Features space analysis for multimodel selection

A. A. Aduenko, V. V. Strijov

Moscow Institute of Physics and Technology
Computing Center of Russian Academy of Sciences

IDP Conference, Barcelona
14th October 2016

Goals

Goal: construct mathematical apparatus to select multimodels for solving recognition and classification tasks.

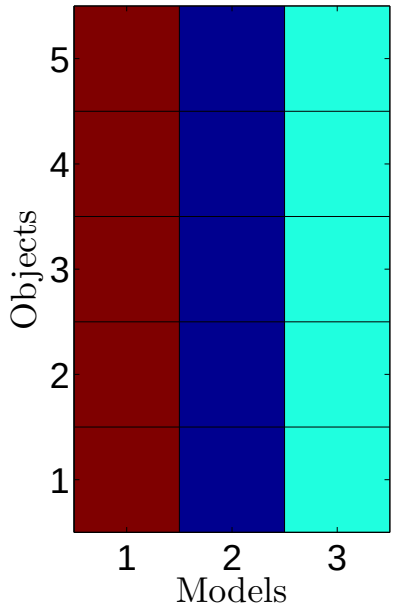
Motivation. Statistical inhomogeneity of a sample arises in recognition and classification problems. Multimodels are used to handle the issue. They contain several models, for which we aim to determine statistical discernability.

Problem. Multimodel may contain many similar models, which results in low forecast quality and lack of interpretability. Models' feature spaces might be different, in particular they can have different dimensionality.

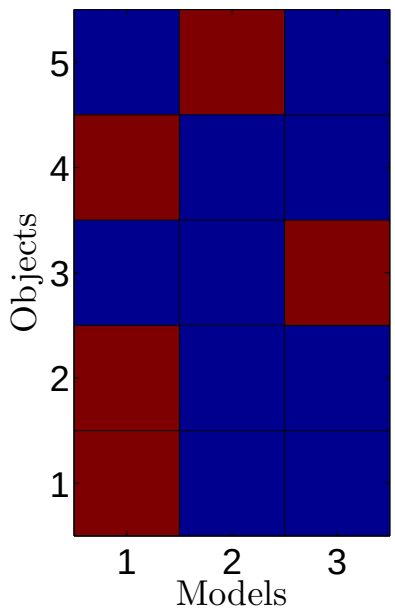
Method. Multidimensional statistics and bayesian inference to construct a method for statistical testing of models' discernability. Similarity function is introduced and analyzed. The function is defined for a pair of distributions possibly with different supports.

Definition 1. Mixture of models is a regression model $f = \sum_{k=1}^{K} \pi_k f_k(\mathbf{w}_k)$,

where $\sum_{k=1}^{K} \pi_k = 1,\ \pi_k \geq 0.$



Definition 2. Multilevel regression models is a union of regression models f_k, $k = 1, \ldots, K$ such that the objects' index set is divided as follows $\mathcal{I} = \sqcup_{k=1}^{K} \mathcal{I}_k$ and for each object with index in $\mathcal{I}_k$ model f_k is used.



Probabilistic model for mixture of models

Data generation hypothesis

- There exists a prior distribution on vector of models' weights $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_K]^\mathsf{T} \sim q(\boldsymbol{\pi}|\alpha)$.
- Models' parameters $\mathbf{w}_1, \ldots, \mathbf{w}_K$ are mutually independent together with models' weight vector $\boldsymbol{\pi}$.
- Each object $\mathbf{x}_i$ is described by a single model k_i, and random variables $k_1, \ldots, k_m$ corresponding to model indices are mutually independent.
- Target variables $y_i|k_i, \mathbf{w}_{k_i} \sim \mathrm{Be}\big(f_{k_i}(\mathbf{x}_i, \mathbf{w}_{k_i})\big)$ are mutually independent together with $\boldsymbol{\pi}$.

Joint distribution for a multimodel

$$p(\mathbf{y}, \mathbf{w}_1, \ldots, \mathbf{w}_K, \boldsymbol{\pi}|\mathbf{X}) =$$

$$q(\boldsymbol{\pi}|\alpha) \prod_{j=1}^{K} p_j(\mathbf{w}_j) \prod_{i=1}^{m} \left(\sum_{k=1}^{K} \pi_k f_k(\mathbf{x}_i, \mathbf{w}_k)^{y_i} \big(1 - f_k(\mathbf{x}_i, \mathbf{w}_k)\big)^{1-y_i} \right).$$

