PLISKA

STUDIA MATHEMATICA BULGARICA

ПЛИСКА

БЪЛГАРСКИ МАТЕМАТИЧЕСКИ СТУДИИ

Pliska Stud. Math. Bulgar. **22** (2013), 129–142

PLISKA
STUDIA MATHEMATICA
BULGARICA

MINIMUM DESCRIPTION LENGTH PRINCIPLE IN DISCRIMINATING MARGINAL DISTRIBUTIONS

Bono Nonchev*

Abstract. In this paper the MDL principle is explored in discriminating between a model with normal marginal distributions vs a model with Student-T marginal distributions. The shape complexity of a distribution is defined with insights from the closed-form solution for model complexity for normal distribution. An optimised numerical approach for the Student-T distribution is devised with the aim of extending it to the fat-tailed distributions commonly found in econometric time series.

1. Introduction. The problem we will discuss in this paper is problem of determining the distribution of a sample, in particular determining whether the sample's joint probability distribution is multivariate uncorrelated normal distribution, or it is multivariate uncorrelated Student-T distribution. For the sake of brevity they will be refered below as the normal model and the Student-t model.

In statistical model selection the competing models are always compared along two somewhat contradictory qualities – goodness of fit and generalizability. Goodness of fit is the quality of explanation of past data, which is obviously desired as we cannot expect to explain future data well if we don't make sence of the past. Generalizability is the ability of the model to explain future data.

Unfortunately these two qualities often give contradictory indications when several competing models are compared. A more complex model with many

2010 *Mathematics Subject Classification:* 94A17, 62B10, 62F03

Key words: MDL, Model Selection, Complexity, Distribution Selection

*The research was partially supported by appropriated state funds for research allocated to Sofia University (contract No 125/2012), Bulgaria.

variables may easily fit any data better than a more parsimonious one, but if the fitted parameters try to explain the noise, instead of the underlying relationship (i.e. overfitting), they can easily provide poor explanation of future data, so generalizability would suffer. The problem is exacerbated by the abstractness of future data, and the key is the *model complexity*.

For the last century there have been plenty of research on the problem of statistical model selection, based on a variety of ideas. One is the frequentist approach to assign hypothesis and alternative. The distrinct disadvantage is that sometimes we don't have good reasons to treat the models on different footing. Another way is using Bayesian factors as in [4], but it requires assignment of prior probabilities, which will also be somewhat arbitrary.

The purpose of this paper is to present the solution of the problem of selecting between the two models using the Minimum Description Length (MDL) principle. The stochastic complexity criterion (SC) will be used to decide which of the two models better fit a given sample. Previous work exploring the statistical hypothesis testing and the complexity of the normal distribution in MDL are [8] and [9], or a broader perspective in [10].

This problem is particularly important in econometric time-series modelling. There has been significant evidence that stock price changes do not follow a normal distribution and on shorter time-scales can exhibit high kurtosis. The several solutions for forecasting that incorporate time-dependence (e.g. GARCH) or different distributions (e.g. Student-T, stable, various tempered stable) are usually justified on the basis that they empirically behave "better" in forecasts. However it would be desirable to show that indeed the time-series are best modelled by using those complex models.

Numerically calculating the model complexity from the definition is problematic, as it involves n-dimensional integration, so accept-reject methods will not work. An optimized numerical approach by rewriting the integral as expectation will be presented in order to compute the stochastic complexity of a non-gaussian scale-location family, namely the Student-T distribution.

More details on the choice of models are presented in Section 2.

Regarding the model selection for linear regression most research has been focused on the number of parameters as a proxy for the model complexity in widely used criteria like AIC and BIC. An MDL treatment on linear regression can be found in [9], where the concept of denoising is discussed in details in the case of normal distribution of the residuals. Even though the framework of information criteria for model selection can meaningfully address also the question of the shape of a distribution of the noise, there is comparatively little research

on the subject. The incorporation of 2non-gaussian distributions for residuals in linear regression is important and is left as future work.

One of the advantages of the SC is its computational simplicity - to apply the criterion we only need a table with the values of the stochastic complexity of the different models and the log-likelihood of the data given the competing models. The scheme will be explored in Section 3.

The classical result of the stochastic complexity of the normal distribution will be shown in Section 4. The setting for a general scale-location family will be explored in Section 5, and in particular for the scale-location family of a Student-T distribution with known degrees of freedom in Section 6.

The numerical results will be shown in Section 7 along with a summary of the results and future work in Section 8.

