University of Tartu
Faculty of Science and Technology
Institute of Mathematics and Statistics

Kristi Ernits

Finite Mixtures with Application on Estonian Meteorological Data

Mathematics and Statistics Curriculum
(Mathematical Statistics)
Master's Thesis (30 ECTS)

Supervisor: Annika Krutto, PhD

Tartu 2019

Finite Mixtures with Application on Estonian Meteorological Data
Master's thesis
Kristi Ernits

Abstract. Finite mixtures have found application in various areas including meteorology. The purpose of this master's thesis is to study finite mixtures with focus on parameter estimation in the concept of the EM algorithm. Overview of finite mixtures and EM algorithm estimators for the parameters of specific finite mixtures are given. In addition, finite mixtures are used to model Estonian meteorological data. Various finite mixtures are fitted on Estonian daily wind speed data and normal mixtures on daily temperature data.

CERCS research specialisation: P160 Statistics, operation research, programming, actuarial mathematics

Keywords: probability distributions, mixture distributions, EM algorithm, wind speed, air temperature

Jaotuste segud rakendusega Eesti ilmaandmetele
Magistritöö
Kristi Ernits

Lühikokkuvõte. Jaotuste segud on leidnud rakendust mitmes valdkonnas kaasa arvatud meteoroloogias. Käesoleva magistritöö eesmärk on vaadelda jaotuste segusid keskendudes nende parameetrite hindamisele EM-algoritmi abil. Töös antakse ülevaade jaotuste segudest ja tuletatakse valemid konkreetsete jaotuste segude parameetrite hinnangute uuendamiseks EM-algoritmi sammudel. Lisaks rakendatakse jaotuste segusid Eesti ilmaandmetele. Erinevaid jaotuste segusid sobitatakse Eesti tuule kiiruse ööpäeva andmetele ja normaaljaotuste segusid temperatuuri ööpäeva andmetele.

CERCS teaduseriala: P160 Statistika, operatsioonanalüüs, programmeerimine, finants- ja kindlustusmatemaatika

Märksõnad: tõenäosusjaotused, jaotuste segud, EM-algoritm, tuule kiirus, õhutemperatuur

Contents

Introduction

A finite mixture arises in a natural way when a heterogeneous population can be divided into homogeneous subgroups while it is not possible to record the group indicator for the subject. For example, imagine observing human height without recording gender. More precisely, the distribution function of a finite mixture is defined as the weighted sum of K distribution functions. Hereby, finite mixtures are also called K-component mixtures, where $K < \infty$ indicates the number of mixture components, $K \in \mathbb{N}$. K-component mixtures with more than one component have often found to give a more accurate model than one-component models (see, e.g., Kollu et al., 2012; Miljkovic and Grün, 2016; Abu Bakar et al., 2018). Indeed, the shape of mixture density is flexible, being able to capture, for example, multimodality, skewness and kurtosis occurring in data.

The concept of a finite mixture is not new, for example, Pearson (1893) used the two-component normal mixture on ratio of forehead to body-length of crabs. Nowadays mixtures have been put into practice for modeling processes occurring in various fields, such as climatology, demographics, economics, healthcare, and others. For example, Trabzuni and Thomson (2014) found finite mixture useful in analysing gene expressions. Pittau et al. (2014) used mixtures to show the existence of levels of international attainment. Punzo et al. (2018) suggested finite mixtures for insurance and economic data with outliers. Antonio et al. (2014), Miljkovic and Grün (2016), Abu Bakar et al. (2018) showed that finite mixtures give good fit for the heavy tailed insurance claim sizes. Dias et al. (2013) analysed HIV/AIDS diagnosis-related data in Portugal by a finite mixture while Pelosi et al. (2015) applied mixtures for clustering human tissues. Antonio et al. (2014) successfully fitted mixtures of Erlangs to unemployment data.

In recent years finite mixtures have been found beneficial for climatological and meteorological data. That arises from the fact that the most accurate fit of weather variable(s) is of great importance in natural hazard risk management. Rainfall level, wind speed and temperature are important in disaster reduction, energy systems, economics, especially agriculture, and in particular in weather index insurance (e.g., Kjellström et al., 2007; Osman et al., 2015; Devis-Morales et al., 2017; Rohrbeck, 2017).

For estimating the parameters of finite mixtures various methods have been proposed. The method of moments and maximum likelihood based estimation are two popular methods. For the maximum likelihood estimation often a numerical algorithm, called the expectation-maximization (EM) algorithm, is used because allocations of objects to components are often not known. For mixtures of normal distributions the before mentioned methods are carefully studied and widely available (e.g., Everitt and Hand, 1981; McLachlan and Peel, 2000). However, for mixtures of other distributions, the exact derivations of necessary parameter estimates are frequently unavailable while the implementation of EM algorithm is freely available in several software packages.

This thesis has two main goals. Firstly, the aim is to study some characteristics of finite mixtures and the concept of the EM algorithm with examples for specific distributions. In addition, special attention is paid to the packages available in R (R Core Team, 2018) for estimating parameters of finite mixtures using the EM algorithm with the aim to possibly extend the range of distributions available. The second goal is to give insight to the behaviour of Estonian long term annual meteorological data, more precisely, to wind speed and temperature data, by applying the finite mixtures.

Main contributions of the thesis are following. Firstly, exact derivations for various specific mixtures, helpful for extended research or as study materials, are given. Secondly, the possibilities of R package `flexmix` (Leisch, 2004) are improved by adding the option for mixtures of Rayleigh and truncated normal mixtures. (The code developed for this thesis is available at `https://github.com/kristiern/codeexample`.) Thirdly, novel insight to Estonian daily wind speed and temperature behaviour is given by applying mixtures to Estonian meteorological data. Up to author's very best knowledge the last has not been done before.

The thesis is divided into three parts. In the first section the basic theory of finite mixtures with special focus on moments is given. In addition, examples for mixtures of specific distributions are provided. In the second section the EM algorithm is introduced and estimators for parameters of specific mixtures (lognormal, inverse Gaussian, gamma, Burr, inverse Burr, Weibull, Rayleigh and truncated normal mixtures) are derived in the concept of the EM algorithm. In Section 3 finite mixtures are applied on Estonian daily maximal hourly mean wind speed data and daily mean temperature data.

The analysis is carried out using R software version 3.5.2 (R Core Team, 2018). Parameters of finite mixtures are estimated using the EM algorithm, for that R package `flexmix` version 2.3-15 (Leisch, 2004) is used and extended.

1 Finite Mixtures

Let $G_1(y|\boldsymbol{\theta}_1)$ and $G_2(y|\boldsymbol{\theta}_2)$ be distribution functions depending on d_1-dimensional and d_2-dimensional parameter vectors $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, respectively, and let $\pi \in (0,1)$. Then function

$$F(y|(\boldsymbol{\theta}_1', \boldsymbol{\theta}_2', \pi)') = \pi G_1(y|\boldsymbol{\theta}_1) + (1-\pi) G_2(y|\boldsymbol{\theta}_2) \quad \text{for all } y \in \mathbb{R}, \tag{1}$$

is a new distribution function and it is called a mixture of G_1 and G_2, while distributions G_1 and G_2 are called the mixture components (e.g., Steutel and Van Harn, 2004, p. 327). Note that generally the mixture components do not have to be of the same distribution family.

The distribution F in (1) is the most straightforward example of a mixture. In general, the mixtures are given as follows (e.g., Steutel and Van Harn, 2004, pp. 327–328),

$$F(y) = \int_{\boldsymbol{\Theta}} G(y|\boldsymbol{\theta}) \mu(\mathrm{d}\boldsymbol{\theta}) \quad \text{for all } y \in \mathbb{R},$$

where μ is a probability measure on a measurable space[1] $(\boldsymbol{\Theta}, \Sigma)$ and $G(y|\boldsymbol{\theta})$ is a distribution function for every $\boldsymbol{\theta}$ such that for all y function $G^*(\boldsymbol{\theta}) = G(y|\boldsymbol{\theta})$, $G^* : \boldsymbol{\Theta} \to [0,1]$, is Σ-measurable. The set $\boldsymbol{\Theta}$ is sometimes called parameter space and in most cases μ is the Lebesgue-Stieltjes measure m_H induced by a distribution function H called the mixing distribution (see, for example, Steutel and Van Harn (2004, pp. 465–470) for details), then

$$F(y) = \int_{\boldsymbol{\Theta}} G(y|\boldsymbol{\theta}) \mathrm{d}H(\boldsymbol{\theta}) \quad \text{for all } y \in \mathbb{R}.$$

If the distribution function $G(y|\boldsymbol{\theta})$ is absolutely continuous with density $g(y|\boldsymbol{\theta})$ for all $\boldsymbol{\theta}$, then the distribution function $F(y)$ is also absolutely continuous with density (e.g., Steutel and Van Harn, 2004, p. 328)

$$f(y) = \int_{\boldsymbol{\Theta}} g(y|\boldsymbol{\theta}) \mathrm{d}H(\boldsymbol{\theta}) \quad \text{for all } y \in \mathbb{R},$$

where $f(y)$ is called mixture density (e.g., Everitt and Hand, 1981, p. 4).

In this thesis a special case of mixtures is considered when H is discrete, called finite mixtures. In this case, the discrete mixing distribution H assigns positive probabilities $\pi_1, \ldots, \pi_K$ only to a finite number of points $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$ (e.g., Everitt and Hand, 1981, p. 4). In addition, in this thesis it is assumed that all mixture components are distribution functions of a same distribution, more precisely, of a same continuous parametric distribution family (with possibly different parameter values). Note that if the mixture components are assumed to be distribution functions of the same family then, in general, the corresponding distribution function of the mixture is not from that distribution family.

1.1 Definition and Some Characteristics

Let Y be a continuous random variable taking values in a sample space $\Omega \subset \mathbb{R}$ and let $F(y|\boldsymbol{\Phi})$ be the distribution function of Y. Following Frühwirth-Schnatter (2006, p. 3–4) and Steutel and Van Harn (2004, p. 327–328) the formal definition of a finite mixture is given.

[1] Consider a nonempty set $\boldsymbol{\Theta}$ and a σ-algebra Σ on $\boldsymbol{\Theta}$. Then $(\boldsymbol{\Theta}, \Sigma)$ is called a measurable space.

Definition 1. A random variable Y is from a finite mixture if

$$F(y|\mathbf{\Phi}) = \sum_{k=1}^{K} \pi_k G(y|\boldsymbol{\theta}_k) \quad \text{for all } y \in \Omega, \tag{2}$$

where K is the number of components, $\mathbf{\Phi} = (\boldsymbol{\theta}'_1, \ldots, \boldsymbol{\theta}'_K, \boldsymbol{\pi}')' = (\boldsymbol{\theta}'_1, \ldots, \boldsymbol{\theta}'_K, (\pi_1, \ldots, \pi_K))'$ is the vector of parameters, $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$ are the component parameters, $G(y|\boldsymbol{\theta}_1), \ldots, G(y|\boldsymbol{\theta}_K)$ are the mixture components and $\pi_1, \ldots, \pi_K$ are the mixing weights such that

$$\pi_k > 0 \quad \text{for all } k \in \{1, \ldots, K\} \quad \text{and} \quad \sum_{k=1}^{K} \pi_k = 1. \tag{3}$$

Finite mixtures are also called K-component mixtures and hereafter mixtures refer to finite mixtures.

If the component distribution functions $G(y|\boldsymbol{\theta}_1), \ldots, G(y|\boldsymbol{\theta}_K)$ in (2) are absolutely continuous with densities $g(y|\boldsymbol{\theta}_1), \ldots, g(y|\boldsymbol{\theta}_K)$ then the distribution function $F(y|\mathbf{\Phi})$ is also absolutely continuous with density (Frühwirth-Schnatter, 2006, p. 4)

$$f(y|\mathbf{\Phi}) = \sum_{k=1}^{K} \pi_k g(y|\boldsymbol{\theta}_k) \quad \text{for all } y \in \Omega. \tag{4}$$

Note that the finite mixture density $f(y|\mathbf{\Phi})$ is indeed a density function: firstly, $f(y|\mathbf{\Phi})$ is non-negative for every $y \in \Omega$, because for all $k \in \{1, \ldots, K\}$ π_k are positive and $g(y|\boldsymbol{\theta}_k)$ are densities, secondly, $f(y|\mathbf{\Phi})$ integrates to 1:

$$\int_\Omega f(y|\mathbf{\Phi})\mathrm{d}y = \int_\Omega \sum_{k=1}^{K} \pi_k g(y|\boldsymbol{\theta}_k)\mathrm{d}y = \sum_{k=1}^{K} \pi_k \int_\Omega g(y|\boldsymbol{\theta}_k)\mathrm{d}y = \sum_{k=1}^{K} \pi_k = 1.$$

Next some basic characteristics of Y are introduced.