Multimodel optimality and adequacy condition

Definition 3. Call a multimodel defined by the joint distribution $p(\mathbf{y},\ \mathbf{w}_1,\ \ldots,\ \mathbf{w}_K,\ \boldsymbol{\pi}|\mathbf{X})$ $(s,\ \alpha)$-**adequate**, if the models constituting the multimodel are pairwise statistically distinguishable with the similarity function s at significance level α.
Denote the set of all $(s,\ \alpha)$-adequate multimodels by $\mathcal{M}_{s,\,\alpha}$.
Definition 4. Call a multimodel **optimal**, if it has the maximum evidence $[q(\pi|\alpha),\ p_1(\mathbf{w}_1),\ \ldots,\ p_K(\mathbf{w}_K)] = \underset{q,\,p_1,\,\ldots,\,p_K}{\arg\max}\ p(\mathbf{y}|\mathbf{X}) =$

$$\underset{q,\,p_1,\,\ldots,\,p_K}{\arg\max} \int p(\mathbf{y},\ \mathbf{w}_1,\ \ldots,\ \mathbf{w}_K,\ \boldsymbol{\pi}|\mathbf{X}) d\mathbf{w}_1 \ldots d\mathbf{w}_K d\boldsymbol{\pi}.$$

Maximum a posteriori probability estimate for models' parameters and multimodel's weights

$$[\boldsymbol{\pi},\ \mathbf{w}_1,\ \ldots,\ \mathbf{w}_K] = \underset{\boldsymbol{\pi},\,\mathbf{w}_1,\,\ldots,\,\mathbf{w}_K}{\arg\max}\ p(\boldsymbol{\pi},\ \mathbf{w}_1,\ \ldots,\ \mathbf{w}_K|\mathbf{X},\ \mathbf{y}).$$

EM-algorithm for the mixture of logistic regression models

Joint distribution for the mixture of models

Introduce hidden variables $\{z_{ik} \in \{0,\ 1\}\}$ where $z_{ik} = 1$ means that an object $(\mathbf{x}_i,\ y_i)$ belongs to model k.

$$p(\mathbf{y},\ \mathbf{Z},\ \boldsymbol{\pi},\ \mathbf{w}_1,\ \ldots,\ \mathbf{w}_K|\mathbf{X},\ \mathbf{A}_1,\ \ldots,\ \mathbf{A}_K) = \prod_{k=1}^{K} p_k(\mathbf{w}_k|\mathbf{0},\ \mathbf{A}_k^{-1})$$

$$\frac{\Gamma(K\alpha)}{\Gamma^K(\alpha)} \prod_{k=1}^{K} \pi_k^{\alpha-1} \prod_{i=1}^{m} \prod_{k=1}^{K} \{\pi_k f(\mathbf{x}_i,\ \mathbf{w}_k)^{y_i}(1 - f(\mathbf{x}_i,\ \mathbf{w}_k))^{1-y_i}\}^{z_{ik}}.$$

E-step

$$\gamma_{ik} = \mathsf{E} z_{ik} = \pi_k f(\mathbf{x}_i,\ \mathbf{w}_k)^{y_i}(1 - f(\mathbf{x}_i,\ \mathbf{w}_k))^{1-y_i} / N_i.$$

M-step

At the M-step models' weights $\boldsymbol{\pi}$ and vectors of models' parameters $\mathbf{w}_1,\ \ldots,\ \mathbf{w}_K$ are defined.

$$\pi_k = \max(0,\ \gamma_k + \alpha - 1) / Z_k, \text{ где } \gamma_k = \sum_{i=1}^{m} \gamma_{ik}$$

M-step

$$\tilde{l}(\mathbf{w}_1, \ldots, \mathbf{w}_K, \boldsymbol{\pi}|\mathbf{X}, \mathbf{y}) =$$
$$-\sum_{k=1}^{K}(\gamma_k + \alpha - 1)\log \pi_k + \sum_{k=1}^{K}\tilde{l}_k(\mathbf{w}_k|\mathbf{X}, \mathbf{y}, \mathbf{A}_k).$$
$$\frac{\partial \tilde{l}_k}{\partial \mathbf{w}_k} = \mathbf{X}^\mathsf{T}\boldsymbol{\Gamma}_k(\mathbf{f} - \mathbf{y}) + \mathbf{A}_k\mathbf{w}_k,\ \mathbf{H}_k = \mathbf{X}^\mathsf{T}\mathbf{R}_k\mathbf{X} + \mathbf{A}_k,$$
$$\mathbf{R}_k = \mathrm{diag}(\gamma_{ik} f(\mathbf{x}_i^\mathsf{T}\mathbf{w}_k) f(-\mathbf{x}_i^\mathsf{T}\mathbf{w}_k)).$$

Properties of the optimized function

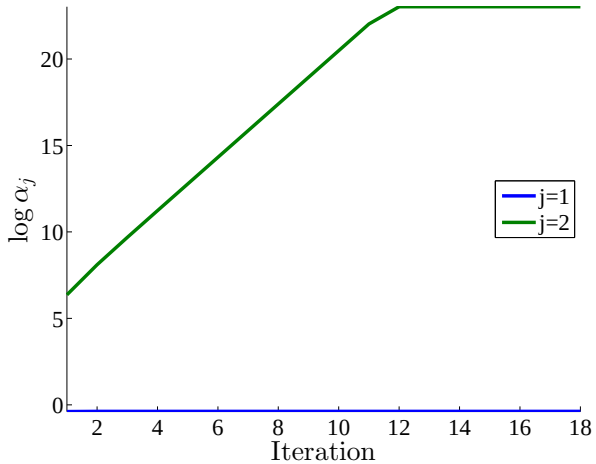
$\tilde{l}_k(\mathbf{y},\ \mathbf{w}_k|\mathbf{X},\ \mathbf{A}_k,\ \boldsymbol{\Gamma}_k)$ with fixed objects' weights $\boldsymbol{\Gamma}_k$ is the logarithm of joint distribution for a standard logistic regression model with weighted objects.

