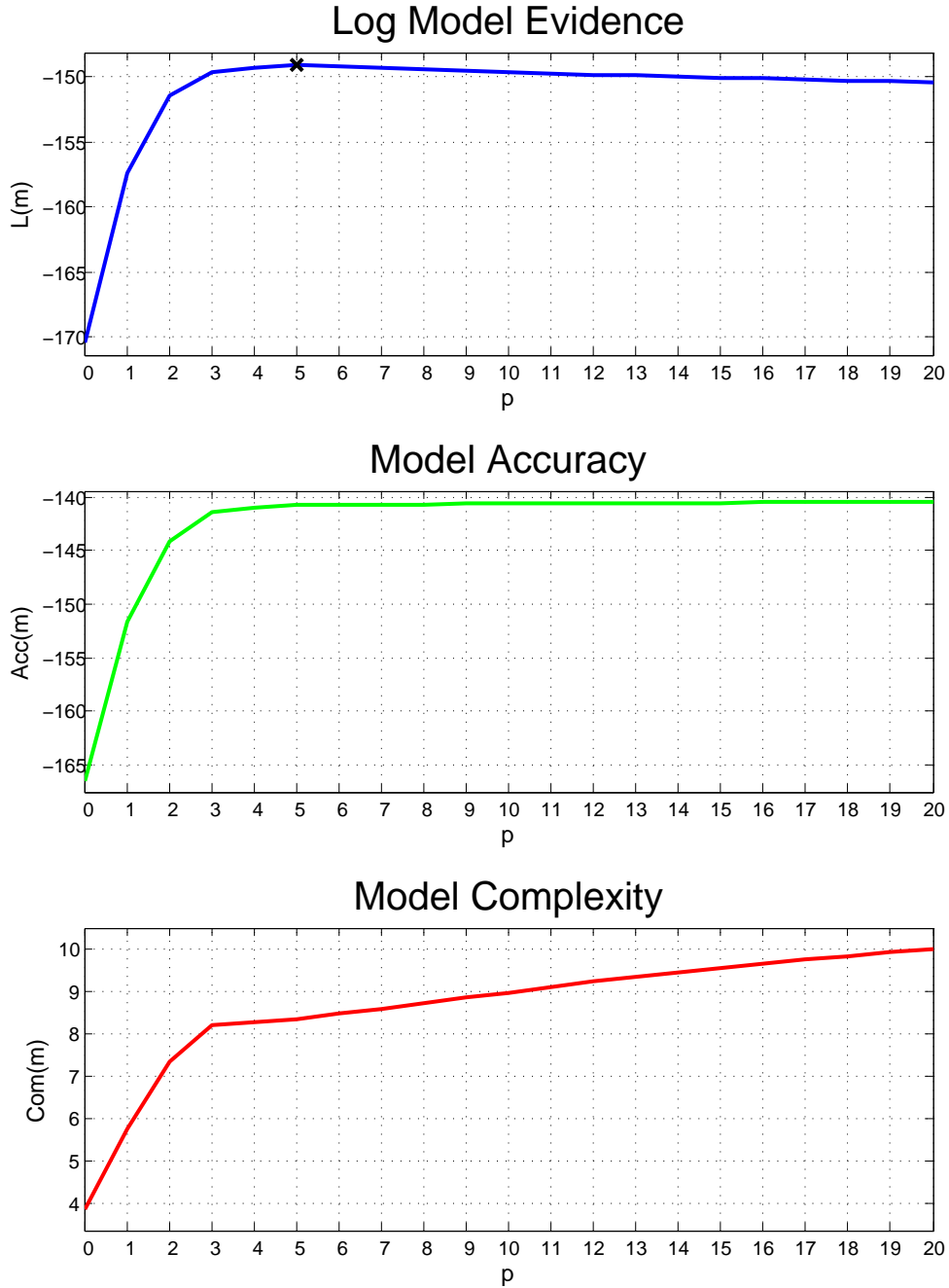
Figure 1. Bayesian model selection for polynomial basis functions. All displays have model order on the x-axis and average model quality measures (across simulations) on the y-axis. Intuitively, the model accuracy (middle panel) increases with model order, but saturates at around -140 with no major increase after $p = 5$. Moreover, the model complexity (lower panel) – which is the KL divergence between prior and posterior distribution – also grows with model order, but switches to a linear increase at around $p = 5$ reaching a value of 10 at $p = 20$. Together, this has the consequence that the log model evidence (upper panel) is maximal for $p = 5$ (black cross) where exact values are: $\mathrm{Acc}(m) = -140.77$, $\mathrm{Com}(m) = 8.34$, $\mathrm{L}(m) = -149.11$.