2. Models of interest. The first model concidered for the sample is a multivariate normal distribution with p.d.f.:

$$f_N\left(\mathbf{x}^n|\mu,\sigma\right) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i-\mu\right)\right)$$

The second model considered for the sample is a multivariate uncorrelated Student-T distribution with fixed degrees of freedom ν_0, having p.d.f.

$$f_T\left(\mathbf{x}^n|\mu,\sigma\right) = \frac{\Gamma\left(\frac{n+\nu_0}{2}\right)}{\Gamma\left(\frac{\nu_0}{2}\right)\left(\nu_0\pi\sigma^2\right)^{\frac{n}{2}}}\left(1+\frac{1}{\nu_0}\frac{1}{\sigma^2}\sum_{i=1}^{n}\left(x_i-\mu\right)\right)^{-\frac{n+\nu_0}{2}}$$

In this paper the Student-T distribution is used as a similar to the normal distribution, although with heavier tails. It is not, as in frequentist hypothesis testing, used as "alternative", as we do not consider the normal to be the default choice of model. Nor do we assign prior probabilities of the models as in a Bayesian setting.

The individual observations in the normal model are independent, in addition to identically distributed. In the Student-T model they are only uncorrelated, so that model cannot be used to model an IID sample (see[6], chapter 1). Other than that, the multivariate Student-T distribution has the advantage that the maximum likelihood estimator for the parameters μ and σ are, as in the case of the normal distribution, the sample mean and variance.

In addition the MLE of a linear regression is the ordinary least squares method (see [6], chapter 11). This means that the discussed Student-T model can easily replace the gaussian distribution as the noise distribution in a linear regression. Exploring that possibility is left as future work.

3. The Minimum Description Length principle. The Minimum Description Length principle (MDL) in its most basic form states that the more we can compress the data generated by a process, the more we know about it. A classic example problem used in the inspirational paper of Kolmogorov [5] is that if you are charged with the task of transmitting three sequences of a million symbols, each 0 or 1, like the following:

- 01010101010101010101010101010101010...
- 11011001111111011111110110011111111111...
- 10101010001110100011101000111010111110...

you can certainly do better than transmitting the whole sequence bit by bit, if you explore the regularities in the data.

The first sequence is just 01 repeated, so sending this instruction instead is quite a lot faster. The second has about 9 ones for each zero, so we can try to encode long strings of ones with shorter codes than strings of zeroes, and transmit shorter codes than the trivial. For the third there is not much we can do, as it is random with equal probabilities of 0 and 1.

In each of those cases knowledge of the patterns in the data allows us to compress it, which is why the MDL principle equates knowledge with compression.

The original idea of Kolmogorov defines a code and compression as the shortest algorithm that output the given sequence. Regrettably there is no algorithm that can find such codes, so in the MDL principle the compression is done using only codes that correspond to probability distributions.

The compression is defined in the sence of Shannon's information theory and the goal is to find the shortest possible codelength that can be used to compress the data. If we have found such a way, we can compute the code for our data for our two models, and declare that the one that compresses the data better is the proper model for that data.

The correspondence of codes with probability distributions is established using Kraft's inequality:

$$\sum_{x \in \Omega} 2^{l(x)} \leq 1$$

as it roughly corresponds to

$$\sum_{x \in \Omega} p\left(x\right) = 1$$

when $p\left(x\right)$ is a probability distribution. The optimal code (in terms of expected code length) for will have codeword with length

$$l\left(x\right) = \log_2 p\left(x\right)$$

so for each p.d.f. there is an optimal code, and for each Kraft-tight code there is a probability distribution for which the code is optimal.

The coding aspect has many details such as how to deal with continuous distributions, which are not essential for the application of MDL in our setting. In fact we are only interested in the codelength, not the actual coding, as the goal is to find which model best fits the data.

There are many obstacles to the application of the above idea (called Crude MDL in the literature), not the least of which is that there is no way to create a shortest code. A major milestone in the application of the MDL principle has been the discovery of "universal models", which clarify the idea of optimal codes. There is extensive literature on the subject and the reader is referred to [2] or [3] for a thorough review. We will present here only the necessary classical results.

Universal model with respect to a class of models is a single probability distribution, whose corresponding code compress a given data "almost as well" as the code for the best distribution in that class. In our setting we are searching for one universal model for the normal model and another universal model for the Student-T model.