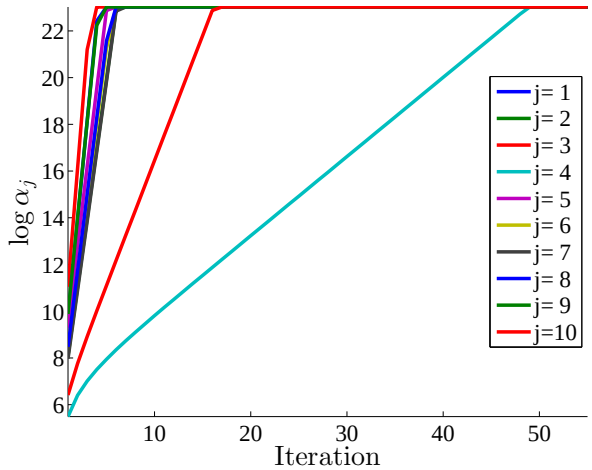
Suggested features' selection method

$\mathbf{A}_k = \arg\max_{\mathbf{A}\in\mathcal{M}} \tilde{p}(\mathbf{y}|\mathbf{X}, \mathbf{A},\ \boldsymbol{\Gamma}_k)$, где

$$\tilde{p}(\mathbf{y}|\mathbf{X},\ \mathbf{A},\ \boldsymbol{\Gamma}_k) = \int \tilde{l}_k(\mathbf{y},\ \mathbf{w}_k|\mathbf{X},\ \mathbf{A},\ \boldsymbol{\Gamma}_k)d\mathbf{w}_k.$$

Features' selection using maximum evidence principle





Theorem 1 (Aduenko, 2016)

Пусть $n = 2$, $k = 1$, $\mathbf{w} = [w_1,\ w_2]$, $w_1,\ w_2 \neq 0$. Denote by

$$\mathbf{\Sigma} = \mathbf{X}^{\mathsf{T}}\mathbf{R}\mathbf{X} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix},\ \mathbf{A} = \begin{pmatrix} s_1^2 & \kappa s_1 s_2 \\ \kappa s_1 s_2 & s_2^2 \end{pmatrix}.$$

If $m \to \infty$ and

- $\sigma_1^2,\ \sigma_2^2 \xrightarrow{\mathsf{P}} \infty$,
- $\exists\, c > 0 :\ \mathsf{P}(1 - \rho^2 \geq c) \to 1$,

then $s_1^*,\ s_2^* \xrightarrow{\mathsf{P}} \infty$, $\kappa^* \xrightarrow{\mathsf{P}} -\mathrm{sign}(w_1 w_2)$.

Model comparison problem statement

Issue

Despite the sparcification of a multimodel, it can still be not (s, α) – adequate, i.e. can contain similar models.

Input

- Two models f_1 and f_2 with vectors of parameters $\mathbf{w}_1$ and $\mathbf{w}_2$.
- Samples $(\mathbf{X}_1,\ \mathbf{y}_1)$ and $(\mathbf{X}_2,\ \mathbf{y}_2)$, $y_{1,i} = f_1(\mathbf{x}_{1,i},\ \mathbf{w}_1), \quad y_{2,i} = f_2(\mathbf{x}_{2,i},\ \mathbf{w}_2)$.
- Prior distributions on models' parameters $\mathbf{w}_1 \sim p_1(\mathbf{w}),\ \mathbf{w}_2 \sim p_2(\mathbf{w})$.
- Posterior distributions on models' parameters $p(\mathbf{w}_1|\mathbf{X}_1,\ \mathbf{y}_1)$ and $p(\mathbf{w}_2|\mathbf{X}_2,\ \mathbf{y}_2)$ denoted further by $g_1(\mathbf{w})$ and $g_2(\mathbf{w})$.

Goal: to construct a similarity function defined on a pair of distributions $g_1(\mathbf{w})$ and $g_2(\mathbf{w})$. It must satisfy several requirements.

Requirements for similarity function s

Similarity function s must

1. be defined in case distributions' supports are different,
2. satisfy $s(g_1,\, g_2) \leq s(g_1,\, g_1)$,
3. satisfy $s \in [0,\, 1]$,
4. satisfy $s(g_1,\, g_1) = 1$,
5. be close to 1, if $g_2(\mathbf{w})$ is non-informative distribution,
6. be symmetric, i.e. $s(g_1,\, g_2) = s(g_2,\, g_1)$.