Assume the mth moment (e.g., Casella and Berger, 2002, p. 59) of kth component exists,

$$\mathbb{E}(Y^m|\boldsymbol{\theta}_k) = \int_\Omega y^m g(y|\boldsymbol{\theta}_k)\mathrm{d}y < \infty, \quad k = 1, \ldots, K,$$

where $\mathbb{E}$ is the expectation operator. Then the mth moment of the finite mixture exists (e.g., Frühwirth-Schnatter, 2006, p. 11; Withers et al., 2015),

$$\begin{aligned}\mathbb{E}(Y^m|\mathbf{\Phi}) = \int_\Omega y^m f(y|\mathbf{\Phi})\mathrm{d}y = \int_\Omega y^m \sum_{k=1}^{K} \pi_k g(y|\boldsymbol{\theta}_k)\mathrm{d}y = \\ = \sum_{k=1}^{K} \pi_k \int_\Omega y^m g(y|\boldsymbol{\theta}_k)\mathrm{d}y = \sum_{k=1}^{K} \pi_k \mathbb{E}(Y^m|\boldsymbol{\theta}_k).\end{aligned} \tag{5}$$

Thus, the mth moment of a K-component finite mixture is the sum of weighted component mth moments.

Assume the first moment of kth component, denoted by $\tilde{\mu}_k$, exists,

$$\tilde{\mu}_k = \mathbb{E}(Y|\boldsymbol{\theta}_k) < \infty, \quad k = 1, \ldots, K.$$

Let $\tilde{\mu}$ denote the first moment of finite mixture, $\tilde{\mu} = \mathbb{E}(Y|\mathbf{\Phi})$. Then based on Equation (5),

$$\tilde{\mu} = \sum_{k=1}^{K} \pi_k \tilde{\mu}_k. \tag{6}$$

The first moment is also known as the expectation or mean (e.g., Casella and Berger, 2002, p. 59). Hereby, the mean of a finite mixture is the sum of weighted means of components.

Assume the mth central moment (e.g., Casella and Berger, 2002, p. 59) of kth component exists,

$$\mathbb{E}((Y-\tilde{\mu}_k)^m|\boldsymbol{\theta}_k) = \int_{\Omega} (y-\tilde{\mu}_k)^m g(y|\boldsymbol{\theta}_k)\mathrm{d}y < \infty, \quad k = 1,\ldots,K.$$

Then the mth central moment of the finite mixture exists (e.g., Frühwirth-Schnatter, 2006, p. 11; Withers et al., 2015),

$$\begin{aligned}
\mathbb{E}((Y-\tilde{\mu})^m|\mathbf{\Phi}) &= \int_{\Omega} (y-\tilde{\mu})^m f(y|\mathbf{\Phi})\mathrm{d}y = \int_{\Omega} (y-\tilde{\mu})^m \sum_{k=1}^{K} \pi_k g(y|\boldsymbol{\theta}_k)\mathrm{d}y = \\
&= \sum_{k=1}^{K} \pi_k \int_{\Omega} (y-\tilde{\mu})^m g(y|\boldsymbol{\theta}_k)\mathrm{d}y = \sum_{k=1}^{K} \pi_k \mathbb{E}((Y-\tilde{\mu})^m|\boldsymbol{\theta}_k) = \\
&= \sum_{k=1}^{K} \pi_k \mathbb{E}((Y-\tilde{\mu}_k+\tilde{\mu}_k-\tilde{\mu})^m|\boldsymbol{\theta}_k) = \\
&= \sum_{k=1}^{K} \pi_k \mathbb{E}\left(\sum_{n=0}^{m} \mathrm{C}_m^n (Y-\tilde{\mu}_k)^n (\tilde{\mu}_k-\tilde{\mu})^{m-n}|\boldsymbol{\theta}_k\right) = \\
&= \sum_{k=1}^{K} \pi_k \sum_{n=0}^{m} \mathrm{C}_m^n (\tilde{\mu}_k-\tilde{\mu})^{m-n} \mathbb{E}((Y-\tilde{\mu}_k)^n|\boldsymbol{\theta}_k).
\end{aligned} \tag{7}$$

The second central moment (that is, in (7) fix $m = 2$) is called the variance (e.g., Casella and Berger, 2002, p. 59). Assume the variance of kth component, denoted by $\tilde{\sigma}_k^2$, exists,

$$\tilde{\sigma}_k^2 = \mathbb{E}((Y-\tilde{\mu}_k)^2|\boldsymbol{\theta}_k) < \infty, \quad k = 1,\ldots,K.$$

Let $\tilde{\sigma}^2$ denote the variance of finite mixture, $\tilde{\sigma}^2 = \mathbb{E}((Y-\tilde{\mu})^2|\mathbf{\Phi})$. Then based on (5),

$$\tilde{\sigma}^2 = \mathbb{E}(Y^2|\mathbf{\Phi}) - \tilde{\mu}^2 = \sum_{k=1}^{K} \pi_k \mathbb{E}(Y^2|\boldsymbol{\theta}_k) - \tilde{\mu}^2 = \sum_{k=1}^{K} \pi_k (\tilde{\mu}_k^2 + \tilde{\sigma}_k^2) - \tilde{\mu}^2. \tag{8}$$

From Equation (8) it is easy to see that the variance of a finite mixture can be expressed through mixing weights, means of components and variances of components.

The normalized third central moment is known as the skewness (e.g., Casella and Berger, 2002, p. 79). Assume the third central moment of kth component exists,

$$\mathbb{E}((Y-\tilde{\mu}_k)^3|\boldsymbol{\theta}_k) < \infty, \quad k = 1,\ldots,K.$$

Let $\tilde{\beta}_1$ denote the skewness of finite mixture, $\tilde{\beta}_1 = \mathbb{E}((Y-\tilde{\mu})^3|\mathbf{\Phi}) / \left(\sqrt{\tilde{\sigma}^2}\right)^3$. Then based on (7),

$$\tilde{\beta}_1 = \frac{\mathbb{E}((Y-\tilde{\mu})^3|\mathbf{\Phi})}{\left(\sqrt{\tilde{\sigma}^2}\right)^3} = \frac{1}{\tilde{\sigma}^3} \sum_{k=1}^{K} \pi_k \sum_{n=0}^{3} \mathrm{C}_3^n (\tilde{\mu}_k-\tilde{\mu})^{3-n} \mathbb{E}((Y-\tilde{\mu}_k)^n|\boldsymbol{\theta}_k) =$$

$$= \frac{1}{\tilde{\sigma}^3} \sum_{k=1}^{K} \pi_k \left[(\tilde{\mu}_k - \tilde{\mu})^3 + 3(\tilde{\mu}_k - \tilde{\mu})\tilde{\sigma}_k^2 + \mathbb{E}((Y - \tilde{\mu}_k)^3 | \boldsymbol{\theta}_k) \right]. \tag{9}$$

The normalized fourth central moment is known as the kurtosis (e.g., Casella and Berger, 2002, p. 79). Assume the fourth central moment of kth component exists,

$$\mathbb{E}((Y - \tilde{\mu}_k)^4 | \boldsymbol{\theta}_k) < \infty, \quad k = 1, \ldots, K.$$

Let $\tilde{\beta}_2$ denote the kurtosis of finite mixture, $\tilde{\beta}_2 = \mathbb{E}((Y - \tilde{\mu})^4 | \boldsymbol{\Phi}) / \left(\tilde{\sigma}^2\right)^2$. Then based on (7),

$$\tilde{\beta}_2 = \frac{\mathbb{E}((Y - \tilde{\mu})^4 | \boldsymbol{\Phi})}{(\tilde{\sigma}^2)^2} = \frac{1}{\tilde{\sigma}^4} \sum_{k=1}^{K} \pi_k \sum_{n=0}^{4} \mathrm{C}_4^n (\tilde{\mu}_k - \tilde{\mu})^{4-n} \mathbb{E}((Y - \tilde{\mu}_k)^n | \boldsymbol{\theta}_k) =$$

$$= \frac{1}{\tilde{\sigma}^4} \sum_{k=1}^{K} \pi_k \left[(\tilde{\mu}_k - \tilde{\mu})^4 + 6(\tilde{\mu}_k - \tilde{\mu})^2 \tilde{\sigma}_k^2 + 4(\tilde{\mu}_k - \tilde{\mu}) \mathbb{E}((Y - \tilde{\mu}_k)^3 | \boldsymbol{\theta}_k) + \mathbb{E}((Y - \tilde{\mu}_k)^4 | \boldsymbol{\theta}_k) \right].$$

The α-quantile of finite mixture is the value q_α that satisfies the equation (e.g., Gibbons and Chakraborti, 2003, p. 34)

$$F(q_\alpha | \boldsymbol{\Phi}) = \alpha, \tag{10}$$

where F is the distribution function (2). However, often (10) has to be solved numerically as generally for mixtures no closed form solution exists (e.g., Miljkovic and Grün, 2016).

1.2 Definitions of Specific Mixtures

In this section definitions of normal, truncated normal, lognormal, inverse Gaussian, gamma, Burr, inverse Burr, Weibull and Rayleigh mixtures are given. Recall that a continuous random variable Y is said to be from a finite mixture if its density is of form (4).

Definition 2 (Normal Mixture). A random variable Y is from a normal mixture if its density is of the form (4), where kth component is of a normal distribution (e.g., Casella and Berger, 2002, p. 102) with the density given by

$$g(y | \mu_k, \sigma_k) = \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{(y-\mu_k)^2}{2\sigma_k^2}}, \quad y \in \mathbb{R}, \quad k = 1, \ldots, K,$$

where $\mu_k \in \mathbb{R}$ is a location parameter and $\sigma_k > 0$ is a scale parameter.

The mean and variance of a normal mixture directly follow from (6) and (8), respectively, where the mean and variance of kth component are $\tilde{\mu}_k = \mu_k$ and $\tilde{\sigma}_k^2 = \sigma_k^2$ (e.g., Casella and Berger, 2002, p. 102), where μ_k and σ_k are component parameters ($k = 1, \ldots, K$). For examples and more discussion on normal mixtures see, for example, Everitt and Hand (1981), McLachlan and Peel (2000), Frühwirth-Schnatter (2006).

The normal distribution and the truncated normal distribution are closely related. In this thesis the truncated normal distribution refers to the truncated normal distribution with left tail truncated at zero and with no truncation for the right tail.

Definition 3 (**Truncated Normal Mixture**). A random variable Y is from a truncated normal mixture if its density is of the form (4), where kth component is of a truncated normal distribution (e.g., Forbes et al., 2011, p. 147) with the density given by

$$g(y|\sigma_k) = \frac{1}{\sqrt{2\pi}\sigma_k \Phi\left(\frac{\mu_k}{\sigma_k}\right)} e^{-\frac{(y-\mu_k)^2}{2\sigma_k^2}}, \quad y \geq 0, \quad k = 1, \ldots, K,$$

where $\mu_k \geq 0$ and $\sigma_k > 0$ are parameters and

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} \mathrm{d}t, \quad x \in \mathbb{R}, \tag{11}$$

is the standard normal distribution function (e.g., Casella and Berger, 2002, p. 102).

The mean and variance of a truncated normal mixture immediately follow from (6) and (8), respectively, where the mean of kth component is (e.g., Forbes et al., 2011, p. 147)

$$\tilde{\mu}_k = \mu_k + \frac{\sigma_k}{\sqrt{2\pi}\Phi\left(\frac{\mu_k}{\sigma_k}\right)} e^{-\frac{\mu_k^2}{2\sigma_k^2}}, \quad k = 1, \ldots, K,$$

and the variance of kth component is (e.g., Forbes et al., 2011, p. 148)

$$\tilde{\sigma}_k^2 = \sigma_k^2 - \frac{\mu_k \sigma_k}{\sqrt{2\pi}\Phi\left(\frac{\mu_k}{\sigma_k}\right)} e^{-\frac{\mu_k^2}{2\sigma_k^2}} - \frac{\sigma_k^2}{2\pi\Phi^2\left(\frac{\mu_k}{\sigma_k}\right)} e^{-\frac{\mu_k^2}{\sigma_k^2}}, \quad k = 1, \ldots, K,$$

where μ_k and σ_k are component parameters.

Definition 4 (**Lognormal Mixture**). A random variable Y is from a lognormal mixture if its density is of the form (4), where kth component is of a lognormal distribution (e.g., Forbes et al., 2011, p. 131) with the density given by

$$g(y|\mu_k, \sigma_k) = \frac{1}{y\sigma_k\sqrt{2\pi}} e^{-\frac{(\ln y-\mu_k)^2}{2\sigma_k^2}}, \quad y > 0, \quad k = 1, \ldots, K,$$

where $\mu_k \in \mathbb{R}$ and $\sigma_k > 0$ are parameters.

The mean and variance of a lognormal mixture directly follow from (6) and (8), respectively, where the mean and variance of kth component are (e.g., Forbes et al., 2011, p. 132)

$$\tilde{\mu}_k = e^{\mu_k + \frac{1}{2}\sigma_k^2} \quad \text{and} \quad \tilde{\sigma}_k^2 = e^{2\mu_k + \sigma_k^2}\left(e^{\sigma_k^2} - 1\right), \quad k = 1, \ldots, K,$$

where μ_k and σ_k are component parameters.

