The Gaussian Onion: Priors for a Spherical World

or

r-priors: Towards unbiased model comparison in Gaussian linear regression problems

Hans Eggers, Michiel de Kock

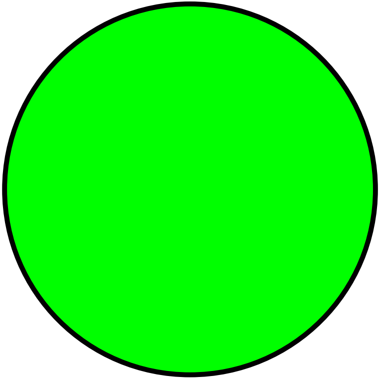
Stellenbosch University and Excellence Cluster "Universe"

Bayes Forum

27 November 2015

Overview

- **Why?**
  - Niche problem
  - Very general application
  - Start with simplest case
  - Try to understand everything
- **Structure of talk**
  - **Layer 1** Basic χ^2 fitting
  - **Layer 2** Rescaled coordinates
  - **Layer 3** Correlations and Eigenvalue scales
  - **Model comparison with evidence**
  - **Layer 4** The hypersphere
    - Projection onto radius
    - Earlier work
  - **Simulation results**
  - **Layer 5** Incomplete or unsolved issues
  - **South Africa**



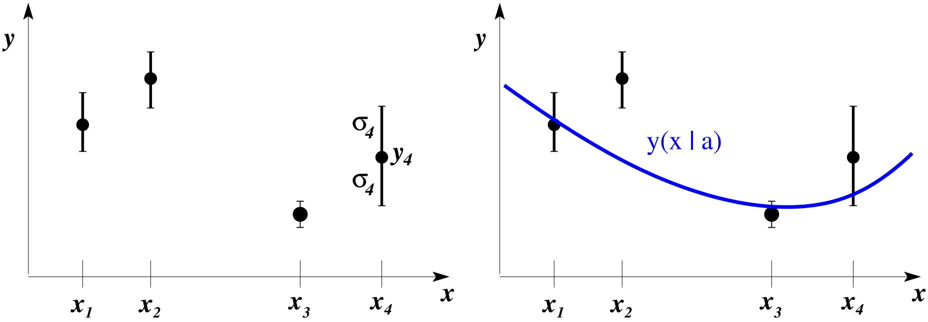
Layer 1

Gaussian regression: dimensioned version

Data, experimental uncertainties, trial function

No. of data points	N	Trial function	$y(x \mid \boldsymbol{a})$
Data means	$\boldsymbol{y} = (y_1, \ldots, y_N)$	with K parameters	$\boldsymbol{a} = (a_1, a_2, \ldots, a_K)$
Data uncertainties, taken as constants	$\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_N)$	Aim: Small χ^2	$\chi^2 \sim (\text{data} - \text{model})^2$
		Aim: best-fit values	$\hat{\boldsymbol{a}}$

Implicit aim: model comparison



Basic χ^2: model assumptions

Trial function $$y(x\,|\,\boldsymbol{a}) = \sum_{k=1}^{K} f_k(x)\, a_k$$

made up of K functions $f_k(x)$ of any kind, differentiable, e.g. $\cos(k\phi)$

with K linear parameters $\boldsymbol{a} = (a_1, \ldots, a_K)$

Additive error model $$\varepsilon_n = y_n - y(x_n\,|\,\boldsymbol{a}) = \text{data} - \text{model}$$

Model errors
assumed gaussian iid $$p(\varepsilon_n\,|\,\sigma_n) = \frac{e^{-\varepsilon_n^2/2\sigma_n^2}}{\sqrt{2\pi}\sigma_n}$$

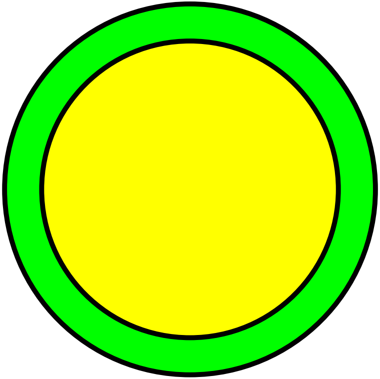
Likelihood $$p(\boldsymbol{y}\,|\,\boldsymbol{a}) = C_{\boldsymbol{\sigma}} \cdot \exp\left[-\frac{1}{2}\sum_{n=1}^{N}\left(\frac{y_n - y(x_n\,|\,\boldsymbol{a})}{\sigma_n}\right)^2\right]$$

$p(\boldsymbol{y}\,|\,\boldsymbol{a}) = C_\sigma\, e^{-\chi^2/2}$ $$C_{\boldsymbol{\sigma}} = \left[(2\pi)^{N/2}\prod_n \sigma_n\right]^{-1}$$

contains the usual χ^2 $$\chi^2(\boldsymbol{a}) = \sum_{n=1}^{N}\left(\frac{y_n - \sum_k f_k(x_n)\, a_k}{\sigma_n}\right)^2$$