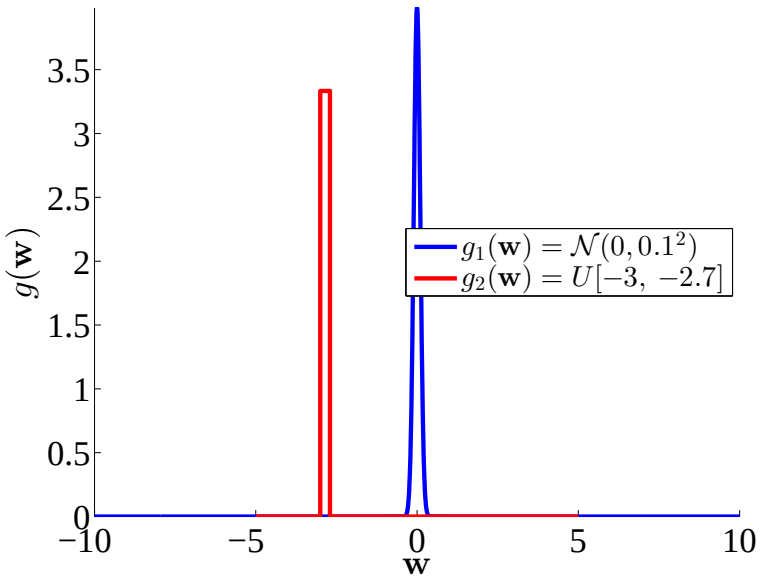
Theorem 2 (Aduenko, 2014)

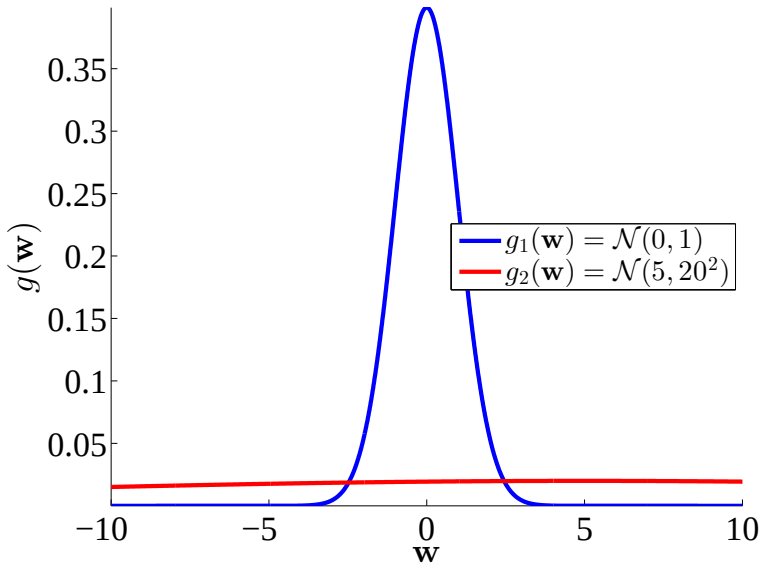
Kullback-Leibler divergence, Jensen-Shannon, Hellinger and Bhattacharyya distances do not meet the requirements for the similarity function.

Illustration of the requirements for similarity function

It is important that the value of s

is close to 1 if $g_2(\mathbf{w})$ is non-informative distribution.





Theorem 2 (Aduenko, 2014)

Kullback-Leibler divergence, Jensen-Shannon, Hellinger and Bhattacharyya distances do not meet the requirements for the similarity function.

Cases 1, 2

Kullback-Leibler divergence and Jensen-Shannon distance

$$D_{KL}(g_1, g_2) = \int g_1(\mathbf{w}) \log \frac{g_1(\mathbf{w})}{g_2(\mathbf{w})} d\mathbf{w}$$

$D_{JS}(g_1, g_2) = \frac{1}{2} D_{KL}(g_1, \frac{1}{2}(g_1 + g_2)) + \frac{1}{2} D_{KL}(g_2, \frac{1}{2}(g_1 + g_2))$ do not meet the requirement for similarity function.

Proof

1. $D_{KL} = \infty$ if $g_1(x) \neq 0$, $g_2(x) = 0$ on a set of positive measure with respect to g_1.
2. $D_{KL}(g_1, g_2) \neq D_{KL}(g_2, g_1)$.
3. $D_{KL} \to \infty$ for a pair of normal distributions $\mathcal{N}(0, 1)$ and $\mathcal{N}(0, \sigma^2)$ when $\sigma^2 \to \infty$.
4. $D_{JS} \not\to 0$ for a pair of normal distributions $\mathcal{N}(0, 1)$ and $\mathcal{N}(0, \sigma^2)$ when $\sigma^2 \to \infty$.

Cases 3, 4

Hellinger and Bhattacharyya distances

$D_H(g_1,\, g_2) = 1 - \int \sqrt{g_1(\mathbf{w})g_2(\mathbf{w})}d\mathbf{w}$,

$D_B(g_1,\, g_2) = -\log \int \sqrt{g_1(\mathbf{w})g_2(\mathbf{w})}d\mathbf{w}$ do not meet the requirement for similarity function.