3.2 Neuroimaging model selection

In neuroimaging, especially functional magnetic resonance imaging (fMRI), GLMs as given by equation (24) are applied to time series of neural data y (Friston et al., 1995). The design matrix X is specified by the temporal occurrence of experimental conditions and the covariance structure V is estimated from residual auto-correlations. Model estimation and statistical inference are performed "voxel-wise", i.e. separately for each measurement location in the brain, usually referred to as the "mass-univariate GLM".

Here, we analyze data from a study on orientation pop-out processing (Bogler et al., 2013). During the experimental paradigm, the screen showed a 3×7 array of homogeneous bars oriented either 0°, 45°, 90° or 135° relative to the vertical axis. This background stimulation changed every second and was interrupted by trials in which one target bar on the left and one target bar on the right were independently rotated either 0°, 30°, 60° or 90° relative to the rest of the stimulus display. Those trials of orientation contrast (OC) lasted 4 seconds and were alternated with inter-trial intervals of 7, 10 or 13 seconds. Each combination of OC on the left side and OC on the right side was presented three times resulting in 48 trials in each of the 5 sessions lasting 672 seconds.

After fMRI data preprocessing (slice-timing, realignment, normalization, smoothing), two different models of hemodynamic activation were applied to the fMRI data. The first model (GLM I) considers the experiment a factorial design with two factors (left vs. right OC) having four levels (0°, 30°, 60°, 90°). This results in $4 \times 4 = 16$ possible combinations or experimental conditions modelled by 16 onset regressors convolved with the canonical hemodynamic response function (HRF). The second model (GLM II) puts all trials from all conditions into one HRF-convolved regressor and encodes orientation contrast using a parametric modulator (PM) that is given as PM = deg/90° with deg = (0°, 30°, 60°, 90°), resulting in PM = $(0, \frac{1}{3}, \frac{2}{3}, 1)$, such that the parametric modulator is proportional to orientation contrast. There was one PM for each factor of the design, i.e. one PM for left OC and one PM for right OC. Note that both models encode the same information and that every signal that can be identified using GLM II can also be characterized using GLM I, but not vice versa, because the first model allows for a greater flexibility of activation patterns across experimental conditions than the second.

For these two models, we performed Bayesian model estimation. To overcome the challenge of having to specify prior distributions on the model parameters, we apply cross-validation across fMRI sessions. This gives rise to a cross-validated log model evidence (cvLME) as well as cross-validated accuracies and complexities for each model in each subject. We then performed a paired t-test to find voxels where GLM II has a significantly higher cvLME than GLM I. Due to the specific assumptions in GLM II and the higher flexibility of GLM I, we assumed that these differences might be primarily based on a complexity advantage of GLM II over GLM I.