χ^2 (Gaussian) fitting in vector notation

Design matrix or "response matrix"	$(\mathbb{A})_{nk} = \dfrac{f_k(x_n)}{\sigma_n}$
Response vector (scaled data)	$\boldsymbol{z} = \left(\dfrac{y_1}{\sigma_1}, \dots \dfrac{y_N}{\sigma_N}\right)$
Likelihood	$p(\boldsymbol{y} \mid \boldsymbol{a}) = C_\sigma e^{-\chi^2/2}$
Chisquared	$\chi^2(\boldsymbol{a}) = (\boldsymbol{z} - \mathbb{A}\boldsymbol{a})^\intercal(\boldsymbol{z} - \mathbb{A}\boldsymbol{a})$
Hessian	$\dfrac{1}{2}\dfrac{\partial^2\chi^2}{\partial \boldsymbol{a}^2} = \mathbb{A}^\intercal\mathbb{A} \;=\; \mathbb{H}_d$ is constant for linear model
Maximum likelihood	$0 = \left.\dfrac{\partial\chi^2}{\partial \boldsymbol{a}}\right\rvert_{\boldsymbol{a}=\hat{\boldsymbol{a}}} = 2\left(-\mathbb{A}^\intercal\boldsymbol{z} + \mathbb{A}^\intercal\mathbb{A}\hat{\boldsymbol{a}}\right)$
Mode (best-fit parameters)	$\hat{\boldsymbol{a}} = (\mathbb{A}^\intercal\mathbb{A})^{-1}\mathbb{A}^\intercal\boldsymbol{z} \;=\; \mathbb{H}_d^{-1}\mathbb{A}^\intercal\boldsymbol{z}$
Covariance matrix	$\mathbb{V}_d = \mathbb{H}_d^{-1}$



Layer 2

Gaussian regression scaled version

Dimensional analysis

Physical dimensions $\Longrightarrow$ very large or small eigenvalues, determinants

Start with usual chisquared

$$\chi^2(\boldsymbol{a}) = \sum_{n=1}^{N} \left(\frac{y_n - \sum_k f_k(x_n)\, a_k}{\sigma_n} \right)^2$$

Physical dimensions:

$\dim(y_n) = Y$ (metres, GeV, ...)

$\dim(\sigma_n) = Y$

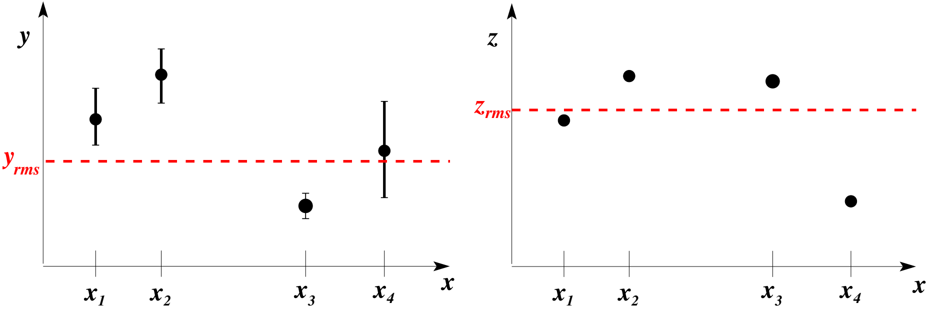
so $z_n = y_n/\sigma_n$ is dimensionless

$\dim(a_k) = Y$ if we assume $f_k(x)$ is dimensionless.

Scaling with N:

$\sum_n$ means that χ^2 grows asymptotically almost linearly with N

There are TWO physical scales



mean scale $y_{\rm rms} = \left[\frac{1}{N} \sum_n y_n^2 \right]^{1/2}$ **precision scale** $z_{\rm rms} = \left[\frac{1}{N} \sum_n z_n^2 \right]^{1/2}$

$$S_N^2 = \sum_n z_n^2 = N \langle \boldsymbol{z}^2 \rangle \qquad \langle \boldsymbol{z}^2 \rangle = z_{\rm rms}^2$$

Double rescaling

mean scale $$y_{\mathrm{rms}} = \left[\frac{1}{N} \sum_n y_n^2 \right]^{1/2}$$

precision scale $$z_{\mathrm{rms}} = \left[\frac{1}{N} \sum_n z_n^2 \right]^{1/2}$$

dimensionless parameters $$\beta_k = \frac{a_k}{y_{\mathrm{rms}}}$$

precision-scaled data $$u_n = \frac{z_n}{z_{\mathrm{rms}}}$$

$$\langle \boldsymbol{u}^2 \rangle = 1$$

precision-scaled uncertainties $$s_n = \frac{\sigma_n \, z_{\mathrm{rms}}}{y_{\mathrm{rms}}}$$

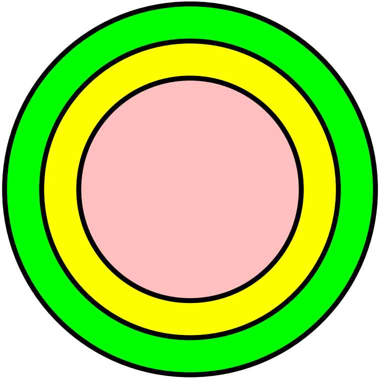
$$\chi^2(\boldsymbol{a}) = \sum_n \left(\frac{y_n - \sum_k f_k(x_n)\, a_k}{\sigma_n} \right)^2$$

$$Q(\boldsymbol{\beta}) = \frac{\chi^2(\boldsymbol{a})}{N \langle \boldsymbol{z}^2 \rangle} = \sum_n \left(u_n - \sum_k g_k(x_n)\, \beta_k \right)^2$$