Proof

Both distances do not have the desired property for non-informative distributions

$$D_H(g_1,\, g_2) \to 1,\; D_B(g_1,\, g_2) \to \infty.$$

Suggested similarity function

The s-score function is suggested to measure the similarity

$$s(g_1,\, g_2) = \frac{\int_{\mathbf{w}} g_1(\mathbf{w}) g_2(\mathbf{w}) d\mathbf{w}}{\max_{\mathbf{b}} \int_{\mathbf{w}} g_1(\mathbf{w}-\mathbf{b}) g_2(\mathbf{w}) d\mathbf{w}}.$$

Theorem 3 (Aduenko, 2014). Suggested similarity function meets all the requirements for the similarity function.

Examples:

$g_1(\mathbf{w})$	$g_2(\mathbf{w})$	$s(g_1,\, g_2)$
$U[0,\, 1]$	$U[0.5,\, 1.5]$	0.5
$U[0,\, 1]$	$U[0.,\, 1.]$	1
$\mathcal{N}(0,\, 1)$	$\mathcal{N}(10, 10^{10})$	1

Expression for $s(g_1,\ g_2)$ for a pair of normal distributions

Theorem 4 (Aduenko, 2014).
Let $g_1 = \mathcal{N}(\mathbf{v}_1,\ \mathbf{\Sigma}_1)$, $g_2 = \mathcal{N}(\mathbf{v}_2,\ \mathbf{\Sigma}_2)$. Then for $s(g_1,\ g_2)$ obtain

$$s(g_1,\ g_2) = \exp\left[\frac{1}{2}(\mathbf{\Sigma}_1^{-1}\mathbf{v}_1 + \mathbf{\Sigma}_2^{-1}\mathbf{v}_2)^{\mathsf{T}}(\mathbf{\Sigma}_1^{-1} + \mathbf{\Sigma}_2^{-1})^{-1}(\mathbf{\Sigma}_1^{-1}\mathbf{v}_1 + \mathbf{\Sigma}_2^{-1}\mathbf{v}_2) - \right.$$
$$\left. - \frac{1}{2}\mathbf{v}_1^{\mathsf{T}}\mathbf{\Sigma}_1^{-1}\mathbf{v}_1 - \frac{1}{2}\mathbf{v}_2^{\mathsf{T}}\mathbf{\Sigma}_2^{-1}\mathbf{v}_2\right].$$

Corollary 1. In case $\mathbf{\Sigma}_2 = \mathbf{0}$ for s-score obtain

$$s(g_1,\ g_2) = \exp\left[-\frac{1}{2}(\mathbf{v}_2 - \mathbf{v}_1)^{\mathsf{T}}\mathbf{\Sigma}_1^{-1}(\mathbf{v}_2 - \mathbf{v}_1)\right].$$

Corollary 2 (s-score expression simplification).
For a pair of normal distribution the expression for s-score is as follows

$$s(g_1,\ g_2) = \exp\left(-\frac{1}{2}(\mathbf{v}_1 - \mathbf{v}_2)^{\mathsf{T}}(\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2)^{-1}(\mathbf{v}_1 - \mathbf{v}_2)\right).$$

Theorem 5 (Aduenko, 2014). Let

- Model f_1 and f_2 coincide, i.e. $\mathbf{w}_1 = \mathbf{w}_2 = \mathbf{w}$.
- Second model parameter values $\mathbf{w}$ are known, i.e. $\mathbf{\Sigma}_2 = \mathbf{O}$.
- Features' values are bounded, i.e. $\exists\, C : \; |x_{1ij}| \leq C$.
- $\mathbf{\Sigma}_1$ is positive definite, and $\lambda_{\max}(\mathbf{\Sigma}_1)/\lambda_{\min}(\mathbf{\Sigma}_1) = O(1)$, $\lambda_{\max}(\mathbf{\Sigma}_1) \to 0$ when $m_1 \to \infty$.

Then the expression for s-score for this two models is

$$s(g_1,\, g_2) = \exp\left[-1/2(\hat{\mathbf{w}}_1 - \mathbf{w})^{\mathsf{T}} \mathbf{\Sigma}_1^{-1} (\hat{\mathbf{w}}_1 - \mathbf{w})\right],$$

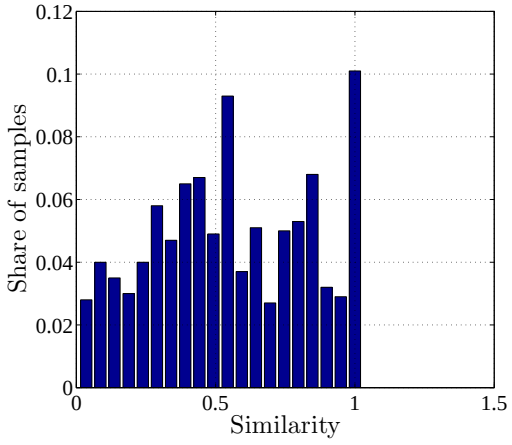
and $s \sim \exp\left[-1/2\xi\right]$, where $\xi \xrightarrow{d} \chi^2(n)$ when $m_1 \to \infty$, n is the number of features.