Example 1. Let Y be a random variable from a three-component lognormal mixture with mixing weights $\pi_1 = 0.4$, $\pi_2 = 0.3$, $\pi_3 = 0.3$ and component parameters $\mu_1 = 1.5$, $\mu_2 = 2.5$, $\mu_3 = 3$, $\sigma_1 = 0.25$, $\sigma_2 = 0.25$, $\sigma_3 = 0.5$. Its mean is

$$\tilde{\mu} = \frac{1}{10} e^{\frac{49}{32}} \left(4 + 3e + 3e^{\frac{51}{32}}\right) \approx 12.448,$$

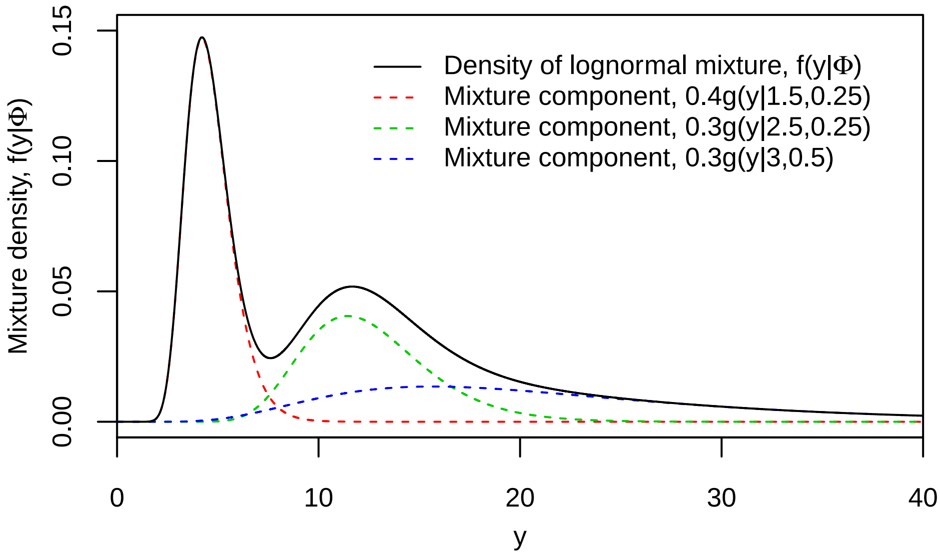


Figure 1: An example of three-component lognormal mixture

while the means of the components are $\tilde{\mu}_1 \approx 4.624$, $\tilde{\mu}_2 \approx 12.569$ and $\tilde{\mu}_3 \approx 22.760$. The variance of the three-component lognormal mixture is

$$\tilde{\sigma}^2 = \frac{1}{10}e^{\frac{25}{8}}\left(4+3e^2+3e^{\frac{27}{8}}\right) - \tilde{\mu}^2 \approx 104.138,$$

while the variances of the components are $\tilde{\sigma}_1^2 \approx 1.379$, $\tilde{\sigma}_2^2 \approx 10.189$ and $\tilde{\sigma}_3^2 \approx 147.129$.

The density of the three-component lognormal mixture is

$$f(y|\mathbf{\Phi}) = 0.4g(y|1.5,0.25) + 0.3g(y|2.5,0.25) + 0.3g(y|3,0.5) =$$
$$= \frac{1}{5\sqrt{2\pi}y}\left(8e^{-8(\ln y-1.5)^2} + 6e^{-8(\ln y-2.5)^2} + 3e^{-2(\ln y-3)^2}\right), \quad y > 0.$$

Based on the graph it can be said that the density is bimodal (has two local maxima) although it is a mixture of three lognormal densities (see Figure 1).

Definition 5 (Inverse Gaussian Mixture). A random variable Y is from an inverse Gaussian mixture if its density is of the form (4), where kth component is of an inverse Gaussian distribution (e.g., Forbes et al., 2011, p. 120) with the density given by

$$g(y|\mu_k,\lambda_k) = \sqrt{\frac{\lambda_k}{2\pi y^3}}e^{-\frac{\lambda_k(y-\mu_k)^2}{2\mu_k^2 y}}, \quad y > 0, \quad k = 1,\ldots,K,$$

where $\mu_k > 0$ is a location parameter and $\lambda_k > 0$ is a scale parameter.

The mean and variance of an inverse Gaussian mixture directly follow, respectively, from (6) and (8), where the mean and variance of kth component are (e.g., Forbes et al., 2011, p. 120)

$$\tilde{\mu}_k = \mu_k \quad \text{and} \quad \tilde{\sigma}_k^2 = \frac{\mu_k^3}{\lambda_k}, \quad k = 1,\ldots,K,$$

where μ_k and λ_k are component parameters.

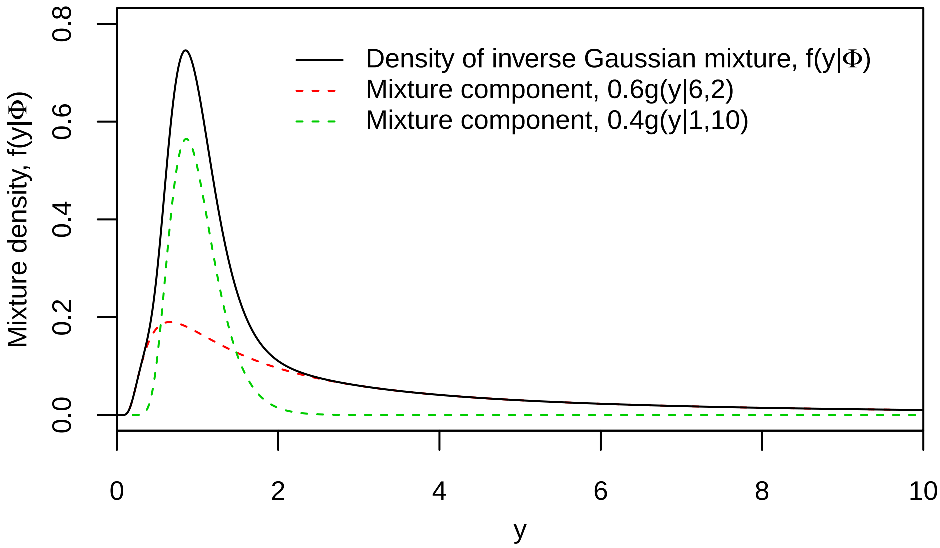


Figure 2: An example of two-component inverse Gaussian mixture

Example 2. Let Y be a random variable from a two-component inverse Gaussian mixture with mixing weights $\pi_1 = 0.6$, $\pi_2 = 0.4$ and component parameters $\mu_1 = 6$, $\mu_2 = 1$, $\lambda_1 = 2$, $\lambda_2 = 10$. Its density is

$$f(y|\mathbf{\Phi}) = 0.6g(y|6,2) + 0.4g(y|1,10) = \frac{1}{\sqrt{\pi y^3}}\left(e^{-\frac{(y-6)^2}{36y}} + \sqrt{5}e^{-\frac{5(y-1)^2}{y}}\right), \quad y > 0.$$

The density is shown in Figure 2. It seems to be unimodal although it is a mixture of two inverse Gaussian densities. The density also shows possible positive skewness, which is confirmed by the value of the skewness coefficient, $\tilde{\beta}_1 \approx 34.038$. The skewness of an inverse Gaussian mixture immediately follows from (9), where the third central moment of kth component is (e.g., Forbes et al., 2011, p. 121)

$$\mathbb{E}((Y - \tilde{\mu}_k)^3|\boldsymbol{\theta}_k) = 3\sqrt{\frac{\mu_k}{\lambda_k}}\left(\sqrt{\tilde{\sigma}_k^2}\right)^3 = \frac{3\mu_k^5}{\lambda_k^2}, \quad k = 1, \ldots, K,$$

where μ_k and λ_k are component parameters. The mean and variance of the two-component inverse Gaussian mixture are $\tilde{\mu} = 4$ and $\tilde{\sigma}^2 = 70.84$.

Definition 6 (**Gamma Mixture**)**.** A random variable Y is from a gamma mixture if its density is of the form (4), where kth component is of a gamma distribution (e.g., Forbes et al., 2011, p. 109) with density given by

$$g(y|\alpha_k, \lambda_k) = \frac{\lambda_k^{\alpha_k} y^{\alpha_k - 1}}{\Gamma(\alpha_k)} e^{-y\lambda_k}, \quad y \geq 0, \quad k = 1, \ldots, K,$$

where $\alpha_k > 0$ is a shape parameter, $\lambda_k > 0$ is a rate parameter and

$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} \mathrm{d}t, \quad x > 0, \tag{12}$$

is the gamma function (e.g., Forbes et al., 2011, p. 56).

The mean and variance of a gamma mixture directly follow from (6) and (8), respectively, where the mean and variance of kth component are (e.g., Forbes et al., 2011, p. 109)

$$\tilde{\mu}_k = \frac{\alpha_k}{\lambda_k} \quad \text{and} \quad \tilde{\sigma}_k^2 = \frac{\alpha_k}{\lambda_k^2}, \quad k = 1, \dots, K,$$

where α_k and λ_k are component parameters.

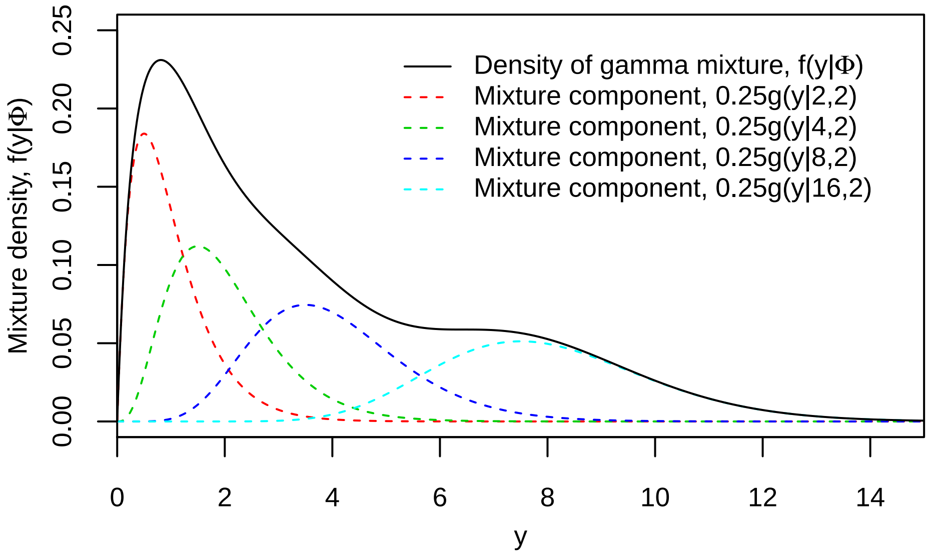


Figure 3: An example of four-component gamma mixture

Example 3. Let Y be a random variable from a four-component gamma mixture with equal mixing weights $\pi_k = 0.25$ ($k = 1,2,3,4$), component shape parameters $\alpha_1 = 2$, $\alpha_2 = 4$, $\alpha_3 = 8$, $\alpha_4 = 16$ and equal component rate parameters $\lambda_k = 2$ ($k = 1,2,3,4$). Its density is

$$f(y|\mathbf{\Phi}) = 0.25g(y|2,2) + 0.25g(y|4,2) + 0.25g(y|8,2) + 0.25g(y|16,2) =$$
$$= ye^{-2y}\left(1 + \frac{2}{3}y^2 + \frac{4}{315}y^6 + \frac{8}{638512875}y^{14}\right), \quad y \geq 0.$$

The density is shown in Figure 3, it is an example of mixture density that is unimodal (based on the graph) although it is a mixture of four unimodal gamma densities.

The Burr distribution, more precisely the Burr type XII distribution is also known as the Singh–Maddala distribution or the Pareto (IV) distribution or the beta-P distribution or the generalized log-logistic distribution (Kleiber and Kotz, 2003, p. 198).

Definition 7 (**Burr Mixture**)**.** A random variable Y is from a Burr mixture if its density is of the form (4), where kth component is of a Burr distribution (e.g., Zimmer et al., 1998) with density given by

$$g(y|\alpha_k, \gamma_k, \theta_k) = \frac{\alpha_k \gamma_k y^{\gamma_k - 1}}{\theta_k^{\gamma_k}\left(1 + \frac{y^{\gamma_k}}{\theta_k^{\gamma_k}}\right)^{\alpha_k + 1}}, \quad y > 0, \quad k = 1, \dots, K,$$

where $\alpha_k > 0$ and $\gamma_k > 0$ are shape parameters and $\theta_k > 0$ is a scale parameter.

Let B denote the beta function (e.g., Forbes et al., 2011, p. 56),

$$B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}\mathrm{d}t = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}, \quad x > 0, \quad y > 0, \tag{13}$$

where Γ is the gamma function (12).

The mean and variance of a Burr mixture immediately follow from (6) and (8), respectively, where the mean of kth component is (e.g., Zimmer et al., 1998)

$$\tilde{\mu}_k = \theta_k \alpha_k B\left(\alpha_k - \frac{1}{\gamma_k}, 1 + \frac{1}{\gamma_k}\right), \quad k = 1, \ldots, K,$$

and the variance of kth component is (e.g., Zimmer et al., 1998)

$$\tilde{\sigma}_k^2 = \theta_k^2 \alpha_k \left[B\left(\alpha_k - \frac{2}{\gamma_k}, 1 + \frac{2}{\gamma_k}\right) - \alpha_k B^2\left(\alpha_k - \frac{1}{\gamma_k}, 1 + \frac{1}{\gamma_k}\right)\right], \quad k = 1, \ldots, K,$$

where α_k, γ_k and θ_k are component parameters. For the mixture mean $\tilde{\mu}$ to exist it must hold that $\alpha_k \gamma_k > 1$ for all $k \in \{1, \ldots, K\}$ and for the mixture variance $\tilde{\sigma}^2$ to exist it must hold that $\alpha_k \gamma_k > 2$ for all $k \in \{1, \ldots, K\}$ (e.g., Kleiber and Kotz, 2003, p. 201).