$$g_k(x_n) = \frac{f_k(x_n)}{s_n} = \mathbb{G}_{kn}$$

Vector-matrix notation

Mode	$\hat{\boldsymbol{\beta}} = (\mathbb{G}^{\intercal}\mathbb{G})^{-1}\mathbb{G}^{\intercal}\boldsymbol{u}$
Hessian	$\mathbb{H} = \dfrac{1}{2}\dfrac{\partial^2 Q}{\partial \boldsymbol{\beta}^2} = \dfrac{\mathbb{G}^{\intercal}\mathbb{G}}{N}$
Scaled χ^2	$Q(\boldsymbol{\beta}) = \dfrac{1}{N}(\boldsymbol{u} - \mathbb{G}\boldsymbol{\beta})^{\intercal}(\boldsymbol{u} - \mathbb{G}\boldsymbol{\beta})$
	$= Q(\hat{\boldsymbol{\beta}}) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^{\intercal}\mathbb{H}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})$
Minimum χ^2	$Q(\hat{\boldsymbol{\beta}}) = 1 - \hat{\boldsymbol{\beta}}^{\intercal}\mathbb{H}\hat{\boldsymbol{\beta}}$
Scaled likelihood	$p(\boldsymbol{u} \mid \boldsymbol{\beta}) = \dfrac{\langle \boldsymbol{z}^2 \rangle^{K/2}}{(2\pi)^{N/2}} \exp\left[-\dfrac{N\langle \boldsymbol{z}^2 \rangle \, Q(\boldsymbol{\beta})}{2}\right]$



Layer 3

Correlation and eigenvalue scales

Diagonalisation: rotation in parameter space

- Eigenvalues λ, orthonormal eigenvectors $\boldsymbol{e}$ of $\mathbb{H}$:

$$\mathbb{H}\, \boldsymbol{e}_\ell = \boldsymbol{e}_\ell \lambda_\ell$$

- Orthogonal eigenvalue matrix

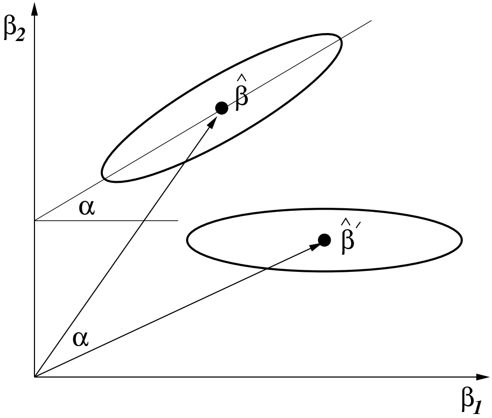
$$\mathbb{S} = (\boldsymbol{e}_1 \ \cdots \ \boldsymbol{e}_K)$$

- Diagonalisation via

$$\mathbb{H}\mathbb{S} = \mathbb{S}\mathbb{L}$$
$$\mathbb{L} = \mathrm{diag}(\lambda_1, \ldots, \lambda_N)$$

rotates Hessian and covariance matrix

$$\boldsymbol{\beta} = \mathbb{S}\boldsymbol{\beta}'$$



$$Q(\hat{\boldsymbol{\beta}}') = 1 - \hat{\boldsymbol{\beta}}'^{\intercal}\mathbb{L}\hat{\boldsymbol{\beta}}'$$

$$Q(\boldsymbol{\beta}') = Q(\hat{\boldsymbol{\beta}}') + (\boldsymbol{\beta}' - \hat{\boldsymbol{\beta}}')^{\intercal}\mathbb{L}(\boldsymbol{\beta}' - \hat{\boldsymbol{\beta}}')$$

Squeeze: rescaling in parameter space

Parameters

$$\boldsymbol{b} = \mathbb{L}^{1/2}\boldsymbol{\beta}' \;=\; \mathbb{L}^{1/2}\mathbb{S}^{\mathsf{T}}\boldsymbol{\beta}$$

Minimum χ^2

$$Q(\hat{\boldsymbol{b}}) = 1 - \hat{\boldsymbol{b}}^2$$

Scaled χ^2

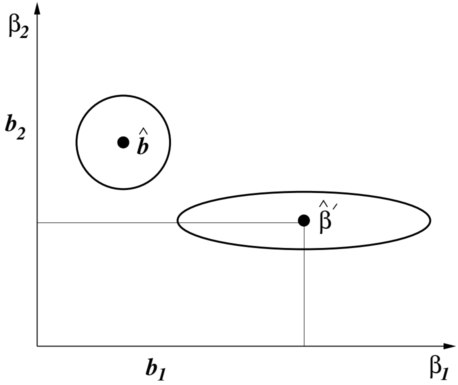
$$Q(\boldsymbol{b}) = Q(\hat{\boldsymbol{b}}) + (\boldsymbol{b} - \hat{\boldsymbol{b}})^2$$
$$= 1 + \boldsymbol{b}^2 - 2\hat{\boldsymbol{b}}^{\mathsf{T}}\boldsymbol{b}$$

Limit on radius

$$Q(\hat{\boldsymbol{b}}) = 1 - \hat{\boldsymbol{b}}^2 \geq 0$$

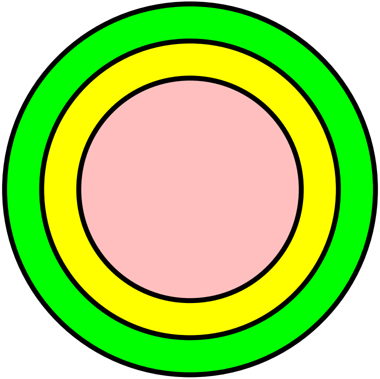
$$\hat{\boldsymbol{b}}^2 \leq 1$$

$$S_N^2 = N\langle \boldsymbol{z}^2 \rangle = \sum_{n=1}^{N} z_n^2$$



Likelihood

$$p(\boldsymbol{u} \mid \boldsymbol{b}) = \frac{\langle \boldsymbol{z}^2 \rangle^{K/2}}{(2\pi)^{N/2}} \exp\left[-\frac{1}{2} S_N^2\, Q(\boldsymbol{b})\right]$$