If we have a scale-location family of distributions with p.d.f. $f(x|\mu,\sigma)$, we could estimate μ and σ using the data, and then try to encode the data with a code corresponding to $f(x|\hat{\mu}(x),\hat{\sigma}(x))$. There is no such code, because $f(x|\hat{\mu}(x),\hat{\sigma}(x))$ is not a p.d.f. for x, as we have used the data to estimate the parameters. However, we can try to compress the data using a code corresponding to

$$f_{NML}(x) = \frac{f(x|\hat{\mu}(x),\hat{\sigma}(x))}{\int f(y|\hat{\mu}(y),\hat{\sigma}(y))\,dy} \tag{1}$$

which is a distribution when

$$COMP_n(f) = \int f(y|\hat{\mu}(y),\hat{\sigma}(y))\,dy < \infty$$

The last term is called the stochastic complexity of a family of distributions. In terms of the log-likelihood equation 1 becomes

$$\ln f_{NML}(x) = -\ln f(x|\hat{\mu}(x),\hat{\sigma}(x)) + \ln COMP_n(f)$$

The $f_{NML}(x)$ defined above is the so-called Normalized Maximum Likelihood model, first introduced in [11] and subsequently thoroughly explored for various problems. It is a universal model, and the basis for the stochastic complexity (SC) criterion for model selection: encode the sample using the NML distribution for one class of distributions, encode with the other and compare the log-likelihood.

The purpose of this paper is to calculate the stochastic complexity $COMP_n(f)$ of the Student-T model and compare it to the classical result for the normal model, as well as to provide insight on its computation for other scale-location families. The problem of infinite complexity is also adressed, and alternative quantity $DC_n(f)$ is proposed, which is finite for the two models discussed in this paper.

The role of complexity can be seen in an example in [7]. Two regression models are considered to explain the relation between two random variables X and Y:

$$Y = aX^b + \epsilon \text{ vs } Y = a\ln(X+b) + \epsilon \tag{2}$$

Although both models have two parameters each, the first one is much more complex in the sense that it fits better arbitrary data. With a small sample size of 4 and artificially generated data, the authors show that the first model fits the data better in terms of log-likelihood in 67% of the cases, even though the second model was used to generate the actual data.

4. Stochastic complexity of the normal distribution. In this section the classical result for the complexity of the normal model will be presented as in [9]. The known ways to deal with infinite model complexity is also addressed below, with justification for our choice.

Let us have an i.i.d. sample of normally distributed random variables from $N\left(\mu,\sigma^2\right)$. The joint distribution of the sample is

$$\begin{aligned} f\left(\mathbf{x}^n|\mu,\sigma\right) &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum\left(x_i-\mu\right)^2\right) \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{ns^2}{2\sigma^2}\right)\exp\left(-\frac{\left(\overline{\mathbf{x}}-\mu\right)^2}{2\left(\frac{\sigma}{\sqrt{n}}\right)^2}\right) \end{aligned}$$

This shows that $\overline{\mathbf{x}}$ and s^2 are sufficient statistics, and from the Fisher-Neyman factorization theorem

$$f\left(\mathbf{x}^n|\mu,\sigma\right) = l\left(\mathbf{x}^n|\overline{\mathbf{x}},s^2\right)h\left(\overline{\mathbf{x}},s^2|\mu,\sigma\right)$$

where $h\left(\overline{\mathbf{x}},s^2|\mu,\sigma\right)$is the p.d.f. of distribution of $\overline{\mathbf{x}}$ and s^2. Using $\overline{x} \in \mathbb{N}\left(\mu,\frac{\sigma^2}{n}\right)$ and $\frac{ns^2}{\sigma^2} \in \chi^2_{n-1}$ and applying Cochran's theorem we arrive at the

joint p.d.f. of the sufficient statitstics:

$$h\left(\overline{\mathbf{x}}, s^2|\mu,\sigma\right) = \frac{1}{2^{\frac{n-1}{2}}\Gamma\left(\frac{n-1}{2}\right)\frac{\sigma^2}{n}}\left(\frac{ns^2}{\sigma^2}\right)^{\frac{n-1}{2}-1} e^{-\frac{ns^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi\left(\frac{\sigma}{\sqrt{n}}\right)^2}} \exp\left[-\frac{\left(\overline{\mathbf{x}}-\mu\right)^2}{2\left(\frac{\sigma}{\sqrt{n}}\right)^2}\right]$$

Using the results from [1], chapter III, section F, for the case where there are sufficient statistics for the parameters we have

$$COMP_n\left(\mathcal{M}\right) = \int_{x^n\in\chi^n} f\left(\mathbf{x}^n|\hat{\theta}\left(x^n\right)\right) dx^n = \int h\left(\hat{\theta}|\hat{\theta}=s\right) ds$$

Unfortunately the above integral is infinite in our model. There are several approaches to deal with this, most notable of which are the *renormalization* by complexity conditional on the data space as presented in [9] and the usage of complexity conditional on the parameter space as in [12]. Both approaches have their merits. However, to compare between models we have to account for the various parametrisations, thus limiting $\mathbf{x}^n$ would allows for more flexibility.