Example 4. Let Y be a random variable from a two-component Burr mixture with mixing weights $\pi_1 = 0.8$, $\pi_2 = 0.2$, component shape parameters $\alpha_1 = 4$, $\alpha_2 = 2$, $\gamma_1 = 1$, $\gamma_2 = 4$ and component scale parameters $\theta_1 = 2$, $\theta_2 = 3$. Its density is

$$f(y|\mathbf{\Phi}) = 0.8g(y|4,1,2) + 0.2g(y|2,4,3) = \frac{256}{5(y+2)^5} + \frac{52488y^3}{5(y^4+81)^3}, \quad y > 0.$$

The density is shown in Figure 4, the shape of the density is strongly influenced by the shape of the component density with the greater weight $\pi_1 = 0.8$.

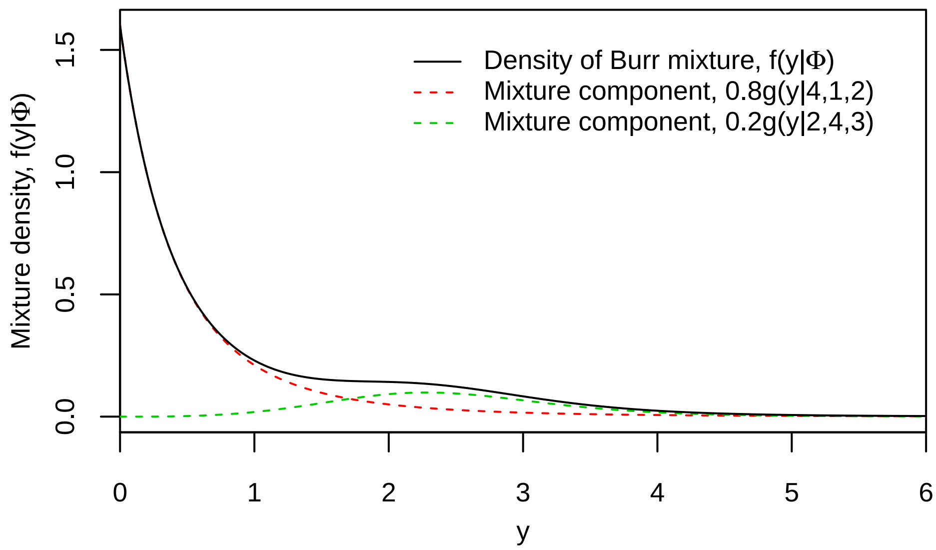


Figure 4: An example of two-component Burr mixture

The mean and variance of the two-component Burr mixture are

$$\tilde{\mu} = \frac{32}{5}B(3,2) + \frac{6}{5}B\left(\frac{7}{4}, \frac{5}{4}\right) = \frac{8}{15} + \frac{9\pi}{40\sqrt{2}} \approx 1.033 \quad \text{and}$$

$$\tilde{\sigma}^2 = \frac{64}{5}B(2,3) + \frac{18}{5}B\left(\frac{3}{2},\frac{3}{2}\right) - \tilde{\mu}^2 = \frac{16}{15} + \frac{9\pi}{20} - \tilde{\mu}^2 \approx 1.413.$$

The inverse Burr distribution, more precisely the Burr type III distribution is also known as the Dagum distribution or the generalized log-logistic distribution or the (three-parameter) kappa distribution or the beta-K distribution (Kleiber and Kotz, 2003, pp. 212–213).

Definition 8 (**Inverse Burr Mixture**)**.** A random variable Y is from an inverse Burr mixture if its density is of the form (4), where kth component is of an inverse Burr distribution (e.g., Kleiber and Kotz, 2003, p. 212) with density given by

$$g(y|\tau_k,\gamma_k,\theta_k) = \frac{\tau_k\gamma_k y^{\tau_k\gamma_k - 1}}{\theta_k^{\tau_k\gamma_k}\left(1 + \frac{y^{\gamma_k}}{\theta_k^{\gamma_k}}\right)^{\tau_k+1}}, \quad y > 0, \quad k = 1,\ldots,K,$$

where $\tau_k > 0$ and $\gamma_k > 0$ are shape parameters and $\theta_k > 0$ is a scale parameter.

The mean and variance of an inverse Burr mixture immediately follow from (6) and (8), respectively, where the mean of kth component is (e.g., Kleiber and Kotz, 2003, p. 214)

$$\tilde{\mu}_k = \theta_k\tau_k B\left(\tau_k + \frac{1}{\gamma_k}, 1 - \frac{1}{\gamma_k}\right), \quad k = 1,\ldots,K,$$

and the variance of kth component is (e.g., Kleiber and Kotz, 2003, p. 214)

$$\tilde{\sigma}_k^2 = \theta_k^2\tau_k\left[B\left(\tau_k + \frac{2}{\gamma_k}, 1 - \frac{2}{\gamma_k}\right) - \tau_k B^2\left(\tau_k + \frac{1}{\gamma_k}, 1 - \frac{1}{\gamma_k}\right)\right], \quad k = 1,\ldots,K.$$

where B is the beta function (13) and τ_k, γ_k and θ_k are component parameters. The mixture mean $\tilde{\mu}$ exists if $\gamma_k > 1$ for all $k \in \{1,\ldots,K\}$ and the mixture variance $\tilde{\sigma}^2$ exists if $\gamma_k > 2$ for all $k \in \{1,\ldots,K\}$ (e.g., Kleiber and Kotz, 2003, p. 213).

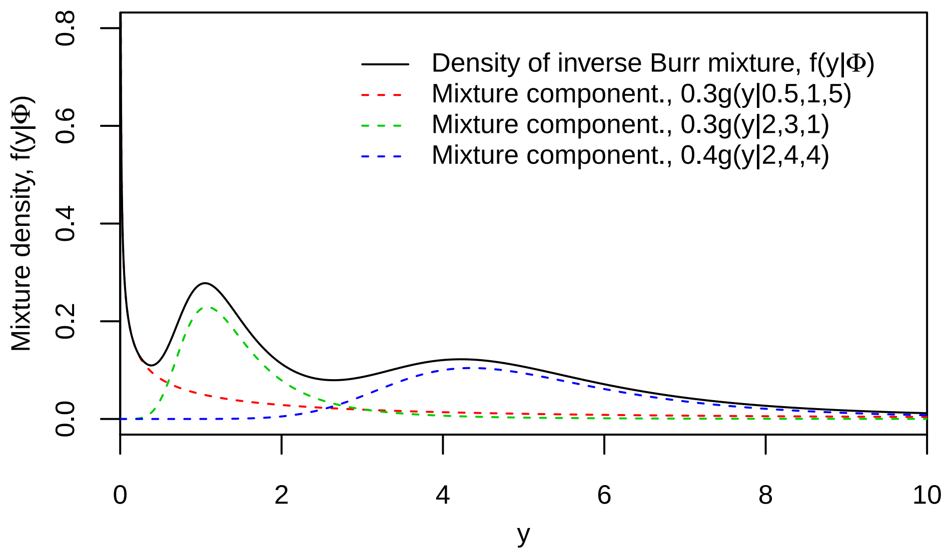


Figure 5: An example of three-component inverse Burr mixture

Example 5. Let Y be a random variable from a three-component inverse Burr mixture with mixing weights $\pi_1 = 0.3$, $\pi_2 = 0.3$, $\pi_3 = 0.4$, component shape parameters $\tau_1 = 0.5$, $\tau_2 = 2$, $\tau_3 = 2$, $\gamma_1 = 1$, $\gamma_2 = 3$, $\gamma_3 = 4$ and component scale parameters $\theta_1 = 5$, $\theta_2 = 1$, $\theta_3 = 4$. Its mean and variance do not exist as because of $\gamma_1 = 1$ it does not hold that $\gamma_k > 1$ for all $k \in \{1,2,3\}$ and $\gamma_k > 2$ for all $k \in \{1,2,3\}$. The density of the inverse Burr mixture is

$$f(y|\mathbf{\Phi}) = 0.3g(y|0.5,1,5) + 0.3g(y|2,3,1) + 0.4g(y|2,4,4) =$$
$$= \frac{3}{4\sqrt{y}(y+5)^{3/2}} + \frac{9y^5}{5(1+y^3)^3} + \frac{4096y^7}{5(y^4+256)^3}, \quad y > 0.$$

The density is shown in Figure 5 and based on the graph it has a rather complicated shape with at least four local extrema.

Definition 9 (Weibull Mixture). A random variable Y is from a Weibull mixture if its density is of the form (4), where kth component is of a two-parameter Weibull distribution (e.g., Forbes et al., 2011, p. 193) with density given by

$$g(y|\alpha_k, \theta_k) = \frac{\alpha_k y^{\alpha_k - 1}}{\theta_k^{\alpha_k}} e^{-\left(\frac{y}{\theta_k}\right)^{\alpha_k}}, \quad y \geq 0, \quad k = 1, \ldots, K,$$

where $\alpha_k > 0$ is a shape parameter and $\theta_k > 0$ is a scale parameter.

The mean and variance of a Weibull mixture immediately follow from (6) and (8), where the mean and variance of kth component are (e.g., Forbes et al., 2011, p. 193)

$$\tilde{\mu}_k = \theta_k \Gamma\left(1 + \frac{1}{\alpha_k}\right) \quad \text{and} \quad \tilde{\sigma}_k^2 = \theta_k^2 \Gamma\left(1 + \frac{2}{\alpha_k}\right) - \left[\theta_k \Gamma\left(1 + \frac{1}{\alpha_k}\right)\right]^2, \quad k = 1, \ldots, K,$$

where Γ is the gamma function (12) and α_k and θ_k are component parameters.

Example 6. Let Y be a random variable from a two-component Weibull mixture with mixing weights $\pi_1 = 0.8$, $\pi_2 = 0.2$ and component parameters $\alpha_1 = 2$, $\alpha_2 = 3$, $\theta_1 = 3$, $\theta_2 = 1$. Its

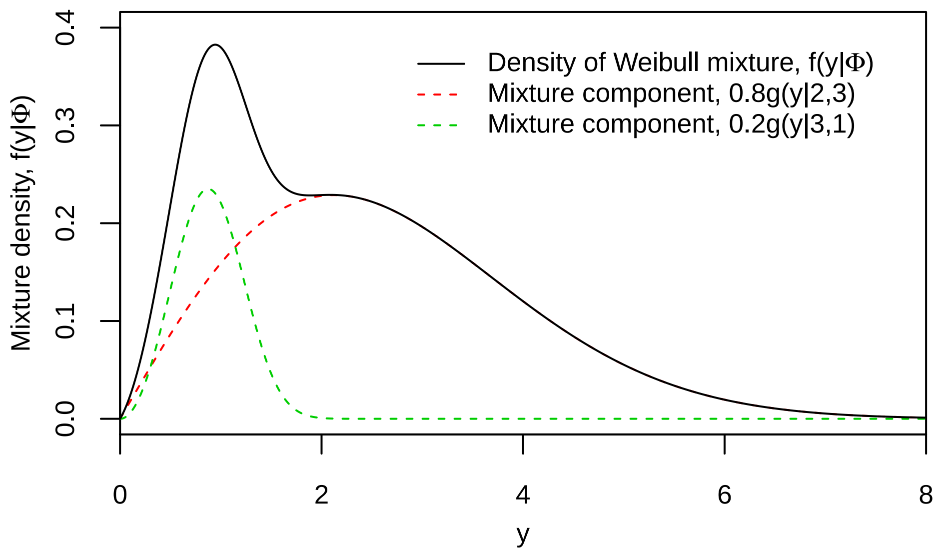


Figure 6: An example of two-component Weibull mixture

mean and variance are

$$\tilde{\mu} = \frac{1}{5}\left[6\sqrt{\pi} + \Gamma\left(\frac{4}{3}\right)\right] \approx 2.306 \quad \text{and} \quad \tilde{\sigma}^2 = \frac{1}{5}\left[36 + \Gamma\left(\frac{5}{3}\right)\right] - \tilde{\mu}^2 \approx 2.065.$$

The density of the two-component Weibull mixture shown in Figure 6 is

$$f(y|\boldsymbol{\Phi}) = 0.8g(y|2,3) + 0.2g(y|3,1) = \frac{8}{45}ye^{-\frac{y^2}{9}} + \frac{3}{5}y^2e^{-y^3}, \quad y \geq 0.$$

Based on the graph the density has one mode despite being a mixture of two Weibull densities.

Definition 10 (**Rayleigh Mixture**). A random variable Y is from a Rayleigh mixture if its density is of the form (4), where kth component is of a Rayleigh distribution (e.g., Forbes et al., 2011, p. 173) with density given by

$$g(y|\sigma_k) = \frac{y}{\sigma_k^2}e^{-\frac{y^2}{2\sigma_k^2}}, \quad y \geq 0, \quad k = 1, \ldots, K,$$

where $\sigma_k > 0$ is a scale parameter.