Interlude

Model comparison, evidence and K

Model comparison with evidence: failure of uniform priors

- Given model $\mathcal{M}$ with parameters $\boldsymbol{b}$ and data $\boldsymbol{y}$, its **evidence** is

$$p(\boldsymbol{y} \,|\, \mathcal{M}) = \int d\boldsymbol{b}\, p(\boldsymbol{y} \,|\, \boldsymbol{b}, \mathcal{M})\, p(\boldsymbol{b} \,|\, \mathcal{M})$$

- Try uniform priors in K dimensions

$$p(\boldsymbol{b} \,|\, \mathcal{M}) = \prod_{k=1}^{K} \frac{1}{\Delta_k}$$

with cutoff parameters such as

$$-\Delta_k/2 \leq a_k \leq \Delta_k/2$$

Model comparison with evidence: failure of uniform priors

- Given evidence with prior in K dimensions

$$p(\boldsymbol{y} \,|\, \mathcal{M}) = \int d\boldsymbol{b}\, p(\boldsymbol{y} \,|\, \boldsymbol{b}, \mathcal{M})\, p(\boldsymbol{b} \,|\, \mathcal{M}) \qquad\qquad p(\boldsymbol{b} \,|\, \mathcal{M}) \sim (\text{uniform})^K$$

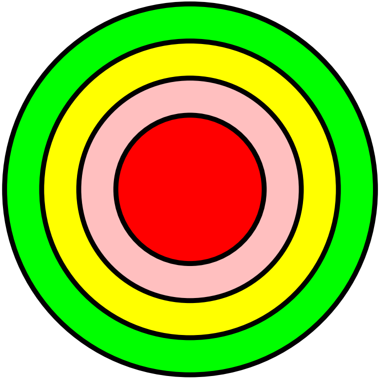
- **Model comparison with Bayes Factors**:
  Compare $\mathcal{M}_1$ **with K_1 parameters** to $\mathcal{M}_2$ **with $K_2 > K_1$ parameters.**
  Even choosing priors identical for all $1 \leq k \leq K_1$, we still get a result that depends on the prior widths for $K_1 < k \leq K_2$:

$$\frac{p(\boldsymbol{y} \,|\, \mathcal{M}_1)}{p(\boldsymbol{y} \,|\, \mathcal{M}_2)} = \frac{\int d\boldsymbol{b}_1\, p(\boldsymbol{y} \,|\, \boldsymbol{b}_1, \mathcal{M}_1)\, p(\boldsymbol{b}_1 \,|\, \mathcal{M}_1)}{\int d\boldsymbol{b}_2\, p(\boldsymbol{y} \,|\, \boldsymbol{b}_2, \mathcal{M}_2)\, p(\boldsymbol{b}_2 \,|\, \mathcal{M}_2)}$$

$$= \frac{\int d\boldsymbol{b}_1\, p(\boldsymbol{y} \,|\, \boldsymbol{b}_1, \mathcal{M}_1)}{\int d\boldsymbol{b}_2\, p(\boldsymbol{y} \,|\, \boldsymbol{b}_2, \mathcal{M}_2)} \left(\prod_{k=K_1+1}^{K_2} \Delta_k \right)$$

  which can take on **any value you like**, depending on the choice of the extra Δ_k. $\Longrightarrow$ **Failure of uniform priors, Failure of Bayes?**

- **How to compare models with different parameter space dimensionalities K?**



Layer 4

The hypersphere and r-priors

Uniform prior for b on the K-hypersphere

- We are now looking for a prior $p(\boldsymbol{b} \,|\, \mathcal{M})$.
- For a **prior**, there is no data, therefore **we do not know $\hat{\boldsymbol{b}}$.**
- We are dealing with a $K-1$ dimensional model on any hyperplane defined by $b_k = 0$. Similarly for double-zero planes $b_k = 0, b_\ell = 0$.
- The **origin** $\boldsymbol{b} = 0$ of the K-dimensional parameter space is special in the sense that it corresponds to zero free parameters. It is therefore a point with unique status.
- Information Symmetry Postulate:
  We have no reason to prefer any particular direction on the hypersphere.
- Hence use the prior centered on $\boldsymbol{b} = 0$ which is **uniform over the surface of the K-hypersphere** at any radius:

$$p(\boldsymbol{b} \,|\, \mathcal{M}) = \frac{p(b)\,\Gamma(K/2)}{b^{K-1}\, 2\pi^{K/2}} \tag{1}$$

Radius-conditional evidence

- The evidence cannot be solved directly. However, by introducing a new degree of freedom r, a **radius** in the K-sphere, we can make progress:

$$
\begin{aligned}
p(\boldsymbol{u} \,|\, \mathcal{M}) &= \int dr \, d\boldsymbol{b} \, p(\boldsymbol{u}, \boldsymbol{b}, r \,|\, \mathcal{M}) \\
&= \int_0^\infty dr \, p(r \,|\, \mathcal{M}) \int_{\mathbb{R}^K} d\boldsymbol{b} \, p(\boldsymbol{u} \,|\, \boldsymbol{b}, r, \mathcal{M}) \, p(\boldsymbol{b} \,|\, r, \mathcal{M}) \\
&= \int_0^\infty dr \, p(r \,|\, \mathcal{M}) \, p(\boldsymbol{u} \,|\, r, \mathcal{M})
\end{aligned}
$$

  in terms of radius-conditional evidence

$$
p(\boldsymbol{u} \,|\, r, \mathcal{M}) = \int_{\mathbb{R}^K} d\boldsymbol{b} \, p(\boldsymbol{u} \,|\, \boldsymbol{b}, r, \mathcal{M}) \, p(\boldsymbol{b} \,|\, r, \mathcal{M})
$$

  projecting K-dimensional parameter space onto 1-dimensional r-space.