We will use the conditional complexity of a model defined as

$$(3)\quad COMP_n\left(\mathcal{M}|\overline{\mathbf{x}}\in[-R;R]\cap s^2\in[D,\infty)\right) = \\ = \int_D^\infty\int_{-R}^R h\left(\overline{\mathbf{x}}, s^2|\mu=\overline{x},\sigma=s\right) dm\, ds^2$$

Substituting $\mu=\overline{x}$ and $\sigma=s$ we get

$$h\left(\overline{\mathbf{x}}, s^2|\mu=\overline{x},\sigma=s\right) = \frac{1}{2^{\frac{n-1}{2}}\Gamma\left(\frac{n-1}{2}\right)\frac{s^2}{n}}\left(n\right)^{\frac{n-3}{2}} e^{-\frac{n}{2}} \frac{\sqrt{n}}{\sqrt{2\pi s^2}}$$

Evaluating the above integral we arrive at an analytic formula for the complexity:

$$\begin{aligned} COMP_n\left(\mathcal{M}|\mathcal{A}\right) &= \int_D^\infty\int_{-R}^R \frac{1}{2^{\frac{n-1}{2}}\Gamma\left(\frac{n-1}{2}\right)\frac{s^2}{n}}\left(n\right)^{\frac{n-3}{2}} e^{-\frac{n}{2}} \frac{\sqrt{n}}{\sqrt{2\pi s^2}} dm\, ds^2 \\ &= 2R\frac{n}{2^{\frac{n-1}{2}}\Gamma\left(\frac{n-1}{2}\right)}\left(n\right)^{\frac{n-3}{2}} e^{-\frac{n}{2}} \frac{\sqrt{n}}{\sqrt{2\pi}}\int_D^\infty (s^2)^{-\frac{3}{2}}\, ds^2 \\ &= 2RD^{-1}\frac{2n^{\frac{n}{2}}e^{-\frac{n}{2}}}{\sqrt{2\pi}2^{\frac{n-1}{2}}\Gamma\left(\frac{n-1}{2}\right)} \end{aligned}$$

The term of interest is $2RD^{-1}$. Since it is a multiplicative constant we can isolate it in the log-complexity:

$$\ln COMP_n\left(\mathcal{M}|\mathcal{A}\right) = \ln 2 + \frac{n}{2}\left(\ln\frac{n}{2} - 1\right) - \frac{\ln\pi}{2} - \ln\Gamma\left(\frac{n-1}{2}\right) + \ln 2RD^{-1}$$

In other papers like [9] and [12] the codes are extended to encode the limits R and D, but this introduces arbitrariness as some parameter values are treated as more likely. Our approach is to use complexity conditional on the data space, but without re-normalization, as it is not needed when comparing the two chosen models, and consequently no arbitrariness arises.

Namely the last term $\ln 2RD^{-1}$ does not depend on the sample size, and captures all of the dependence on the boundaries of integration. The rest of the terms capture the model complexity. We justify SC criterion usage across model classes such as different location and scale families by subtracting the common term $\ln 2RD^{-1}$.

5. Stochastic complexity of location-scale families. The idea explored in this paper is to separate the terms and treat them differently. If we cancel the common terms between two distributions, we can compare their complexity for fixed R and D. Moreover, the same relationship will hold for all R and D. To do that we need a more general expression for the complexity of a scale-location family.

We will work with the entire sample $\mathbf{x}^n$, so the definition is given in the multivariate case:

Definition 1. *A scale-location family is a family of distributions having p.d.f.* $f\left(\mathbf{x}^n|\mu,\sigma\right)$ *for which a function* $g(\mathbf{y}^n)$ *exists satisfying*

$$f\left(\mathbf{x}^n|\mu,\sigma\right) = \sigma^{-n} g\left(\frac{\mathbf{x}^n - \mu}{\sigma}\right)$$

The standard member of the distribution, i.e. the one having $\mu = 0$ and $\sigma = 1$, will feature more prominently in the analysis, and by the definition its p.d.f. is $g(\mathbf{x}^n)$.

The following lemma shows an important characterization of the scale-location families and their corresponding maximum likelihood estimators, a well-known fact for which the proof is also provided.

Lemma 1. *If* $\hat{\mu}\left(\mathbf{x}^n\right)$ *and* $\hat{\sigma}\left(\mathbf{x}^n\right)$ *are MLE for a scale-location family (i.e. they exist and are unique), then* $\hat{\mu}\left(\sigma\mathbf{y}^n + \mu\right) = \sigma\hat{\mu}\left(\mathbf{y}^n\right) + \mu$ *and* $\hat{\sigma}\left(\sigma\mathbf{y}^n + \mu\right) = \sigma\hat{\mu}\left(\mathbf{y}^n\right)$.