Note that a Rayleigh distribution with parameter σ_k is equivalent to a Weibull distribution with parameters $\alpha_k = 2$ and $\theta_k = \sqrt{2}\sigma_k$ (e.g., Forbes et al., 2011, p. 173).

The mean and variance of a Rayleigh mixture directly follow from (6) and (8), respectively, where the mean and variance of kth component are (e.g., Forbes et al., 2011, p. 173)

$$\tilde{\mu}_k = \sqrt{\frac{\pi}{2}}\sigma_k, \quad \text{and} \quad \tilde{\sigma}_k^2 = \left(2 - \frac{\pi}{2}\right)\sigma_k^2, \quad k = 1, \ldots, K,$$

where σ_k is the component parameter.

2 Estimation of Finite Mixtures

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ be a random sample of n independent identically distributed random variables from a finite mixture distribution and let its realization be denoted by $\boldsymbol{y} = (y_1, \ldots, y_n)'$. To use the realization of the sample to make inferences about the underlying mixture structure, for example to estimate parameters of the base finite mixture, concept of a standard finite mixture model should be introduced (e.g., Frühwirth-Schnatter, 2006, p. 12). The standard mixture model is often written out in form of a density of the random variable Y_i $(i = 1, \ldots, n)$ (e.g., McLachlan and Peel, 2000, p. 6):

$$f(y_i|\boldsymbol{\Phi}) = \sum_{k=1}^{K} \pi_k g(y_i|\boldsymbol{\theta}_k). \tag{14}$$

Note that the standard finite mixture model can be viewed as a hierarchical model where the distribution of the observations $\boldsymbol{y} = (y_1, \ldots, y_n)'$ depends on hidden component-allocations

$$\boldsymbol{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n), \tag{15}$$

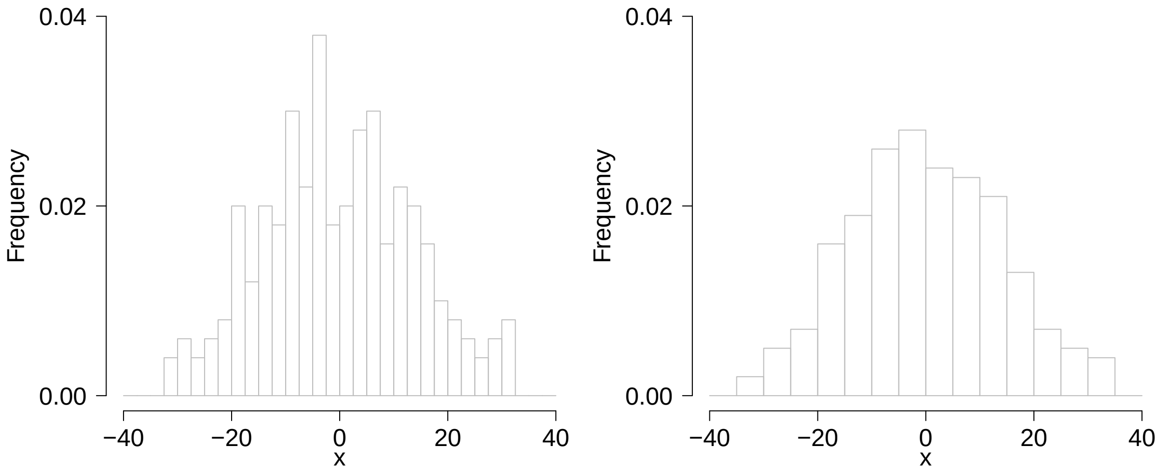
where $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{iK})'$ $(i = 1, \ldots, n)$ are independent identically distributed random component-indicator vectors, wherein $Z_{ik} = 1$ with probability π_k $(k = 1, \ldots, K)$. The vectors $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ are from multinomial distribution $Mult(1; \pi_1, \ldots, \pi_K)$ with parameter for number of trials equal to 1 and mixing weights $\pi_1, \ldots, \pi_K$ as event probabilities. For more details see, for example, McLachlan and Peel (2000, pp. 6–7, 19–21) and Frühwirth-Schnatter (2006, pp. 11–12, 25).

2.1 Estimation of Number of Components

In some applications the number of components, K, or even the component-allocations are known, but mostly in practice the K is unknown and considered as part of the model estimation procedure. Estimating the number of components is one of the key aspects for finding the most suitable distribution among finite mixtures to model the data.

Prior information on components might be visible through informal methods, the initial natural approach for deciding on the possible value of K is to inspect the histogram for multimodality (e.g., Everitt and Hand, 1981, p. 108). One of the pitfalls of using the histogram is that even if the underlying distribution is unimodal the histogram may appear multimodal. In Figure 7 the appearance of the histogram of a random sample from a normal distribution depends also on the chosen bin width for the histogram: with width 2.5 the histogram appears multimodal, with width 5 unimodal. Additionally, as seen in Examples 2 and 3, the mixture distribution itself can be unimodal even if it is a mixture of two or more components. For two-component normal mixtures conditions of the density being unimodal or multimodal are given for example in Everitt and Hand (1981, Chapter 2.2).

Thus, examining the sample histogram might not be helpful when detecting presence of a mixture, it might even be misleading (e.g., Everitt and Hand, 1981, p. 108). Because



(a) Bin width: 2.5

(b) Bin width: 5

Figure 7: Histograms of 200 observations from normal distribution ($\mu = 0$, $\sigma = 15$)

of that also other informal techniques have been developed, for example Everitt and Hand (1981, p. 108) bring out checking for existence of more than two points of inflexion in the frequency curve and different probability plotting methods, McLachlan and Peel (2000, p. 184) mention residual diagnostics and Frühwirth-Schnatter (2006, pp. 108–109) describes mode hunting in the mixture posterior density.

In addition to the informal methods some more formal methods exist for estimating the number of components, K. For example, nonparametric tests for the number of modes (the components should be quite far apart to be distinguishable) (e.g., McLachlan and Peel, 2000, p. 176), the method of moments and Bayesian posterior predictive model checking brought out in Frühwirth-Schnatter (2006, pp. 110–112, 123) can be used. One other way to find a suitable number of components is to conduct a hypothesis test, for that the likelihood ratio test or its modifications are most widely used (e.g., Frühwirth-Schnatter, 2006, pp. 114–115).

However, in practice the number of components is often determined after the procedure of estimating the parameters of mixtures with different number of components, for example by using some likelihood-based criteria. The mixture likelihood (see also Equation (14)) is expressed as

$$L(\mathbf{\Phi}) = \prod_{i=1}^{n} f(y_i|\mathbf{\Phi}) = \prod_{i=1}^{n} \left(\sum_{k=1}^{K} \pi_k g(y_i|\boldsymbol{\theta}_k) \right). \tag{16}$$

The selection of K is determined by the best goodness of fit characteristic (e.g., Frühwirth-Schnatter, 2006, pp. 116–117), for example by Akaike information criterion (23) or Bayesian information criterion (24). Then the model with the lowest criterion value is chosen. Frühwirth-Schnatter (2006, pp. 116–117) and McLachlan and Peel (2000, pp. 175) bring out that the Bayesian information criterion is preferred over the Akaike information criterion to make the selection, because the AIC favours models with more parameters than BIC if $\ln n > 2$ or equivalently $n > 7$. They add, it has been shown asymptotically that under mild conditions the AIC and BIC do not underestimate the correct number of components and are considered to be less demanding than the likelihood ratio test.

The last approach, using the BIC, is also followed in this thesis to estimate the number of components, it is not estimated independently but within the parameter estimation procedure.

2.2 Estimation of Parameters

In this thesis it is assumed that all mixture components are from the same continuous parametric distribution family so that the parameter vectors $\boldsymbol{\theta}_k$ ($k = 1, \ldots, K$) are of length d, note also that because of (3)

$$\pi_K = 1 - \sum_{k=1}^{K-1} \pi_k.$$

Thus, there are maximally $dK + (K-1) = (d+1)K - 1$ free parameters.

Estimation of parameters of a finite mixture usually subsumes estimation of the component parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$ and the mixing weights $\pi_1, \ldots, \pi_K$ as commonly they are unknown. According to Frühwirth-Schnatter (2006, pp. 25–56) estimating parameters of a finite mixture can be divided into two (if the number of components is fixed):

1. Estimation of parameters $\boldsymbol{\Phi}$ when the allocations $\boldsymbol{Z}$ in (15) of objects $\boldsymbol{y} = (y_1, \ldots, y_n)'$ to certain components are known.
2. Estimation of parameters $\boldsymbol{\Phi}$ when the allocations are not observed.

In this thesis the last approach is viewed as more commonly, and also in the application in Section 3, the allocations are not observable or not known. Additionally, by the estimation of parameters the number of components, K, is considered fixed.

For estimating parameters of K-component mixtures a wide range of different methods exist, including both formal approaches and informal graphical techniques. Method of moments is considered to be the oldest estimation method for parameters of finite mixtures, for example, by Everitt and Hand (1981), McLachlan and Peel (2000) and Frühwirth-Schnatter (2006), because Pearson (1893) used it to estimate the five parameters of a two-component normal mixture. Everitt and Hand (1981, pp. 17–18, 31–35) and Frühwirth-Schnatter (2006, pp. 42–43) give a short overview of the method of moments for estimating mixture parameters with examples and bring out some pitfalls of the method: it might be computationally heavy and the method is inferior to the maximum likelihood estimation method.

Frühwirth-Schnatter (2006, p. 49) states that the maximum likelihood estimation became the most used method for estimating parameters of finite mixtures when numerical algorithms became available, with methods such as Newton's or gradient method being used at first for maximizing the likelihood. Nowadays the expectation-maximization (EM) algorithm is the most commonly applied method to find the maximum likelihood estimators for parameters of K-component mixtures (e.g., Frühwirth-Schnatter, 2006, p. 49; McLachlan and Krishnan, 2008, p. 40). The EM algorithm can handle the estimation of parameters of finite mixtures when incomplete data problem is present, that is, the allocations $\boldsymbol{Z}$ in (15) are not known (e.g., McLachlan and Peel, 2000, p. 20).

In this thesis the expectation-maximization algorithm is applied to estimate parameters of K-component mixtures (for fixed number of components and unknown allocations) and for clarity derivations of the estimates are given.

According to Everitt and Hand (1981, p. 7–22, 48–57), McLachlan and Peel (2000, pp. 35–37), Frühwirth-Schnatter (2006, p. 41–56) the mixture parameters $\mathbf{\Phi}$ can also be estimated using graphical methods (suggested to use for initial examination of data), Bayesian estimation, different distance-based methods, error minimization methods and other methods. Some of the methods are developed for special cases or specific mixtures, for example, Fourier transformation methods for normal mixtures (e.g., Everitt and Hand, 1981, pp. 48–57). Withers et al. (2015) suggest using cumulants to estimate the mixture parameters.

2.2.1 Expectation–Maximization Algorithm

In this section overview of using the expectation–maximization algorithm for estimating parameters of K-component mixtures is given based on Everitt and Hand (1981), McLachlan and Peel (2000), Frühwirth-Schnatter (2006) and McLachlan and Krishnan (2008). Recall the assumptions made earlier: mixture components are from the same continuous parametric distribution family, the number of components, K, is fixed and the allocations $\boldsymbol{Z}$ in (15) are unknown.

The vector of observed data $\boldsymbol{y} = (y_1, \ldots, y_n)'$ is named the incomplete data in the framework of the EM algorithm as the allocations are not known. The complete data vector, denoted by $\boldsymbol{c}$, consists of the observed data vector $\boldsymbol{y}$ and vectors $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iK})'$, where $z_{ik} = 1$ if observation y_i is from mixture component k and $z_{ik} = 0$ otherwise ($i = 1, \ldots, n$; $k = 1, \ldots, K$):

$$\boldsymbol{c} = (\boldsymbol{y}', \boldsymbol{z}')' = (\boldsymbol{y}', \boldsymbol{z}_1', \ldots, \boldsymbol{z}_n')',$$

where $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ are realizations of independent identically distributed random vectors $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ from multinomial distribution $Mult(1; \pi_1, \ldots, \pi_K)$.

Taking into account the assumptions made, the complete data likelihood, denoted by $L_c(\mathbf{\Phi})$, can be shown to be equal to (e.g., Frühwirth-Schnatter, 2006, pp. 29–30)

$$L_c(\mathbf{\Phi}) = \prod_{i=1}^{n} \prod_{k=1}^{K} (\pi_k g(y_i|\boldsymbol{\theta}_k))^{z_{ik}}.$$

Thus, the complete data log-likelihood function, denoted by $l_c(\mathbf{\Phi})$, can be expressed as

$$l_c(\mathbf{\Phi}) = \ln L_c(\mathbf{\Phi}) = \ln \left[\prod_{i=1}^{n} \prod_{k=1}^{K} (\pi_k g(y_i|\boldsymbol{\theta}_k))^{z_{ik}} \right] = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \left[\ln \pi_k + \ln(g(y_i|\boldsymbol{\theta}_k)) \right].$$

The EM algorithm is an iterative method for which every iteration p consists of two steps: the expectation step (E-step) and the maximization step (M-step). In the E-step (in the concept of finite mixtures) the conditional expectation of log-likelihood $l_c(\mathbf{\Phi})$ given the observed data and current parameter estimates from step $p-1$, denoted by Q and called the Q-function, is found. In the M-step the Q-function is maximized to obtain new estimates, denoted by $\mathbf{\Phi}^{(p)}$, for the parameters $\mathbf{\Phi}$.