We focus on visual area 4 (V4) that is known to be sensitive to orientation contrast. Within left V4, specified by a mask from a separate localizer paradigm (Bogler et al., 2013), we identified the peak voxel ([x y z] = $[-15, -73, -5]$ mm) defined by the maximal t-value ($t = 3.95$) and extracted log model evidence as well as model accuracy and model complexity from this voxel for each subject. Differences in LME, accuracy and complexity are shown in Figure 2. Again, the model complexity enables a model selection that would not be possible based on the model accuracy alone.
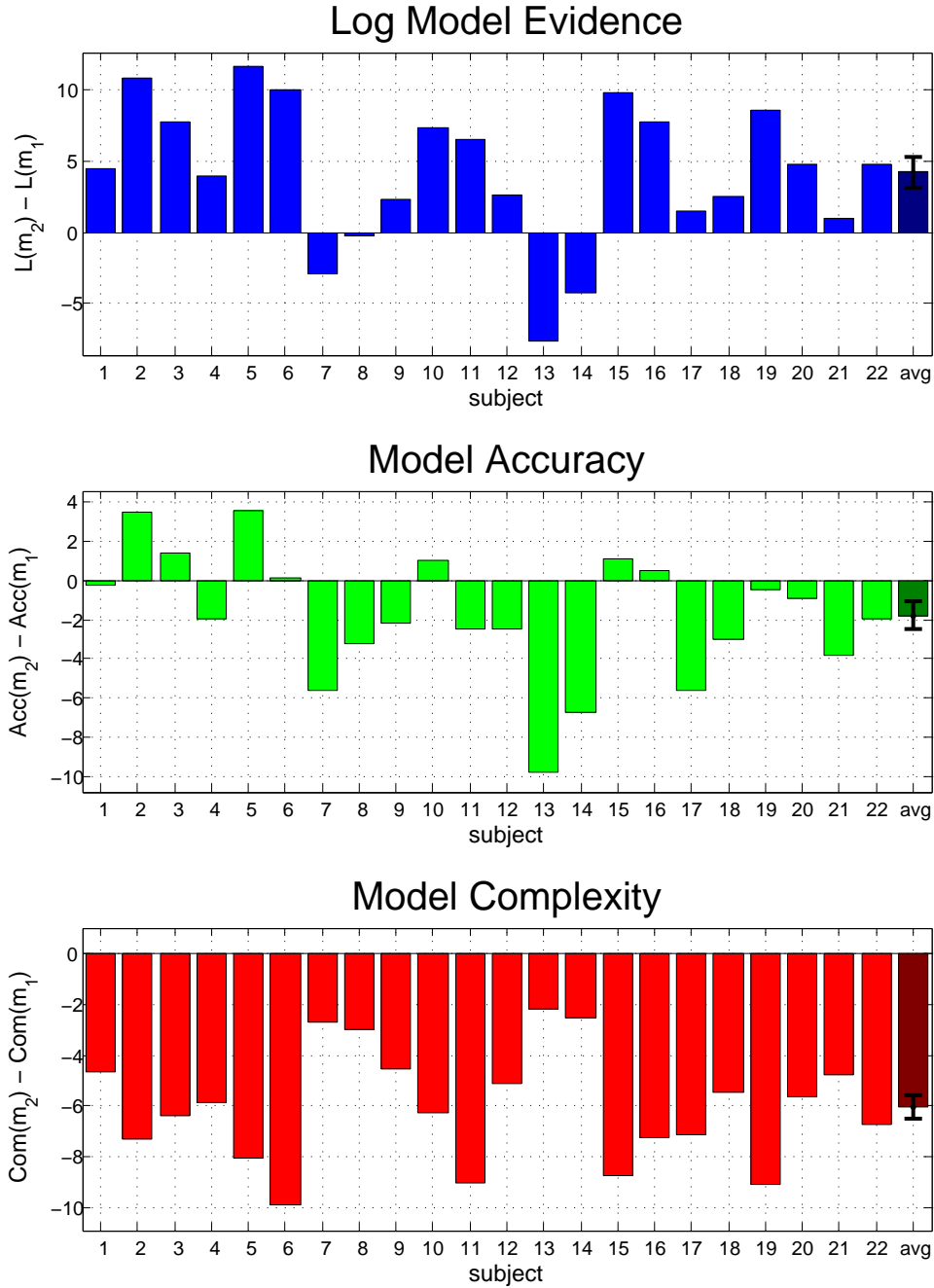
Figure 2. Bayesian model selection for orientation pop-out processing. All displays have subject on the x-axis and difference in model qualities (m_1 = GLM I, m_2 = GLM II) on the y-axis. Interestingly, there is a slight disadvantage for GLM II regarding only the model accuracy (middle panel), its mean difference across subjects being smaller than zero. However, model complexity (lower panel) – measured as the KL divergence between prior and posterior distribution – is consistently higher for GLM I. Together, this has the consequence that the log model evidence (upper panel) most often favors GLM II. Average values are: $\Delta \text{Acc} = -1.78$, $\Delta \text{Com} = -6.02$, $\Delta L = 4.24$.

4 Conclusion

We have derived the Kullback-Leibler divergence of two normal-gamma distributions using earlier results on the KL divergence for multivariate normal and univariate gamma distributions. Moreover, we have shown that the KL divergence for the NG distribution occurs as the complexity term in the univariate general linear model when using conjugate priors.

Analysis of simulated and empirical data demonstrates that the complexity penalty has the desired theoretical features, namely to quantify the relative informational content of two generative models and to detect model differences that cannot be detected by just relying on model accuracy, e.g. given by the maximum log-likelihood (as in information criteria like AIC or BIC) or the posterior log-likelihood (as in the Bayesian log model evidence).

5 References

[1] Bishop CM (2006): *Pattern Recognition and Machine Learning.* 1st Edition, Springer, New York.

[2] Bogler C, Bode S, Haynes JD (2013): "Orientation pop-out processing in human visual cortex". *NeuroImage*, vol. 81, pp. 73-80.

[3] Duchi J (2014): "Derivations for Linear Algebra and Optimization". University of California, Berkeley; URL: http://www.eecs.berkeley.edu/~jduchi/projects/general_notes.pdf.

[4] Friston KJ, Holmes AP, Worsley KJ, Poline JP, Frith CD, Frackowiak RSJ (1995): "Statistical parametric maps in functional imaging: A general linear approach". *Human Brain Mapping*, vol. 2, iss. 4, pp. 189-210.

[5] Koch KR (2007): *Introduction to Bayesian Statistics.* 2nd Edition, Springer, Berlin/ Heidelberg, 2007.

[6] Penny WD (2001): "KL-Divergences of Normal, Gamma, Dirichlet and Wishart densities". University College, London; URL: www.fil.ion.ucl.ac.uk/~wpenny/ publications/densities.ps.