Corollary 1. For the case when $n = 2$ s-score has asymptotically uniform distribution with $[0,\, 1]$ range.

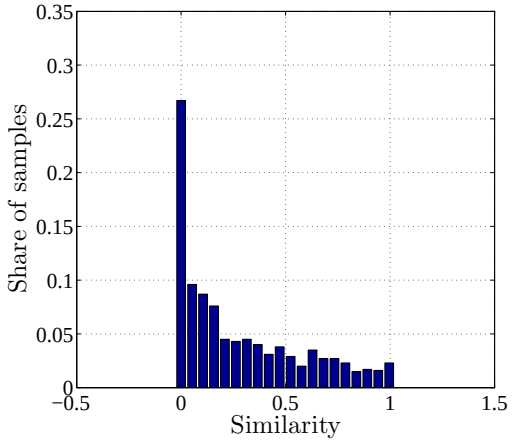
Illustration of s-score application for distinguishing two models, $\rho = 0.9$

Consider two models similar in terms of $\|\mathbf{w}_1 - \mathbf{w}_2\|$, $\|\mathbf{w}_1\| = \|\mathbf{w}_2\| = 1$, $\mathrm{corr}(\mathbf{w}_1, \mathbf{w}_2) = \rho$.

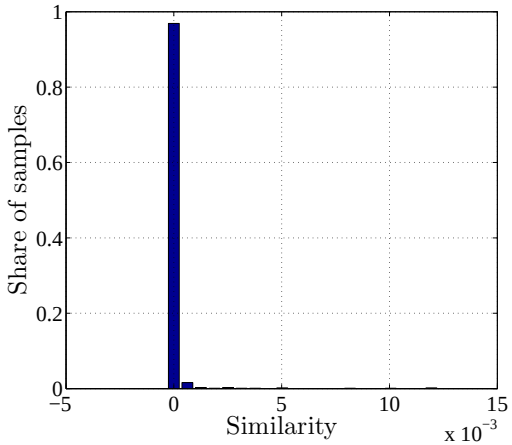
$N_1 = 10000,\ N_2 = 10$



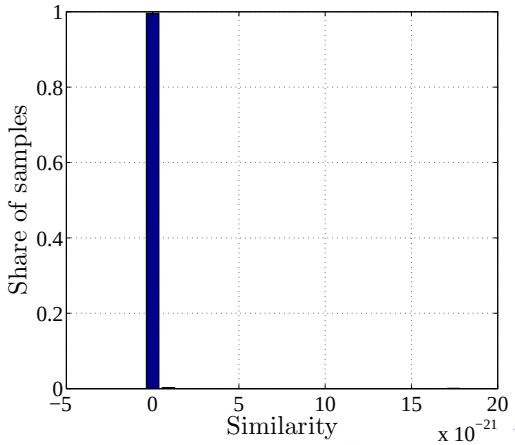
$N_1 = 10000,\ N_2 = 100$



$N_1 = 10000,\ N_2 = 1000$



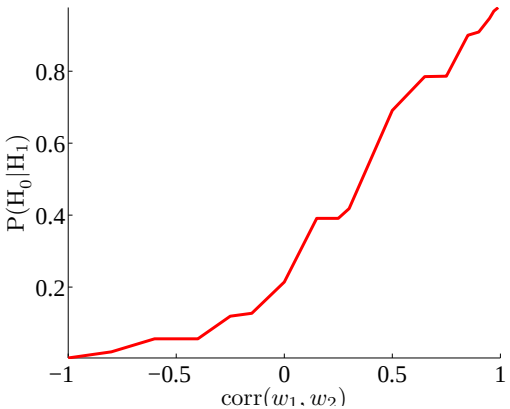
$N_1 = 10000,\ N_2 = 10000$



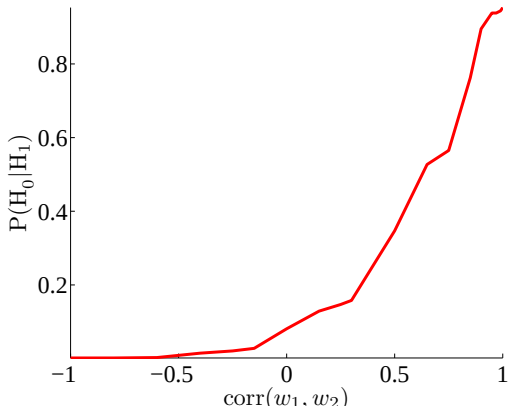
$\mathrm{P}(H_0|H_1)$ dependence on correlation between true models parameters.

Consider two models similar in terms of $\|\mathbf{w}_1 - \mathbf{w}_2\|$, $\|\mathbf{w}_1\| = \|\mathbf{w}_2\| = 1$, $\mathrm{corr}(\mathbf{w}_1, \mathbf{w}_2) = \rho$.