E-step

The conditional expectation of log-likelihood is found, because the component indicators z_{ik} $(i = 1,\dots,n;\ k = 1,\dots,K)$ are not observed. On step p the Q-function is

$$\begin{aligned} Q(\mathbf{\Phi}|\mathbf{\Phi}^{(p-1)}) = \mathbb{E}\left(l_c(\mathbf{\Phi})|\mathbf{y},\mathbf{\Phi}^{(p-1)}\right) = \mathbb{E}\left[\sum_{i=1}^{n}\sum_{k=1}^{K} Z_{ik}\left[\ln\pi_k + \ln(g(y_i|\boldsymbol{\theta}_k))\right] \middle| y_i,\mathbf{\Phi}^{(p-1)}\right] = \\ = \sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{E}\left(Z_{ik}|y_i,\mathbf{\Phi}^{(p-1)}\right)\left[\ln\pi_k + \ln(g(y_i|\boldsymbol{\theta}_k))\right] = \\ = \sum_{i=1}^{n}\sum_{k=1}^{K}\hat{w}_{ik}^{(p)}\left[\ln\pi_k + \ln(g(y_i|\boldsymbol{\theta}_k))\right], \end{aligned} \tag{17}$$

where $\hat{w}_{ik}^{(p)}$ is the posterior probability that the observation y_i belongs to the mixture component k (also called the responsibility) and it is equal to

$$\begin{aligned} \hat{w}_{ik}^{(p)} = \mathbb{E}\left(Z_{ik}|y_i,\mathbf{\Phi}^{(p-1)}\right) = \mathbb{P}\left(Z_{ik}=1|y_i,\mathbf{\Phi}^{(p-1)}\right) = \frac{\pi_k^{(p-1)}g(y_i|\boldsymbol{\theta}_k^{(p-1)})}{f(y_i|\mathbf{\Phi}^{(p-1)})} = \\ = \frac{\pi_k^{(p-1)}g(y_i|\boldsymbol{\theta}_k^{(p-1)})}{\sum_{j=1}^{K}\pi_j^{(p-1)}g(y_i|\boldsymbol{\theta}_j^{(p-1)})}, \quad i=1,\dots,n, \quad k=1,\dots,K. \end{aligned}$$

M-step

The Q-function (17) is maximized to obtain estimates for parameters $\mathbf{\Phi}$. The estimates are derived from a system of $(d+1)K$ equations subject to that the mixing weights sum up to 1:

$$\begin{cases} \dfrac{\partial Q(\mathbf{\Phi}|\mathbf{\Phi}^{(p-1)})}{\partial\theta_{ks}} = 0, \quad k=1,\dots,K, \quad s=1,\dots,d, \\ \dfrac{\partial Q(\mathbf{\Phi}|\mathbf{\Phi}^{(p-1)})}{\partial\pi_k} = 0, \quad k=1,\dots,K, \quad \text{if} \quad \sum_{k=1}^{K}\pi_k = 1. \end{cases}$$

Because of the form of the Q-function the estimates of $\boldsymbol{\theta}_k$ can be found independently for each $k = 1,\dots,K$, additionally, without the condition for the mixing weights. Note that the exact estimates depend on the parametric family chosen for the component distribution. The step p estimates of the component parameters of normal mixtures are given in McLachlan and Peel (2000, p. 82) and for lognormal, inverse Gaussian, gamma, Burr, inverse Burr, Weibull, Rayleigh and truncated normal mixtures derived in Section 2.2.2.

The estimates of component weights $\boldsymbol{\pi}$ can be found subject to the equality constraint

$$\sum_{k=1}^{K}\pi_k = 1$$

independently from the estimates of the component parameters. For this the method of Lagrange multipliers (see, e.g., Fletcher, 1987, pp. 195–200) is applied. Following system of equations

$$\begin{cases} \dfrac{\partial \mathscr{L}(\pi_1,\ldots,\pi_k,\lambda)}{\partial \pi_k} = 0, \quad \lambda \neq 0, \quad k = 1,\ldots,K, \\ \sum\limits_{k=1}^{K} \pi_k = 1, \end{cases}$$

where

$$\mathscr{L}(\pi_1,\ldots,\pi_k,\lambda) = Q(\mathbf{\Phi}|\mathbf{\Phi}^{(p-1)}) - \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right)$$

is the Lagrange function, is solved by first finding the derivative,

$$\frac{\partial \mathscr{L}(\pi_1,\ldots,\pi_k,\lambda)}{\partial \pi_k} = \frac{\partial}{\partial \pi_k} \left[\sum_{i=1}^{n} \sum_{j=1}^{K} \hat{w}_{ij}^{(p)} \left[\ln(\pi_j) + \ln(g(y_i|\boldsymbol{\theta}_j)) \right] - \lambda \left(\sum_{j=1}^{K} \pi_j - 1 \right) \right] =$$

$$= \frac{\partial \ln \pi_k}{\partial \pi_k} \sum_{i=1}^{n} \hat{w}_{ik}^{(p)} - \lambda \frac{\partial \pi_k}{\partial \pi_k} = \frac{1}{\pi_k} \sum_{i=1}^{n} \hat{w}_{ik}^{(p)} - \lambda, \quad k = 1,\ldots,K.$$

The system of equations takes then the form

$$\begin{cases} \pi_k = \dfrac{1}{\lambda} \sum\limits_{i=1}^{n} \hat{w}_{ik}^{(p)}, \quad \lambda \neq 0, \quad k = 1,\ldots,K, \\ \sum\limits_{k=1}^{K} \pi_k = 1, \end{cases}$$

and thus from the following equalities

$$1 = \sum_{k=1}^{K} \pi_k = \frac{1}{\lambda} \sum_{k=1}^{K} \sum_{i=1}^{n} \hat{w}_{ik}^{(p)} = \frac{1}{\lambda} \sum_{i=1}^{n} \sum_{k=1}^{K} \hat{w}_{ik}^{(p)} = \frac{1}{\lambda} \sum_{i=1}^{n} \frac{1}{f(y_i|\mathbf{\Phi}^{(p-1)})} \sum_{k=1}^{K} \pi_k^{(p-1)} g(y_i|\boldsymbol{\theta}_k^{(p-1)}) = \frac{n}{\lambda}$$

it is obtained that $\lambda = n$. Thereby, estimates of the mixing weights on step p are

$$\hat{\pi}_k^{(p)} = \frac{1}{n} \sum_{i=1}^{n} \hat{w}_{ik}^{(p)}, \quad k = 1,\ldots,K. \tag{18}$$

The parameter estimates are updated by repeating E-steps and M-steps until the difference or relative difference between the values of log-likelihoods or likelihoods for consecutive steps is smaller than a pre-specified tolerance value (e.g., 10^{-8}).

There are several aspects that have to be kept in mind when using the EM algorithm for estimating parameters of K-component mixtures. More details are given in McLachlan and Peel (2000), Frühwirth-Schnatter (2006), McLachlan and Krishnan (2008), but some of the most important points to remember are:

- The likelihood function does not decrease with an iteration of the EM algorithm.
- The likelihood function value corresponding to the solution obtained with the EM algorithm is indeed under very weak conditions local maximum of the likelihood function and for finite mixtures often multiple local maxima exist.

- The likelihood function might not be bounded above while in most cases at least one local maximum exists.
- The solutions might be spurious i.e. some mixture components contain very few observations and might cause high likelihood values.

Thus, to avoid non-convergence of the expectation-maximization algorithm and spurious solutions, as also brought out by Miljkovic and Grün (2016), detail must be paid to the initialization of the algorithm as for most iteration methods and optimization problems. In some applications running the EM algorithm several times (up to 1000) with different initial values is applied to find the solution corresponding to the highest likelihood value. Miljkovic and Grün (2016) also use different methods to determine the initial partitioning of the data to give initial values to the EM algorithm in addition to disregarding initial partitioning if any of its partitions contains less than 1% of the observations.

2.2.2 Expectation-Maximization Algorithm for Specific Mixtures

As mentioned before the estimates of parameters $\boldsymbol{\theta}_k$ $(k = 1,\dots,K)$ on the iteration step p in the M-step of the EM algorithm for finding parameter estimates of K-component mixtures depend on the parametric family chosen for the component distribution. In the following form of the estimates of parameters $\boldsymbol{\theta}_k$ $(k = 1,\dots,K)$ of eight K-component mixtures on step p of the EM algorithm are derived. The closed-form solutions are without derivations given in various studies (e.g., Miljkovic and Grün, 2016) for lognormal, inverse Gaussian, gamma, Burr, inverse Burr and Weibull mixtures, but in this thesis they are given for clarity with author's derivations. In addition, estimates of parameters of Rayleigh and truncated normal mixtures are derived by the author.

Definitions of the eight distributions (lognormal, inverse Gaussian, gamma, Burr, inverse Burr, Weibull, Rayleigh and truncated normal mixtures) are given in Section 1.2. All mentioned distributions are non-negative and thus can be applied on wind speed data in the practical part of the thesis.

Lognormal Mixtures

The Q-function (17) that is maximized in the M-step of the EM algorithm used for estimating the parameters of a lognormal mixture (see Definition 4) takes the form

$$Q := Q(\boldsymbol{\Phi}|\boldsymbol{\Phi}^{(p-1)}) = \sum_{i=1}^{n}\sum_{k=1}^{K}\hat{w}_{ik}^{(p)}\left[\ln\pi_k + \ln\left(\frac{1}{y_i\sigma_k\sqrt{2\pi}}e^{-\frac{(\ln y_i-\mu_k)^2}{2\sigma_k^2}}\right)\right] =$$
$$= \sum_{i=1}^{n}\sum_{k=1}^{K}\hat{w}_{ik}^{(p)}\left(A_{ik} - \ln\sigma_k - \frac{(\ln y_i-\mu_k)^2}{2\sigma_k^2}\right),$$

where $A_{ik} = \ln\pi_k - \ln(y_i\sqrt{2\pi})$ $(i = 1,\dots,n;\ k = 1,\dots,K)$ does not contain parameters μ_k and σ_k. The component index k is fixed in the following as the estimates of parameters μ_k and σ_k can be found independently for each k.

Taking the derivative of the Q-function with respect to μ_k resolves in

$$\frac{\partial Q}{\partial \mu_k} = \frac{\partial}{\partial \mu_k}\left[\sum_{i=1}^{n}\sum_{j=1}^{K}\hat{w}_{ij}^{(p)}\left(A_{ij} - \ln\sigma_j - \frac{(\ln y_i - \mu_j)^2}{2\sigma_j^2}\right)\right] = -\frac{1}{2\sigma_k^2}\sum_{i=1}^{n}\hat{w}_{ik}^{(p)}\frac{\partial(\ln y_i - \mu_k)^2}{\partial \mu_k} =$$
$$= \frac{1}{\sigma_k^2}\sum_{i=1}^{n}\hat{w}_{ik}^{(p)}(\ln y_i - \mu_k) = \frac{1}{\sigma_k^2}\sum_{i=1}^{n}\hat{w}_{ik}^{(p)}\ln y_i - \frac{1}{\sigma_k^2}n\hat{\pi}_k^{(p)}\mu_k,$$

because following (18) it holds that

$$\sum_{i=1}^{n}\hat{w}_{ik}^{(p)} = n\hat{\pi}_k^{(p)}. \tag{19}$$

After equalizing the derivative with zero and multiplying it by $\sigma_k^2 > 0$ it is clear that

$$\sum_{i=1}^{n}\hat{w}_{ik}^{(p)}\ln y_i = n\hat{\pi}_k^{(p)}\mu_k,$$

and thus the estimate of the parameter μ_k on step p is

$$\hat{\mu}_k^{(p)} = \frac{1}{n\hat{\pi}_k^{(p)}}\sum_{i=1}^{n}\hat{w}_{ik}^{(p)}\ln y_i.$$

The derivative of the Q-function with respect to σ_k^2 after fixing μ_k is

$$\frac{\partial Q}{\partial \sigma_k^2} = \frac{\partial}{\partial \sigma_k^2}\left[\sum_{i=1}^{n}\sum_{j=1}^{K}\hat{w}_{ij}^{(p)}\left(A_{ij} - \ln\sigma_j - \frac{(\ln y_i - \mu_j)^2}{2\sigma_j^2}\right)\right] =$$
$$= -\sum_{i=1}^{n}\hat{w}_{ik}^{(p)}\left(\frac{\partial \ln\sigma_k}{\partial \sigma_k^2} + \frac{1}{2}(\ln y_i - \mu_k)^2\frac{\partial(1/\sigma_k^2)}{\partial \sigma_k^2}\right) = \sum_{i=1}^{n}\hat{w}_{ik}^{(p)}\left(\frac{(\ln y_i - \mu_k)^2}{2\sigma_k^4} - \frac{1}{2\sigma_k^2}\right) =$$
$$= \frac{1}{2\sigma_k^4}\sum_{i=1}^{n}\hat{w}_{ik}^{(p)}(\ln y_i - \mu_k)^2 - \frac{1}{2\sigma_k^2}n\hat{\pi}_k^{(p)}.$$

When the derivative is equalized with zero and multiplied by $2\sigma_k^4 > 0$ it takes the form

$$n\hat{\pi}_k^{(p)}\sigma_k^2 = \sum_{i=1}^{n}\hat{w}_{ik}^{(p)}(\ln y_i - \mu_k)^2$$

and thus estimate of the parameter σ_k^2 on step p is

$$(\hat{\sigma}_k^2)^{(p)} = \frac{1}{n\hat{\pi}_k^{(p)}}\sum_{i=1}^{n}\hat{w}_{ik}^{(p)}(\ln y_i - \hat{\mu}_k^{(p)})^2.$$

Inverse Gaussian Mixtures

The Q-function (17) that is maximized in the M-step of the EM algorithm used for estimating the parameters of an inverse Gaussian mixture (see Definition 5) is

$$Q := Q(\mathbf{\Phi}|\mathbf{\Phi}^{(p-1)}) = \sum_{i=1}^{n}\sum_{k=1}^{K}\hat{w}_{ik}^{(p)}\left[\ln\pi_k + \ln\left(\sqrt{\frac{\lambda_k}{2\pi y_i^3}}e^{-\frac{\lambda_k(y_i-\mu_k)^2}{2\mu_k^2 y_i}}\right)\right] =$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{K} \hat{w}_{ik}^{(p)} \left(A_{ik} + \frac{1}{2} \ln \lambda_k - \frac{\lambda_k (y_i - \mu_k)^2}{2\mu_k^2 y_i} \right),$$

where $A_{ik} = \ln \pi_k - \ln \sqrt{2\pi y_i^3}$ ($i = 1, \dots, n$; $k = 1, \dots, K$) does not depend on parameters μ_k and λ_k. The component index k is fixed in the following as the estimates of parameters μ_k and λ_k can be found independently for each k.