P r o o f. By the definition of a MLE and the properties of the p.d.f. of a scale-location family we get

$$\left(\hat{\sigma}\left(\mathbf{x}^n\right), \hat{\mu}\left(\mathbf{x}^n\right)\right) = \operatorname{argmax}_{\mu,\sigma} f\left(\mathbf{x}^n|\mu,\sigma\right) = \operatorname{argmax}_{\mu,\sigma} \sigma^{-n} g\left(\frac{\mathbf{x}^n - \mu}{\sigma}\right)$$

$$\begin{aligned}\left(\hat{\sigma}\left(a\mathbf{x}^n + b\right), \hat{\mu}\left(a\mathbf{x}^n + b\right)\right) &= \operatorname{argmax}_{\mu,\sigma} f\left(a\mathbf{x}^n + b|\mu,\sigma\right) \\ &= \operatorname{argmax}_{\mu,\sigma} \sigma^{-n} g\left(\frac{a\mathbf{x}^n + b - \mu}{\sigma}\right) \\ &= \operatorname{argmax}_{\mu,\sigma} \left(\frac{\sigma}{a}\right)^{-n} g\left(\frac{\mathbf{x}^n + (b-\mu)/a}{\sigma/a}\right)\end{aligned}$$

Combined with the definition we get the following equalities

$$\begin{aligned}\hat{\sigma}\left(a\mathbf{x}^n + b\right)/a &= \hat{\sigma}\left(\mathbf{x}^n\right) \\ \left(\hat{\mu}\left(a\mathbf{x}^n + b\right) - \mu\right)/a &= \hat{\mu}\left(\mathbf{x}^n\right)\end{aligned}$$

completing the proof. □

Definition 2. *The distribution complexity is defined as*

$$DC_n\left(\mathcal{M}\right) = \mathbb{E}_{\mathbf{Y}^n}\left[\delta\left(\hat{\mu}\left(\mathbf{Y}^n\right)\left(1 - \hat{\sigma}\left(\mathbf{Y}^n\right)\right)\right)\right]$$

Now the conditional complexity will be defined, as in 3, as conditional on the set

$$\mathcal{A} = \left\{\hat{\mu}\left(\mathbf{x}^n\right) \in [-R; R] \cap \hat{\sigma}\left(\mathbf{x}^n\right) \in [D, \infty)\right\}$$

$$COMP_n\left(\mathcal{M}|\mathcal{A}\right) = \int_{\mathbf{x}^n \in \mathcal{A}} f\left(\mathbf{x}^n|\mu = \hat{\mu}\left(\mathbf{x}^n\right), \sigma = \hat{\sigma}\left(\mathbf{x}^n\right)\right) d\mathbf{x}^n \tag{4}$$

Note that the distribution complexity does not depend on the restriction or parameters. This integral is also hard to solve for large n using numerical integration, because it is n-dimensional, so a better approach is needed.

We will use the Dirac delta δ for brevity of notation to prove the following

Theorem 1. *For a scale-location family the conditional complexity can be decomposed as*

$$COMP_n\left(\mathcal{M}|\mathcal{A}\right) = 2RD^{-1} \times DC_n\left(\mathcal{M}\right)$$

P r o o f. The first step is to rewrite the integral using the standard density $g(\mathbf{x}^n)$:

$$\begin{aligned}
&COMP_n\left(\mathcal{M}|\mathcal{A}\right)\\
&=\int_{\mathbf{x}^n\in\mathcal{A}} f\left(\mathbf{x}^n|\mu=\hat{\mu}\left(\mathbf{x}^n\right),\sigma=\hat{\sigma}\left(\mathbf{x}^n\right)\right)d\mathbf{x}^n\\
&=\int_{\mathbf{x}^n\in\mathcal{A}}\int\int \delta\left(\mu-\hat{\mu}\left(\mathbf{x}^n\right)\right)\delta\left(\sigma-\hat{\sigma}\left(\mathbf{x}^n\right)\right) f\left(\mathbf{x}^n|\mu,\sigma\right)d\mu\,d\sigma\,d\mathbf{x}^n\\
&=\int_{\mathbf{x}^n\in\mathcal{A}}\int\int \delta\left(\mu-\hat{\mu}\left(\mathbf{x}^n\right)\right)\delta\left(\sigma-\hat{\sigma}\left(\mathbf{x}^n\right)\right)\sigma^{-n}g\left(\frac{\mathbf{x}^n-\mu}{\sigma}\right)d\mu\,d\sigma\,d\mathbf{x}^n
\end{aligned}$$