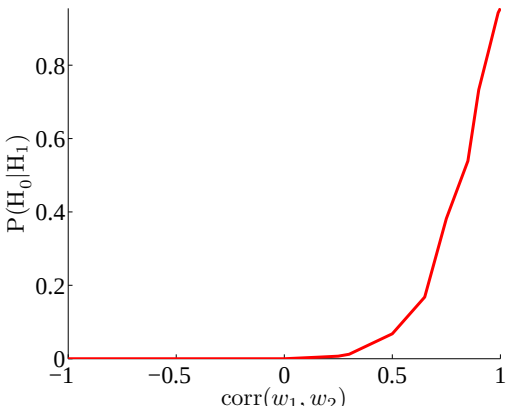
$N_1 = 10000,\ N_2 = 30$



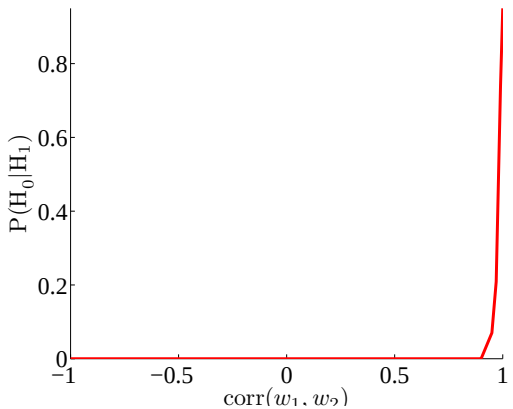
$N_1 = 10000,\ N_2 = 50$



$N_1 = 10000,\ N_2 = 100$



$N_1 = 10000,\ N_2 = 1000$



Generalization of s-score distribution theorem for the case of two finite samples

Theorem 6 (Aduenko, 2016).
If for the models f_1 and f_2

- Models f_1 and f_2 coincide, i.e. $\mathbf{w}_1 = \mathbf{w}_2 = \mathbf{w}$;
- Features' values are bounded, i.e. $\exists\, C : \ |x_{kij}| \leq C,\ k = 1,\ 2$;
- $\hat{\mathbf{\Sigma}}_k$ is positive definite in some neighbourhood of $\mathbf{w}$, and $\lambda_{\max}(\mathbf{\Sigma}_k)/\lambda_{\min}(\mathbf{\Sigma}_k) = O(1),\ \lambda_{\max}(\mathbf{\Sigma}_k) \to 0$ when $m_k \to \infty,\ k = 1,\ 2$;
- $\|\mathbf{\Sigma}_1^{-1}\|\|\mathbf{\Sigma}_2\| \xrightarrow{\mathsf{P}} 0$ when $m_1,\ m_2 \to \infty$;

Then when $m_1,\ m_2 \to \infty$

$$-2\log(s(g_1,\, g_2)) = (\hat{\mathbf{w}}_2 - \hat{\mathbf{w}}_1)^{\mathsf{T}}(\hat{\mathbf{\Sigma}}_1 + \hat{\mathbf{\Sigma}}_2)^{-1}(\hat{\mathbf{w}}_2 - \hat{\mathbf{w}}_1) \xrightarrow{d} \chi^2(n).$$

Theorem 7 (Aduenko, 2014). Let the models defined by $(\mathbf{v}_1, \mathbf{\Sigma}_1)$ and $(\mathbf{v}_2, \mathbf{\Sigma}_2)$ be considered distinguishable if

$$s\left(\mathcal{N}(\mathbf{v}_1, \mathbf{\Sigma}_1), \mathcal{N}(\mathbf{v}_2, \mathbf{\Sigma}_2)\right) \leq C \in (0, 1).$$

Then if the models are distinguishable according to this criterion, then

- models defined by $(\mathbf{v}_1, \mathbf{\Sigma}_1)$ and $(\mathbf{v}_2, \mathbf{O})$ are also distinguishable according to this criterion,
- models defined by $(\mathbf{v}_1, \mathbf{\Sigma}_1)$ and $(\mathbf{v}_2, \lambda\mathbf{\Sigma}_2)$, $\lambda \in [0, 1]$ are also distinguishable according to this criterion.

Theorem 8 (Aduenko, 2014). Consider K models with $\|\mathbf{v}_1\| = \ldots = \|\mathbf{v}_K\| = \lambda_1 > 0$ и $\mathbf{\Sigma}_1 = \ldots = \mathbf{\Sigma}_K = \lambda_2 \mathbf{I}$. Let the following criterion be used to distinguish the models: models $i \neq j$ are distinguishable, if

$$s\left(\mathcal{N}(\mathbf{v}_i,\, \mathbf{\Sigma}_i), \mathcal{N}(\mathbf{v}_j,\, \mathbf{\Sigma}_j)\right) \leq C \in (0,\, 1).$$

Then the maximum number of pairwise distinguishable models in a set is

$$K_{\max} = \left\lfloor \sqrt{\pi} \frac{n\Gamma(\frac{n+1}{2})}{(n-1)\Gamma(\frac{n}{2}+1)} \frac{1}{\int_0^{\theta/2} \sin^{n-2}\varphi d\varphi} \right\rfloor,$$

where $\theta \in [0,\, \pi]$, $\cos\theta = \rho = \max(-1,\, 1 + 2\lambda_2/\lambda_1^2 \ln C)$, n is the features' space dimensionality. Moreover it is feasible to construct K_{min} pairwise distinguishable models, where