- **Drop the $\mathcal{M}$ in the notation, use $\mathcal{H}$ for different r-priors.**

Projection onto radius-conditional evidence I

- We want to find $p(\boldsymbol{u} \,|\, r) = \int_{\mathbb{R}^K} d\boldsymbol{b}\, p(\boldsymbol{u} \,|\, \boldsymbol{b}, r)\, p(\boldsymbol{b} \,|\, r)$
- Choose

$$p(b \,|\, r) = \delta(b - r) \;=\; 2r\,\delta(\boldsymbol{b}^2 - r^2)$$

to get a prior radially symmetric about 0 and uniform on the K-sphere

$$p(\boldsymbol{b} \,|\, r) = \frac{p(b \,|\, r)\,\Gamma(K/2)}{r^{K-1}\, 2\pi^{K/2}} \;=\; \frac{\delta(\boldsymbol{b}^2 - r^2)\,\Gamma(K/2)}{r^{K-2}\, \pi^{K/2}}$$

- Fourier representation of delta function is close to integral representation of confluent hypergeometric function:

$$\delta(\boldsymbol{b}^2 - r^2) = \int_{-i\infty}^{+i\infty} \frac{ds}{2\pi i} \exp\left[sr^2 - s\boldsymbol{b}^2\right]$$

$${}_0F_1(c\,;\,z) = \frac{\Gamma(c)}{2\pi i} \int_{-i\infty}^{+i\infty} dv\, v^{-c} \exp\left(v + \frac{z}{v}\right),$$

Projection onto radius-conditional evidence II

- The likelihood $p(\boldsymbol{u}\,|\,\boldsymbol{b},r) = p(\boldsymbol{u}\,|\,\boldsymbol{b})$ is independent of r,

$$p(\boldsymbol{u}\,|\,\boldsymbol{b}) = \frac{\langle \boldsymbol{z}^2\rangle^{K/2}\, e^{-S_N^2/2}}{(2\pi)^{N/2}} \left[-\frac{1}{2}S_N^2\boldsymbol{b}^2 + S_N^2\hat{\boldsymbol{b}}^\top\boldsymbol{b}\right],$$

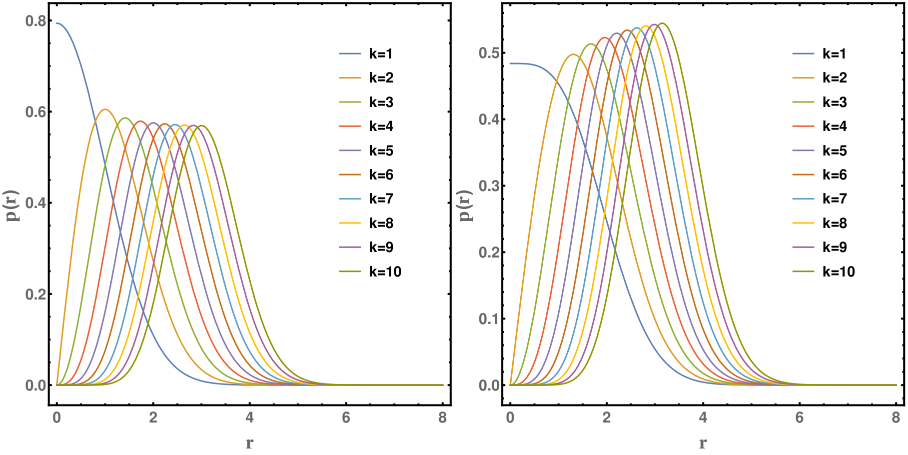
- As a result

$$p(\boldsymbol{u}\,|\,r) = \int_{\mathbb{R}^K} d\boldsymbol{b}\, p(\boldsymbol{u}\,|\,\boldsymbol{b})\, p(\boldsymbol{b}\,|\,r)$$
$$= \frac{\langle \boldsymbol{z}^2\rangle^{K/2}}{(2\pi)^{N/2}} \exp\left[-\frac{1}{2}S_N^2(1+r^2)\right] {}_0F_1\left(\frac{K}{2}\,;\,\frac{1}{4}S_N^4\,\hat{\boldsymbol{b}}^2 r^2\right)$$

so the evidence has reduced to the one-dimensional r-integral

$$\boxed{p(\boldsymbol{u}\,|\,\mathcal{M}) = \frac{\langle \boldsymbol{z}^2\rangle^{(K/2)-1}}{(2\pi)^{N/2}} \int dr\, p(r)\, \exp\left[-\frac{1}{2}S_N^2(1+r^2)\right] {}_0F_1\left(\frac{K}{2}\,;\,\frac{S_N^4}{4}\,\hat{\boldsymbol{b}}^2 r^2\right)}$$

Plot of radius-conditional evidence



Representative plots of $\exp\left[-\frac{1}{2}S_N^2(1+r^2)\right]\ {}_0F_1\left(\frac{K}{2}\,;\ \frac{S_N^4}{4}\,\hat{\boldsymbol{b}}^2 r^2\right)$

Precursors for r-priors I

- Notation: $\mathcal{H}$ for different r-priors
- **Jeffreys (1967)**: *null hypothesis* vs λ *unknown* hypothesis; obtained a one-dimensional Cauchy prior:

$$f(\lambda \,|\, \sigma, \mathcal{H}_J) = \frac{1}{\pi(1+\lambda^2/\sigma^2)}$$

- **Zellner and Siow (1980)** generalised to a K-dimensional Cauchy prior,

$$p(\boldsymbol{\beta} \,|\, \sigma, \mathcal{H}_{ZS}) = \frac{\Gamma[(K+1)/2]}{\pi^{(K+1)/2}} \, \frac{\sqrt{\det \mathbb{H}/N\sigma^2}}{1+\boldsymbol{\beta}^{\intercal}\mathbb{H}\boldsymbol{\beta}/N\sigma^2}$$

  Problem: evidence based on Cauchy cannot be found in closed form.