Now let us turn our attention to $\mathcal{A}$, so that we can move the boundaries of the integral from $\hat{\mu}$ and $\hat{\sigma}$ to μ and σ. Since $\mathcal{A}$ is defined as those $\mathbf{x}^n$ for which $\hat{\mu}\left(\mathbf{x}^n\right)\in[-R;R]$ and $\hat{\sigma}\left(\mathbf{x}^n\right)\in[D,\infty)$, and the delta function has support only $\{0\}$, we can change the inner integral limits to $[-R;R]$and $[D;\infty)$:

$$\begin{aligned}
&COMP_n\left(\mathcal{M}|\mathcal{A}\right)\\
&=\int_{\mathbf{x}^n\in\mathbb{R}^n}\int_{-R}^{R}\int_{D}^{\infty}\delta\left(\mu-\hat{\mu}\left(\mathbf{x}^n\right)\right)\delta\left(\sigma-\hat{\sigma}\left(\mathbf{x}^n\right)\right)\sigma^{-n}g\left(\frac{\mathbf{x}^n-\mu}{\sigma}\right)d\mu\,d\sigma\,d\mathbf{x}^n
\end{aligned}\tag{5}$$

We make the substitution $\mathbf{y}^n=\dfrac{\mathbf{x}^n-\mu}{\sigma}$, for which $|J|=\sigma^n$. Since $\hat{\mu}\left(\mathbf{x}^n\right)$ and $\hat{\sigma}\left(\mathbf{x}^n\right)$ are MLE for a scale-location family we can use the Lemma 1 to simplify 5. Combining that with the fact that δ is homogeneous of degree -1, we can isolate the dependence on the boundaries and σ and μ as follows:

$$\begin{aligned}
&COMP_n\left(\mathcal{M}|\mathcal{A}\right)\\
&=\int_{\mathbf{y}^n\in\mathbb{R}^n}\int_{-R}^{R}\int_{D}^{\infty}\delta\left(\mu-\sigma\hat{\mu}\left(\mathbf{y}^n\right)-\mu\right)\delta\left(\sigma-\sigma\hat{\sigma}\left(\mathbf{y}^n\right)\right)\sigma^{-n}g\left(\mathbf{y}^n\right)\sigma^n d\mu\,d\sigma\,d\mathbf{y}^n\\
&=\int_{\mathbf{y}^n\in\mathbb{R}^n}\int_{-R}^{R}\int_{D}^{\infty}\sigma^{-2}\delta\left(\hat{\mu}\left(\mathbf{y}^n\right)\right)\delta\left(1-\hat{\sigma}\left(\mathbf{y}^n\right)\right)g\left(\mathbf{y}^n\right)d\mu\,d\sigma\,d\mathbf{y}^n\\
&=\left[\int_{-R}^{R}\int_{D}^{\infty}\sigma^{-2}d\mu\,d\sigma\right]\int_{\mathbf{y}^n\in\mathbb{R}^n}\delta\left(\hat{\mu}\left(\mathbf{y}^n\right)\right)\delta\left(1-\hat{\sigma}\left(\mathbf{y}^n\right)\right)g\left(\mathbf{y}^n\right)\,d\mathbf{y}^n\\
&=2RD^{-1}\int_{\mathbf{y}^n\in\mathbb{R}^n}\delta\left(\hat{\mu}\left(\mathbf{y}^n\right)\right)\delta\left(1-\hat{\sigma}\left(\mathbf{y}^n\right)\right)g\left(\mathbf{y}^n\right)\,d\mathbf{y}^n
\end{aligned}$$

So we have arrived at the quantity of interest

$$\frac{COMP_n\left(\mathcal{M}|\mathcal{A}\right)}{2RD^{-1}}=\int_{\mathbf{y}^n\in\mathbb{R}^n}\delta\left(\hat{\mu}\left(\mathbf{y}^n\right)\right)\delta\left(1-\hat{\sigma}\left(\mathbf{y}^n\right)\right)g\left(\mathbf{y}^n\right)\,d\mathbf{y}^n$$

$$
\begin{aligned}
&= \int_{\mathbf{y}^n \in \mathbb{R}^n} \delta\left(\hat{\mu}\left(\mathbf{y}^n\right)\right) \delta\left(1-\hat{\sigma}\left(\mathbf{y}^n\right)\right)\, dG\left(\mathbf{y}^n\right) \\
&= \mathbb{E}_{\mathbf{Y}^n}\left[\delta\left(\hat{\mu}\left(\mathbf{Y}^n\right)\left(1-\hat{\sigma}\left(\mathbf{Y}^n\right)\right)\right)\right] \qquad (6) \\
&= DC_n\left(\mathcal{M}\right)
\end{aligned}
$$

□

The derivation above is the reason we call $DC_n\left(\mathcal{M}\right)$ the distribution complexity, as it depends only on the shape of the marginal distribution and the dependence structure.