Taking the derivative of the Q-function with respect to μ_k resolves in

$$\frac{\partial Q}{\partial \mu_k} = \frac{\partial}{\partial \mu_k} \left[\sum_{i=1}^{n} \sum_{j=1}^{K} \hat{w}_{ij}^{(p)} \left(A_{ij} + \frac{1}{2} \ln \lambda_j - \frac{\lambda_j (y_i - \mu_j)^2}{2\mu_j^2 y_i} \right) \right] =$$
$$= -\frac{\lambda_k}{2} \sum_{i=1}^{n} \hat{w}_{ik}^{(p)} y_i^{-1} \frac{\partial}{\partial \mu_k} \left(\frac{y_i^2 - 2y_i \mu_k + \mu_k^2}{\mu_k^2} \right) = -\frac{\lambda_k}{2} \sum_{i=1}^{n} \hat{w}_{ik}^{(p)} \left(y_i \frac{\partial (1/\mu_k^2)}{\partial \mu_k} - \frac{\partial (2/\mu_k)}{\partial \mu_k} \right) =$$
$$= \lambda_k \sum_{i=1}^{n} \hat{w}_{ik}^{(p)} \left(\frac{y_i}{\mu_k^3} - \frac{1}{\mu_k^2} \right) = \frac{\lambda_k}{\mu_k^3} \sum_{i=1}^{n} \hat{w}_{ik}^{(p)} y_i - \frac{\lambda_k}{\mu_k^2} n \hat{\pi}_k^{(p)}.$$

After equalizing the derivative with zero and multiplying it by $\mu_k^3 / \lambda_k > 0$ it is clear that

$$\sum_{i=1}^{n} \hat{w}_{ik}^{(p)} y_i = n \hat{\pi}_k^{(p)} \mu_k, \tag{20}$$

and thus the estimate of the parameter μ_k on step p is

$$\hat{\mu}_k^{(p)} = \frac{1}{n \hat{\pi}_k^{(p)}} \sum_{i=1}^{n} \hat{w}_{ik}^{(p)} y_i.$$

The derivative of the Q-function with respect to λ_k after fixing μ_k is

$$\frac{\partial Q}{\partial \lambda_k} = \frac{\partial}{\partial \lambda_k} \left[\sum_{i=1}^{n} \sum_{j=1}^{K} \hat{w}_{ij}^{(p)} \left(A_{ij} + \frac{1}{2} \ln \lambda_j - \frac{\lambda_j (y_i - \mu_j)^2}{2\mu_j^2 y_i} \right) \right] =$$
$$= \frac{1}{2} \sum_{i=1}^{n} \hat{w}_{ik}^{(p)} \left(\frac{\partial \ln \lambda_k}{\partial \lambda_k} - \frac{(y_i - \mu_k)^2}{\mu_k^2 y_i} \right) = \frac{n \hat{\pi}_k^{(p)}}{2\lambda_k} - \frac{1}{2} b_{ik},$$

where, since (19) and (20),

$$b_{ik} = \sum_{i=1}^{n} \hat{w}_{ik}^{(p)} \frac{(y_i - \mu_k)^2}{\mu_k^2 y_i} = \sum_{i=1}^{n} \hat{w}_{ik}^{(p)} \left(\frac{y_i}{\mu_k^2} - \frac{2}{\mu_k} + \frac{1}{y_i} \right) =$$
$$= \frac{1}{\mu_k^2} \sum_{i=1}^{n} \hat{w}_{ik}^{(p)} y_i - \frac{2}{\mu_k} \sum_{i=1}^{n} \hat{w}_{ik}^{(p)} + \sum_{i=1}^{n} \hat{w}_{ik}^{(p)} \frac{1}{y_i} = -\frac{n \hat{\pi}_k^{(p)}}{\mu_k} + \sum_{i=1}^{n} \hat{w}_{ik}^{(p)} \frac{1}{y_i} = \sum_{i=1}^{n} \hat{w}_{ik}^{(p)} \left(\frac{1}{y_i} - \frac{1}{\mu_k} \right).$$

When the derivative is equalized with zero the estimate of parameter λ_k on step p is easily obtained as

$$\hat{\lambda}_k^{(p)} = \frac{n \hat{\pi}_k^{(p)}}{\sum_{i=1}^{n} \hat{w}_{ik}^{(p)} \left(\frac{1}{y_i} - \frac{1}{\mu_k^{(p)}} \right)}.$$

Gamma Mixtures

The Q-function (17) that is maximized in the M-step of the EM algorithm used for estimating the parameters of a gamma mixture (see Definition 6) takes the form

$$Q := Q(\mathbf{\Phi}|\mathbf{\Phi}^{(p-1)}) = \sum_{i=1}^{n}\sum_{k=1}^{K}\hat{w}_{ik}^{(p)}\left[\ln\pi_k + \ln\left(\frac{\lambda_k^{\alpha_k}y_i^{\alpha_k-1}}{\Gamma(\alpha_k)}e^{-y_i\lambda_k}\right)\right] =$$
$$= \sum_{i=1}^{n}\sum_{k=1}^{K}\hat{w}_{ik}^{(p)}\left(A_{ik} + \alpha_k\ln\lambda_k + \alpha_k\ln y_i - \ln\Gamma(\alpha_k) - y_i\lambda_k\right),$$

where Γ is the gamma function (12) and $A_{ik} = \ln\pi_k - \ln y_i$ ($i = 1,\ldots,n$; $k = 1,\ldots,K$) does not depend on parameters α_k and λ_k. The component index k is fixed in the following as the estimates of parameters α_k and λ_k can be found independently for each k.

The derivative of the Q-function with respect to λ_k after fixing α_k is

$$\frac{\partial Q}{\partial\lambda_k} = \frac{\partial}{\partial\lambda_k}\left[\sum_{i=1}^{n}\sum_{j=1}^{K}\hat{w}_{ij}^{(p)}\left(A_{ij} + \alpha_j\ln\lambda_j + \alpha_j\ln y_i - \ln\Gamma(\alpha_j) - y_i\lambda_j\right)\right] =$$
$$= \sum_{i=1}^{n}\hat{w}_{ik}^{(p)}\left(\alpha_k\frac{\partial\ln\lambda_k}{\partial\lambda_k} - y_i\frac{\partial\lambda_k}{\partial\lambda_k}\right) = \sum_{i=1}^{n}\hat{w}_{ik}^{(p)}\left(\frac{\alpha_k}{\lambda_k} - y_i\right) = \frac{\alpha_k}{\lambda_k}n\hat{\pi}_k^{(p)} - \sum_{i=1}^{n}\hat{w}_{ik}^{(p)}y_i,$$

which after being equalized with zero gives the parameter λ_k estimate on step p,

$$\hat{\lambda}_k^{(p)} = \frac{\hat{\alpha}_k^{(p)}n\hat{\pi}_k^{(p)}}{\sum\limits_{i=1}^{n}\hat{w}_{ik}^{(p)}y_i}.$$

Estimate of the parameter α_k can be found using numerical optimization as no closed-form solution exists to the equation $\partial Q/\partial\alpha_k = 0$ (with fixed λ_k), because

$$\frac{\partial Q}{\partial\alpha_k} = \frac{\partial}{\partial\alpha_k}\left[\sum_{i=1}^{n}\sum_{j=1}^{K}\hat{w}_{ik}^{(p)}\left(A_{ij} + \alpha_j\ln\lambda_j + \alpha_j\ln y_i - \ln\Gamma(\alpha_j) - y_i\lambda_j\right)\right] =$$
$$= \sum_{i=1}^{n}\hat{w}_{ik}^{(p)}\left((\ln\lambda_k + \ln y_i)\frac{\partial\alpha_k}{\partial\alpha_k} - \frac{\partial\ln\Gamma(\alpha_k)}{\partial\alpha_k}\right) = n\hat{\pi}_k^{(p)}(\ln\lambda_k - \psi(\alpha_k)) + \sum_{i=1}^{n}\hat{w}_{ik}^{(p)}\ln y_i,$$

where ψ is the di-gamma function (e.g., Forbes et al., 2011, p. 56)

$$\psi(x) = \frac{\partial\ln\Gamma(x)}{\partial x}, \quad x > 0,$$

and to where λ_k is inserted as

$$\lambda_k = \frac{\alpha_k n\hat{\pi}_k^{(p)}}{\sum_{i=1}^{n}\hat{w}_{ik}^{(p)}y_i}.$$

Burr Mixtures

The Q-function (17) that is maximized in the M-step of the EM algorithm used for estimating the parameters of a Burr mixture (see Definition 7) takes the form

$$Q := Q(\mathbf{\Phi}|\mathbf{\Phi}^{(p-1)}) = \sum_{i=1}^{n}\sum_{k=1}^{K}\hat{w}_{ik}^{(p)}\left[\ln\pi_k + \ln\left(\frac{\alpha_k\gamma_k y_i^{\gamma_k-1}}{\theta_k^{\gamma_k}\left(1+\frac{y_i^{\gamma_k}}{\theta_k^{\gamma_k}}\right)^{\alpha_k+1}}\right)\right] =$$

$$= \sum_{i=1}^{n}\sum_{k=1}^{K}\hat{w}_{ik}^{(p)}\left(\ln\pi_k + \ln\alpha_k + \ln\left(\frac{\gamma_k y_i^{\gamma_k-1}}{\theta_k^{\gamma_k}}\right) - (\alpha_k+1)\ln\left(1+\frac{y_i^{\gamma_k}}{\theta_k^{\gamma_k}}\right)\right),$$

where $\ln\pi_k$ ($k=1,\ldots,K$) does not depend on parameters α_k, γ_k and θ_k. The component index k is fixed in the following as the estimates of parameters α_k, γ_k and θ_k can be found independently for each k.