- **Zellner (1986)** simplified g-prior based on knowledge of $\mathbb{H}$ ($\sigma = 1$ here)

$$p(\boldsymbol{b} \,|\, g, \sigma, \mathcal{H}_Z) = \frac{e^{-N\boldsymbol{b}^2/2\sigma^2 g}}{(2\pi\sigma^2 g)^{K/2}}.$$

Precursors for r-priors II

- Problem: Zellner's g-prior leads to closed form for evidence; **however**, it suffers from logical contradictions for limiting cases.
- **Liang et al (2008)**: A **hyper-g prior** which is a mixture of g-priors were proposed in 2008. Similar to our r-prior, Liang et al propose a g-prior with hyperparameter a

$$p(g \mid a, \mathcal{H}_g) = \frac{a-2}{2}(1+g)^{-a/2}$$

  with $a \leq 2$ leading to improper and $a > 2$ to proper priors.
- Zellner's g-prior and Liang's hyper-g prior are **widely used** in the statistics community but almost unknown in physics/astronomy.
- **All of the above can be expressed as special forms of r-priors**, i.e. r-priors provide a **general framework** for all previous results which can be compared and analysed.

Example: Parabolic r-prior

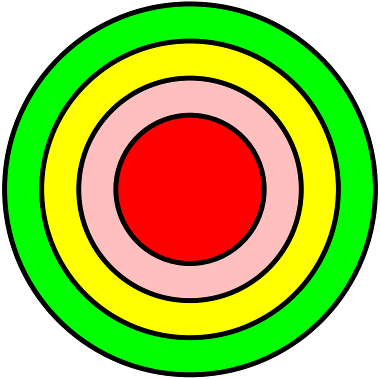
- The **parabolic r-prior** was chosen to resemble the hyper-g asymptotically,

$$p(r \,|\, K, \mathcal{H}_r) = \frac{K}{r\sqrt{\pi}} \left(\frac{Nr^2}{2}\right)^{K/2} U\left[\frac{K+1}{2}; \frac{1}{2}; \frac{Nr^2}{2}\right]$$

  with U the second solution of the confluent hypergeometric function differential equation
- This yields **evidence**

$$p(\boldsymbol{u} \,|\, \mathcal{H}_r) = \frac{\langle \boldsymbol{z}^2 \rangle^{K/2}}{(2\pi)^{N/2} 2^K} e^{-S_N^2/2} \, {}_1F_1\left(\frac{K+1}{2}\,;\, K+1\,;\, \frac{S_N^2 \hat{\boldsymbol{b}}^2}{2}\right)$$

- **Dependent on choice of r-prior, the evidence is a closed solution in terms of hypergeometric functions ${}_1F_1$ or ${}_2F_1$.**
- **Once the evidence is known, posteriors and characteristic functions can be found trivially.**



Results

Model comparison with simulated data

Simulation of data

- For the purposes of the comparison, a "model" consists of
  - Choice of r-prior or information criterion
  - Assumption of a trial function $y(x \mid \beta_1, \ldots, \beta_j)$ with j free parameters $\beta_1, \ldots, \beta_j$ and functions $f_1, \ldots, f_j$ (use cosines).
- **Generation of simulated data**
  1. Generate a set of parameters $\boldsymbol{\beta}_i$ from a Cauchy distribution

$$\boldsymbol{\beta} \sim \prod_k \frac{1}{\pi(1+\beta_k^2/B^2)} \qquad B = \begin{cases} 1 & \text{weak signal} \\ 5 & \text{strong signal} \end{cases}$$

  2. **For given fixed choice of** K_{true} and fixed parameter set $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)$ generate $N = 100$ simdata points with fluctuation

$$y_n(K_{\text{true}}) = \varepsilon_n + \sum_{k=1}^{K-true} f_k(x_n)\, \beta_k \qquad \varepsilon_n \sim \mathcal{N}(0, E) \quad \text{(Normal distribution)}$$

  3. Repeat for different K_{true} and 1000 different parameter sets and hence simulated datasets.

Comparison of models using simulated data

For given model prior or information criterion

e.g. Zellner g, Liang hyper-g, Parabolic r-prior, Information criteria

1. Run different **trial dimensions** $K_{\mathrm{trial}} = 1, 2, \ldots, 50$ and use trial function

$$y(x \mid \boldsymbol{\beta}, K_{\mathrm{trial}}) = \sum_{k=1}^{K-\mathrm{trial}} f_k(x)\, \beta_k$$

2. Find **evidence** for each, $p(\boldsymbol{y} \mid K_{\mathrm{trial}}, \mathcal{H})$
3. Find the **best dimension** K_{best} which yields the best evidence

$$K_{\mathrm{best}} = \max_{K\mathrm{trial}} p(\boldsymbol{y} \mid K_{\mathrm{trial}}, \mathcal{H})$$

   and the corresponding value of the parameters $\hat{\beta}(K_{\mathrm{best}})$.