The SC criterion in the special case of selecting between our two models is rewriten as

$$\tilde{L}(x) = \ln f_{NML}(x) - \ln 2RD^{-1} = -\ln f\left(x|\hat{\mu}(x), \hat{\sigma}(x)\right) + \ln DC_n\left(\mathcal{M}\right)$$

and comparison is done using the adjusted codelength $\tilde{L}(x)$.

6. Stochastic complexity of Student-T distribution. The representation 6 gives us an optimized way to calculate the distributional complexity numerically via Monte Carlo simulations using a sample from the distribution.

For both models considered in this paper the MLE estimators of the parameters are the sample mean and variance:

$$\hat{\mu}\left(\mathbf{y}^n\right) = \frac{1}{n}\sum_i y_i \text{ and } \hat{\sigma}\left(\mathbf{y}^n\right) = \frac{1}{n}\sum_i \left(y_i - \hat{\mu}\left(\mathbf{y}^n\right)\right)^2$$

In this case we can try to split the integral as follows:

$$
\begin{aligned}
DC_n\left(\mathcal{M}\right) &= \int_{\mathbf{y}^{n-2} \in \mathbb{R}^{n-2}} I(\mathbf{y}^{n-2})\, dG\left(\mathbf{y}^{n-2}\right) \\
I(\mathbf{y}^{n-2}) &= \int_{(y_{n-1}, y_n) \in \mathbb{R}^2} \delta\left(\hat{\mu}\left(\mathbf{y}^n\right)\right) \delta\left(1-\hat{\sigma}\left(\mathbf{y}^n\right)\right) g\left(y_{n-1}, y_n|\mathbf{y}^n\right) dy_{n-1} dy_n \qquad (7)
\end{aligned}
$$

$I(\mathbf{y}^{n-2})$ can be calculated via change in varibles: $y_{n-1}, y_n \rightarrow \hat{\mu}, \hat{\sigma}$. The actual range of integration in the first integral is

$$B = \left\{\mathbf{y}^{n-2} : \exists y_{n-1}, y_n \text{ for which } \hat{\mu}\left(\mathbf{y}^n\right) = 0, \hat{\sigma}\left(\mathbf{y}^n\right) = 1\right\}$$

because if $\mathbf{y}^{n-2} \notin B$, then $I(\mathbf{y}^{n-2}) = 0$.

In addition there is a symmetry in the equations $\hat{\mu}\left(\mathbf{y}^n\right) = 0, \hat{\sigma}\left(\mathbf{y}^n\right) = 1$ - if one solution is $\left(y^*_{n-1}, y^*_n\right)$, then $\left(y^*_n, y^*_{n-1}\right)$ is the other solution. Then from the property of the delta function for composition

$$\delta\left(g(x)\right) = \sum_i \frac{\delta\left(x - x_i\right)}{\left|\frac{\partial g}{\partial x}\left(x_i\right)\right|}$$

over all i solutions of $g(x) = 0$ and a bit of calculus we can simplify 7 as

$$I(\mathbf{y}^{n-2}) \quad = \quad \begin{cases} 2n^2 g\left(y^*_{n-1}, y^*_n | \mathbf{y}^{n-2}\right) D & \text{if } \mathbf{y}^{n-2} \in B \\ 0 & \text{if } \mathbf{y}^{n-2} \notin B \end{cases}$$

where $D = \sqrt{2n - 2\left(\sum_{i=1}^{n-2} y_i^2\right) - \left(\sum_{i=1}^{n-2} y_i\right)^2}$.

7. Numerical results. To compute $DC_n(\mathcal{M})$ we just simulate a sample $\left\{\mathbf{y}_i^{n-2}\right\}_{i=1}^{T}$ from an $n-2$ dimensional Student-T random variable and average $I(\mathbf{y}^{n-2})$ to obtain

$$DC_n(\mathcal{M}) \approx \frac{1}{T} \sum_{i=1}^{T} I(\mathbf{y}_i^{n-2})$$

This approach is much better than any accept-reject method applied on equation (4) for large n, as less simulations are needed to achieve the same number of non-zero summands and consequently achieve higher accuracy.

The numerical computation has been done with MATLAB on a standard quad core Intel processor. The visible noise is due to the relatively low number of simulations (100000). The results are summarized on Figure 1.

We can see that the model complexity of a the Student-T distribution is in fact lower than the complexity of the normal distribution. To justify the usage of T-distribution for a sample we may have a smaller likelihood for the parameters at the MLE than for the normal distribution, or in other words the T-distribution (with fixed d.f.) is actually *less complex* than the normal distribution.

In terms of the model selection this means that if we have a sample for which the log-likelihood in the normal model is equal to the log-likelihood in the Student-T model, the MDL principle suggests that we should choose the Student-T model as more parsimonious.