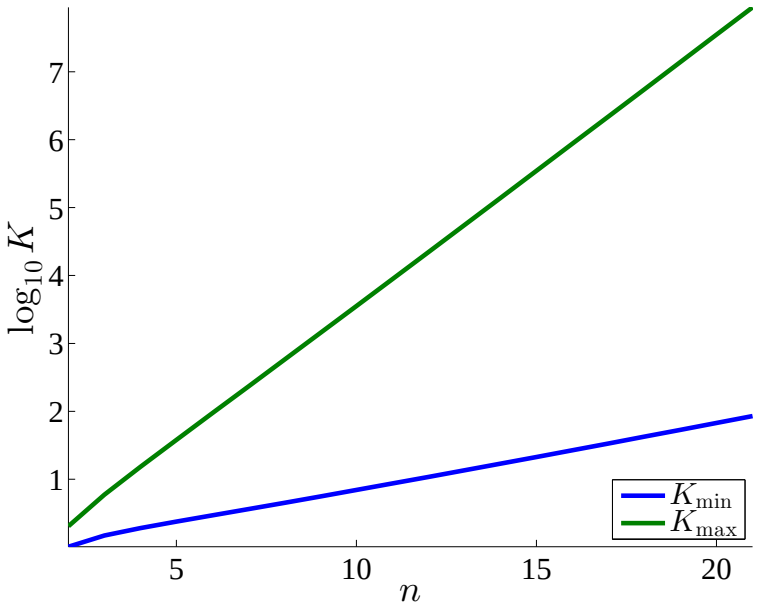
$$K_{\min} = \left\lfloor \sqrt{\pi} \frac{n\Gamma(\frac{n+1}{2})}{(n-1)\Gamma(\frac{n}{2}+1)} \frac{1}{\int_0^{\theta} \sin^{n-2}\varphi d\varphi} \right\rfloor.$$

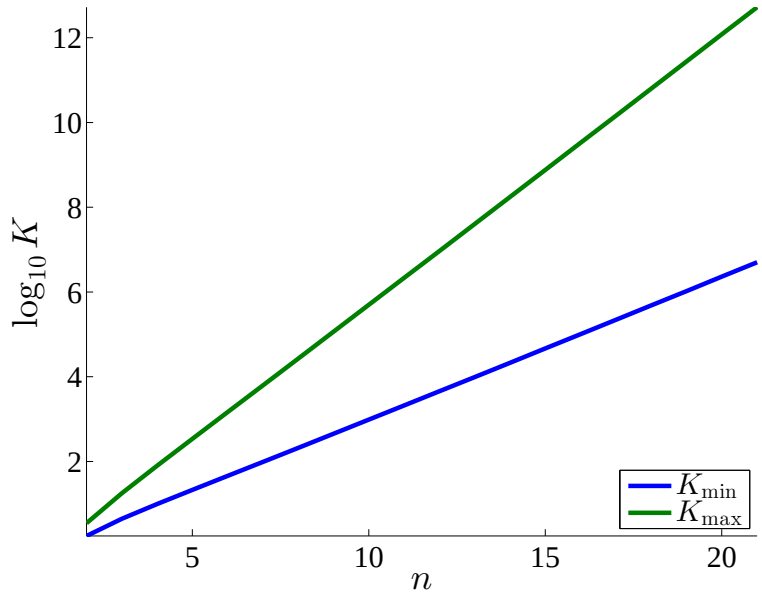
Estimates on the number of models (continued)

Theorem 8 (continued). For C close to 1 obtain

$$K_{\max} \approx \left\lfloor \sqrt{\pi} \frac{n\Gamma(\frac{n+1}{2})}{(n-1)\Gamma(\frac{n}{2}+1)} \frac{2^{\frac{n-1}{2}}}{(1-\rho)^{\frac{n-1}{2}}} \right\rfloor,$$

$$K_{\min} \approx \left\lfloor \sqrt{\pi} \frac{n\Gamma(\frac{n+1}{2})}{(n-1)\Gamma(\frac{n}{2}+1)} \frac{1}{2^{\frac{n-1}{2}}(1-\rho)^{\frac{n-1}{2}}} \right\rfloor.$$





Conclusion

- The theory for selection of $(s,\ \alpha)$ – adequate multimodels containing pairwise distinguishable models was constructed.
- Similarity function s-score which enables checking for similarity between two models was suggested. Asymptotic properties for distributions of $s(g_1,\ g_2)$ and $\log(s(g_1,\ g_2))$ were proved for generalized linear models.
- The method for statistical comparison of models based on introduced similarity function is suggested.
- Using the introduced s-score lower and upper bounds on the number of pairwise distinguishable models were obtained.
- Features' selection algorithm based on maximum evidence estimate of models' parameters' covariance matrix is suggested.
- Structural constraints on covariance matrix in features' selection algorithm were considered. It was shown that the non-diagonal max-evidence estimate of covariance matrix is asymptotically degenerate.