4. K_{best} should ideally be equal to K_{true} but the model must find K_{best} without knowing K_{true}.

Comparison of models using simulated data

5. Compute the **squared error** between this "best shot" of a given prior $\mathcal{H}$ and the simdata

$$SE(K_{\mathrm{true}}, i) = \sum_n \left(y_n(K_{\mathrm{true}}) - \sum_{k=1}^{K-\mathrm{best}} f_k(x_n)\, \hat{\beta}(K_{\mathrm{best}}) \right)^2$$
$$= \left\| \mathbb{A}\beta(K_{\mathrm{true}}) - \mathbb{A}\hat{\beta}(K_{\mathrm{best}}) \right\|_i^2$$

6. For each K_{true} average the above over all datasets i to get the **Mean Square Error**

$$MSE(K_{\mathrm{true}}) = \frac{1}{\mathrm{I}} \sum_i SE(K_{\mathrm{true}})$$

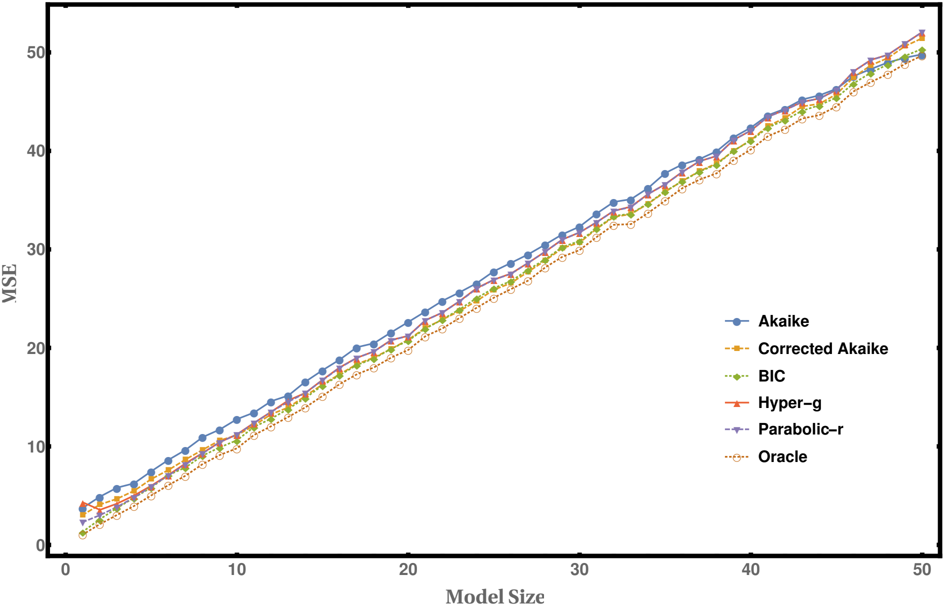
7. The **oracle** case is defined by eliminating the uncertainty in the dimensionality of the appropriate model,

$$K_{\mathrm{best}} = K_{\mathrm{true}}$$

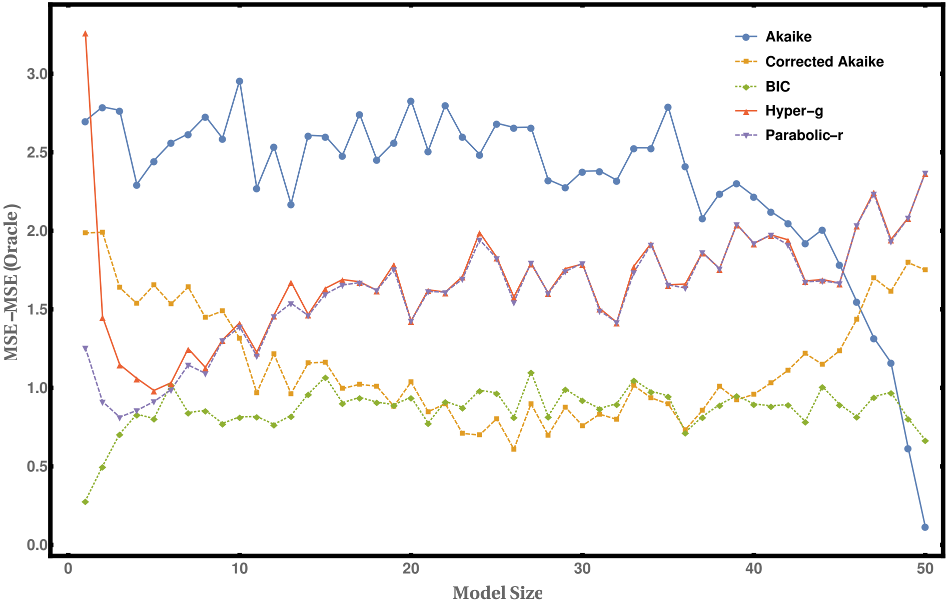
Effective information criteria

- **Akaike Information Criterion (AIC)** is an information-theory (non-Bayesian) based Occam's penalty.
- **Bayes Information Criterion (BIC)** effectively uses a r^{K-1} improper prior.
- **Corrected Akaike Information Criterion (AICc)** makes additional assumptions, without improving the result.