Conversely, a sample that fits equally well the two models will have higher log-likelihood for the normal model, as the normal model is more complex and has a higher chance of fitting noise in the sample.

8. Conclusions. The numerical results show that with the approach outlined in this paper the numerical integration does enable us to compute a table of the stochastic complexity values of Student-T distribution for a fixed degrees of freedom and size of sample up to 100 in a feasible amount of time.

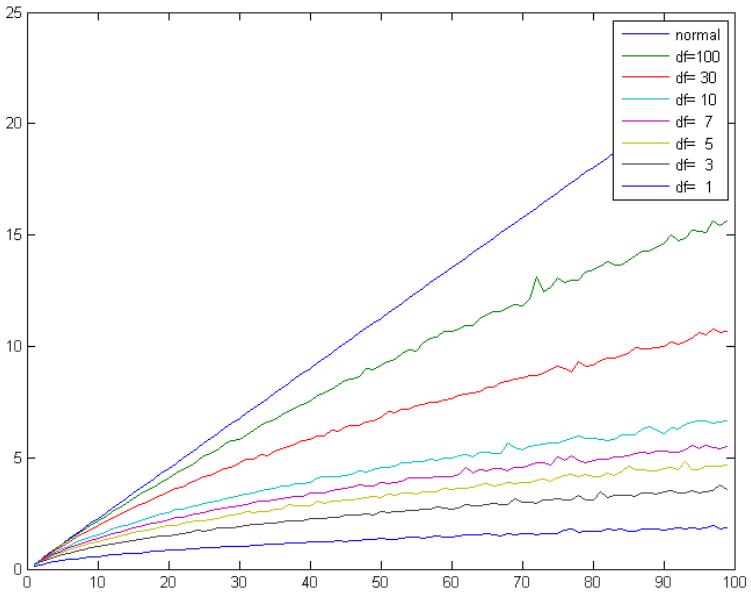

Fig. 1. Plot of the distribution complexity DC_n on the y-axis for normal vs Student-T with various degrees of freedom, relative to the size of the n on the x-axis

The future work will consist of calculating the stochastic complexity of a Student-T model for an independently distributed sample, enabling varying degrees of freedom instead of fixed, as well as extending the calculation to linear regression and econometric time-series models of interest.

Indeed the question of the complexity of the time-series models is also interesting, since in the practical implementation a Stable distributions and GARCH effects are often called complex models, and their dismissal is usually on the ground of parsimony. A thorough investigation and quantification of their complexity can clarify the extend to which this is justified.

Acknowledgements. I also wish to express my gratitude to the rigorous comments by the reviewer.

REFERENCES

[1] A. Barron, J. Rissanen, B. Yu. The Minimum Description Length Principle in Coding and Modeling. *IEEE Transactions on Information Theory* **44** (1998), No 6, 2743–2760.

[2] P. Grunwald. A Tutorial Introduction to the Minimum Description Length Principle. arXiv preprint math/0406077, 2004.

[3] P. Grünwald, J. I. Myung, M. Pitt. Advances in Minimum Description Length: Theory and Applications. The MIT Press, 2005.

[4] R. Kass, A. Raftery. Bayes factors. *Journal of the American Statistical Association* **90** (1995), No 430, 773–795.

[5] A. N. Kolmogorov. On Tables of Random Numbers. *Sankhya: The Indian Journal of Statistics, Series A* **25** (1963), No 4, 369–376.

[6] S. Kotz, S. Nadarajah. Multivariate t Distributions and Their Applications. Cambridge, New York, Madrid, Cambridge University Press, 2004.

[7] I. J. Myung, V. Balasubramanian, M. A. Pitt. Counting Probability Distributions: Differential Geometry and Model Selection. *Proceedings of the National Academy of Sciences of the United States of America* **97** (2000), No 21, 11170–11175.

[8] J. Rissanen. Hypothesis Selection and Testing by the MDL Principle. *The Computer Journal* **42** (1999), No 4, 260–269.

[9] J. Rissanen. MDL Denoising. *IEEE Transactions on Information Theory* **46** (2000), No 7, 2537–2543.

[10] J. Rissanen. Information and Complexity in Statistical Modeling (Information Science and Statistics). Springer, 2007.

[11] Y. Shtarkov. Universal Sequential Coding of Single Messages. *Problems of Information Transmission* **23** (1987), No 3, 175–186.

[12] R. Stine, D. Foster. The Competitive Complexity Ratio. In: Proceedings of the 2001 Conference on Information Sciences and Systems, WP8, 2001, 1–6.

Bono Nonchev
Faculty of Mathematics and Informatics,
Sofia University "St. Kliment Ohridski"
Sofia, Bulgaria
e-mail: bono.nonchev@gmail.com