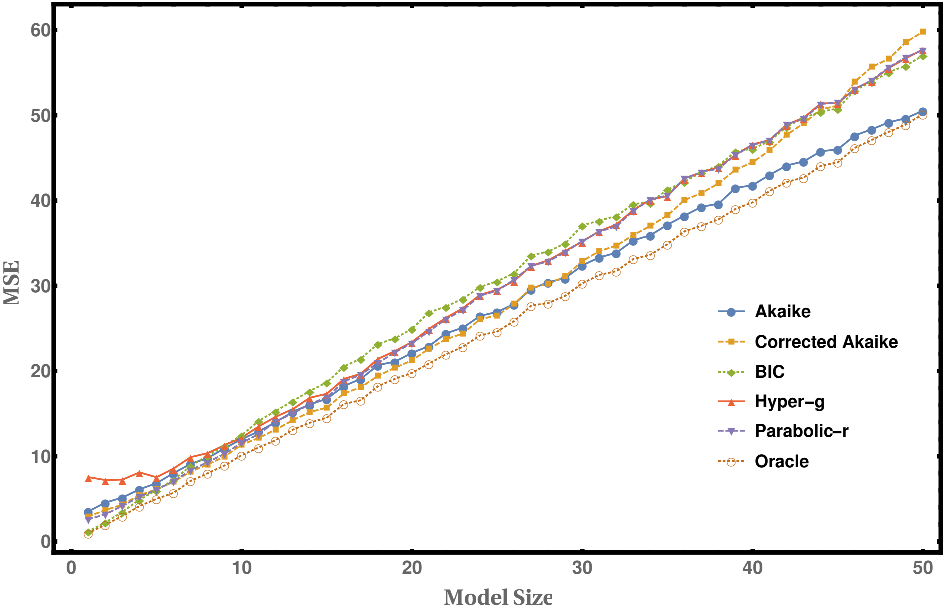
Strong Signal MSE



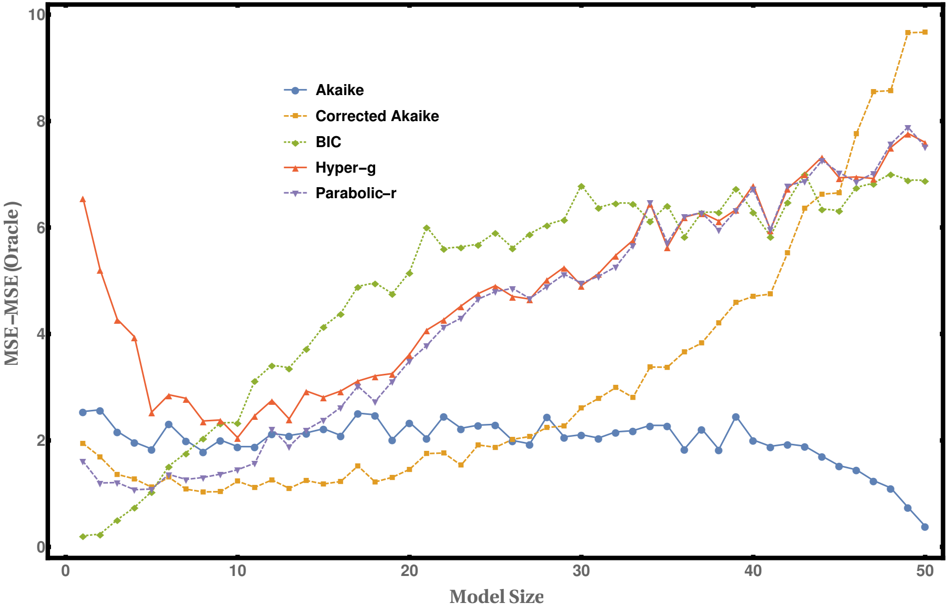
Strong Signal, Differences with Oracle



Weak Signal MSE

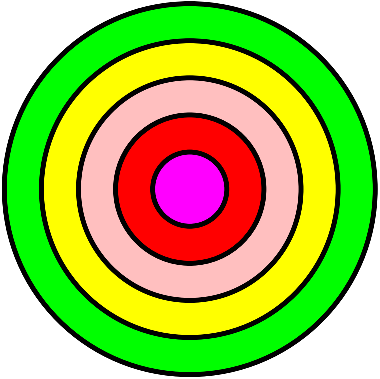


Weak Signal, Differences with Oracle



Conclusions from simulations

- **Strong differences depending on signal strength** (variability of parameters in model mix)
- ***r*-prior converges to hyper-*g* for large *K*** as expected
- ***r*-prior much better than hyper-*g* for small *K***
- **Room for improvement!**



Layer 5

Incomplete or unsolved issues

Incomplete or unsolved issues

INCOMPLETE

- **Work for variable** σ already done; needs updating
- **Alternative model with offset component**

$$y(x \mid \alpha, \boldsymbol{\beta}) = \alpha + \sum_{k=1}^{K} f_k(x)\, \beta_k \; = \; \mathbb{I}\alpha + \mathbb{A}\boldsymbol{\beta}$$

- **Relation between correlations and eigenvalues:** should be known from advanced linear algebra.

UNSOLVED

- **Metapriors on the origin of the hypersphere?**
- Implicit and explicit K**-dependence in delta function and prior**: are the scales correct?

Nachrichten in 100 Sekunden

- **Model comparison across different parameter space dimensions is surprisingly subtle**
- **r-priors represent major progress**
  - Analytical results
  - Framework for everything that has been done so far
- **Some stuff still not understood**

References

- M.B. de Kock and H.C. Eggers
  "Bayesian variable selection with spherically symmetric priors"
  Communications in Statistics (Theory and Methods) (in press)
  arXiv:1410.0891 [math.ST] (only version 1 online so far)
- H. Jeffreys *Theory of Probability* (1967)
- A. Zellner and A. Siow
  "Posterior odds ratios for selected regression hypotheses"
  Trabajos de estadística y de investigación operativa, 31(1):585–603 (1980).
- A. Zellner
  "On assessing prior distributions and bayesian regression analysis with g-prior distributions"
  Bayesian Inference and Decision Techniques: Essays in Honor of Bruno De Finetti 6:233–243 (1986)
- F. Liang, R. Paulo, G. Molina, M.A. Clyde and J.O. Berger,
  "Mixtures of g-priors for bayesian variable selection",
  Journal of the American Statistical Association, 103(481) (2008)