The derivative of the Q-function with respect to α_k after fixing γ_k and θ_k is

$$\frac{\partial Q}{\partial\alpha_k} = \frac{\partial}{\partial\alpha_k}\left[\sum_{i=1}^{n}\sum_{j=1}^{K}\hat{w}_{ij}^{(p)}\left(\ln\alpha_j - (\alpha_j+1)\ln\left(1+\frac{y_i^{\gamma_j}}{\theta_j^{\gamma_j}}\right)\right)\right] =$$

$$= \sum_{i=1}^{n}\hat{w}_{ik}^{(p)}\left(\frac{\partial\ln\alpha_k}{\partial\alpha_k} - \ln\left(1+\frac{y_i^{\gamma_k}}{\theta_k^{\gamma_k}}\right)\right) = \frac{1}{\alpha_k}n\hat{\pi}_k^{(p)} - \sum_{i=1}^{n}\hat{w}_{ik}^{(p)}\ln\left(1+\frac{y_i^{\gamma_k}}{\theta_k^{\gamma_k}}\right).$$

After equalizing the derivative with zero the estimate of parameter α_k on step p is obtained,

$$\hat{\alpha}_k^{(p)} = \frac{n\hat{\pi}_k^{(p)}}{\sum_{i=1}^{n}\hat{w}_{ik}^{(p)}\ln\left(1+\left(\frac{y_i}{\hat{\theta}_k^{(p)}}\right)^{\hat{\gamma}_k^{(p)}}\right)}.$$

Estimates of parameters γ_k and θ_k can be found using numerical optimization as no closed-form solution exists to the system of two equations $\partial Q/\partial\gamma_k = 0$ and $\partial Q/\partial\theta_k = 0$, to where α_k is inserted as

$$\alpha_k = \frac{n\hat{\pi}_k^{(p)}}{\sum_{i=1}^{n}\hat{w}_{ik}^{(p)}\ln\left(1+\left(\frac{y_i}{\theta_k}\right)^{\gamma_k}\right)},$$

because

$$\frac{\partial Q}{\partial\gamma_k} = \frac{\partial}{\partial\gamma_k}\left[\sum_{i=1}^{n}\sum_{j=1}^{K}\hat{w}_{ij}^{(p)}\left(\ln\left(\frac{\gamma_j y_i^{\gamma_j-1}}{\theta_j^{\gamma_j}}\right) - (\alpha_j+1)\ln\left(1+\frac{y_i^{\gamma_j}}{\theta_j^{\gamma_j}}\right)\right)\right] =$$

$$= \sum_{i=1}^{n}\hat{w}_{ik}^{(p)}\frac{\partial}{\partial\gamma_k}\left(\ln\gamma_k + \gamma_k\ln\frac{y_i}{\theta_k} - (\alpha_k+1)\ln\left(1+\frac{y_i^{\gamma_k}}{\theta_k^{\gamma_k}}\right)\right) =$$

$$= \sum_{i=1}^{n}\hat{w}_{ik}^{(p)}\left(\frac{1}{\gamma_k} + \ln\frac{y_i}{\theta_k} - \frac{\alpha_k+1}{1+\frac{y_i^{\gamma_k}}{\theta_k^{\gamma_k}}}\frac{\partial\,(y_i/\theta_k)^{\gamma_k}}{\partial\gamma_k}\right) =$$

$$= \sum_{i=1}^{n} \hat{w}_{ik}^{(p)} \left(\frac{1}{\gamma_k} + \ln \frac{y_i}{\theta_k} - \frac{\alpha_k + 1}{\theta_k^{\gamma_k} + y_i^{\gamma_k}} y_i^{\gamma_k} \ln \frac{y_i}{\theta_k} \right) =$$

$$= \frac{n\hat{\pi}_k^{(p)}}{\gamma_k} + \sum_{i=1}^{n} \hat{w}_{ik}^{(p)} \left(1 - \frac{\alpha_k + 1}{\theta_k^{\gamma_k} + y_i^{\gamma_k}} y_i^{\gamma_k} \right) \ln \frac{y_i}{\theta_k}$$

and

$$\frac{\partial Q}{\partial \theta_k} = \frac{\partial}{\partial \theta_k} \left[\sum_{i=1}^{n} \sum_{j=1}^{K} \hat{w}_{ij}^{(p)} \left(-\gamma_j \ln \theta_j - (\alpha_j + 1) \ln \left(1 + \frac{y_i^{\gamma_j}}{\theta_j^{\gamma_j}} \right) \right) \right] =$$

$$= - \sum_{i=1}^{n} \hat{w}_{ik}^{(p)} \left(\gamma_k \frac{\partial \ln \theta_k}{\partial \theta_k} + (\alpha_k + 1) \frac{\partial}{\partial \theta_k} \ln \left(1 + \frac{y_i^{\gamma_k}}{\theta_k^{\gamma_k}} \right) \right) =$$

$$= - \sum_{i=1}^{n} \hat{w}_{ik}^{(p)} \left(\frac{\gamma_k}{\theta_k} + (\alpha_k + 1) \frac{y_i^{\gamma_k}}{1 + \frac{y_i^{\gamma_k}}{\theta_k^{\gamma_k}}} \frac{\partial (1/\theta_k^{\gamma_k})}{\partial \theta_k} \right) =$$

$$= \sum_{i=1}^{n} \hat{w}_{ik}^{(p)} \left(\frac{(\alpha_k + 1) y_i^{\gamma_k} \theta_k^{\gamma_k} \gamma_k \theta_k^{-\gamma_k - 1}}{y_i^{\gamma_k} + \theta_k^{\gamma_k}} - \frac{\gamma_k}{\theta_k} \right) = \frac{\gamma_k}{\theta_k} \sum_{i=1}^{n} \hat{w}_{ik}^{(p)} \left(\frac{y_i^{\gamma_k} (\alpha_k + 1)}{y_i^{\gamma_k} + \theta_k^{\gamma_k}} - 1 \right) =$$

$$= \frac{\gamma_k (\alpha_k + 1)}{\theta_k} \sum_{i=1}^{n} \frac{\hat{w}_{ik}^{(p)}}{\left(\frac{\theta_k}{y_i} \right)^{\gamma_k} + 1} - \frac{\gamma_k}{\theta_k} n \hat{\pi}_k^{(p)}.$$

Inverse Burr Mixtures

The Q-function (17) that is maximized in the M-step of the EM algorithm used for estimating the parameters of an inverse Burr mixture (see Definition 8) takes the form

$$Q := Q(\mathbf{\Phi} | \mathbf{\Phi}^{(p-1)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \hat{w}_{ik}^{(p)} \left[\ln \pi_k + \ln \left(\frac{\tau_k \gamma_k y_i^{\tau_k \gamma_k - 1}}{\theta_k^{\tau_k \gamma_k} \left(1 + \frac{y_i^{\gamma_k}}{\theta_k^{\gamma_k}} \right)^{\tau_k + 1}} \right) \right] =$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{K} \hat{w}_{ik}^{(p)} \left(\ln \frac{\pi_k \gamma_k}{y_i} + \ln \tau_k + \tau_k \gamma_k \ln \frac{y_i}{\theta_k} - (\tau_k + 1) \ln \left(1 + \frac{y_i^{\gamma_k}}{\theta_k^{\gamma_k}} \right) \right),$$

where $\ln \pi_k$ ($k = 1, \ldots, K$) does not depend on parameters τ_k, γ_k and θ_k. The component index k is fixed in the following as the estimates of parameters τ_k, γ_k and θ_k can be found independently for each k.

The derivative of the Q-function with respect to τ_k after fixing γ_k and θ_k is

$$\frac{\partial Q}{\partial \tau_k} = \frac{\partial}{\partial \tau_k} \left[\sum_{i=1}^{n} \sum_{j=1}^{K} \hat{w}_{ij}^{(p)} \left(\ln \tau_j + \tau_j \gamma_j \ln \frac{y_i}{\theta_j} - (\tau_j + 1) \ln \left(1 + \frac{y_i^{\gamma_j}}{\theta_j^{\gamma_j}} \right) \right) \right] =$$

$$= \sum_{i=1}^{n} \hat{w}_{ik}^{(p)} \left(\frac{1}{\tau_k} + \gamma_k \ln \frac{y_i}{\theta_k} - \ln \left(1 + \frac{y_i^{\gamma_k}}{\theta_k^{\gamma_k}} \right) \right) = \frac{n\hat{\pi}_k^{(p)}}{\tau_k} - \sum_{i=1}^{n} \hat{w}_{ik}^{(p)} \ln \frac{\left(1 + \frac{y_i^{\gamma_k}}{\theta_k^{\gamma_k}} \right) \theta_k^{\gamma_k}}{y_i^{\gamma_k}} =$$

$$= \frac{n\hat{\pi}_k^{(p)}}{\tau_k} - \sum_{i=1}^{n} \hat{w}_{ik}^{(p)} \ln\left(1 + \left(\frac{\theta_k}{y_i}\right)^{\gamma_k}\right).$$

When the derivative is equalized with zero the estimate of parameter τ_k on step p is obtained,

$$\hat{\tau}_k^{(p)} = \frac{n\hat{\pi}_k^{(p)}}{\sum_{i=1}^{n} \hat{w}_{ik}^{(p)} \ln\left(1 + \left(\frac{\hat{\theta}_k^{(p)}}{y_i}\right)^{\hat{\gamma}_k^{(p)}}\right)}.$$

Estimates of parameters γ_k and θ_k can be found using numerical optimization as no closed-form solution exists to the system of two equations $\partial Q/\partial \gamma_k = 0$ and $\partial Q/\partial \theta_k = 0$ to where τ_k is inserted as

$$\tau_k = \frac{n\hat{\pi}_k^{(p)}}{\sum_{i=1}^{n} \hat{w}_{ik}^{(p)} \ln\left(1 + \left(\frac{\theta_k}{y_i}\right)^{\gamma_k}\right)},$$

because

$$\frac{\partial Q}{\partial \gamma_k} = \frac{\partial}{\partial \gamma_k}\left[\sum_{i=1}^{n}\sum_{j=1}^{K} \hat{w}_{ij}^{(p)} \left(\ln \gamma_j + \tau_j \gamma_j \ln \frac{y_i}{\theta_j} - (\tau_j + 1)\ln\left(1 + \frac{y_i^{\gamma_j}}{\theta_j^{\gamma_j}}\right)\right)\right] =$$
$$= \sum_{i=1}^{n} \hat{w}_{ik}^{(p)} \left(\frac{1}{\gamma_k} + \tau_k \ln \frac{y_i}{\theta_k} - \frac{(\tau_k+1) y_i^{\gamma_k}}{(\theta_k^{\gamma_k} + y_i^{\gamma_k})} \ln \frac{y_i}{\theta_k}\right) =$$
$$= \frac{n\hat{\pi}_k^{(p)}}{\gamma_k} + \sum_{i=1}^{n} \hat{w}_{ik}^{(p)} \ln \frac{y_i}{\theta_k}\left(\tau_k - \frac{(\tau_k+1) y_i^{\gamma_k}}{(\theta_k^{\gamma_k} + y_i^{\gamma_k})}\right)$$

and

$$\frac{\partial Q}{\partial \theta_k} = \frac{\partial}{\partial \theta_k}\left[\sum_{i=1}^{n}\sum_{j=1}^{K} \hat{w}_{ij}^{(p)} \left(-\tau_j \gamma_j \ln \theta_j - (\tau_j + 1)\ln\left(1 + \frac{y_i^{\gamma_j}}{\theta_j^{\gamma_j}}\right)\right)\right] =$$
$$= -\frac{\gamma_k \tau_k}{\theta_k} n\hat{\pi}_k^{(p)} + \frac{\gamma_k(\tau_k + 1)}{\theta_k} \sum_{i=1}^{n} \frac{\hat{w}_{ik}^{(p)}}{\left(\frac{\theta_k}{y_i}\right)^{\gamma_k} + 1},$$

analogously to the derivations for the Burr mixture on page 25.

Weibull Mixtures

The Q-function (17) that is maximized in the M-step of the EM algorithm used for estimating the parameters of a Weibull mixture (see Definition 9) takes the form

$$Q := Q(\mathbf{\Phi}|\mathbf{\Phi}^{(p-1)}) = \sum_{i=1}^{n}\sum_{k=1}^{K} \hat{w}_{ik}^{(p)} \left[\ln \pi_k + \ln\left(\frac{\alpha_k y_i^{\alpha_k - 1}}{\theta_k^{\alpha_k}} \exp\left(-\left(\frac{y_i}{\theta_k}\right)^{\alpha_k}\right)\right)\right] =$$
$$= \sum_{i=1}^{n}\sum_{k=1}^{K} \hat{w}_{ik}^{(p)} \left(A_{ik} + \ln \alpha_k + \alpha_k \ln y_i - \alpha_k \ln \theta_k - \left(\frac{y_i}{\theta_k}\right)^{\alpha_k}\right),$$

where $A_{ik} = \ln \pi_k - \ln y_i$ ($i = 1, \ldots, n$; $k = 1, \ldots, K$) does not depend on parameters α_k and θ_k. The component index k is fixed in the following as the estimates of parameters α_k and θ_k can be found independently for each k.

Estimates of parameters α_k and θ_k can be found using numerical optimization as no closed-form solution exists to the system of two equations $\partial Q/\partial \alpha_k = 0$ and $\partial Q/\partial \theta_k = 0$, because

$$\frac{\partial Q}{\partial \alpha_k} = \frac{\partial}{\partial \alpha_k}\left[\sum_{i=1}^{n}\sum_{j=1}^{K}\hat{w}_{ij}^{(p)}\left(A_{ij} + \ln \alpha_j + \alpha_j \ln y_i - \alpha_j \ln \theta_j - \left(\frac{y_i}{\theta_j}\right)^{\alpha_j}\right)\right] =$$
$$= \sum_{i=1}^{n}\hat{w}_{ik}^{(p)}\left(\frac{1}{\alpha_k} + \ln \frac{y_i}{\theta_k} - \left(\frac{y_i}{\theta_k}\right)^{\alpha_k} \ln \frac{y_i}{\theta_k}\right) = \frac{n\hat{\pi}_k^{(p)}}{\alpha_k} + \sum_{i=1}^{n}\hat{w}_{ik}^{(p)} \ln \frac{y_i}{\theta_k}\left(1 - \left(\frac{y_i}{\theta_k}\right)^{\alpha_k}\right)$$

and

$$\frac{\partial Q}{\partial \theta_k} = \frac{\partial}{\partial \theta_k}\left[\sum_{i=1}^{n}\sum_{j=1}^{K}\hat{w}_{ij}^{(p)}\left(\ln \pi_j + \ln \alpha_j + (\alpha_j - 1)\ln y_i - \alpha_j \ln \theta_j - \left(\frac{y_i}{\theta_j}\right)^{\alpha_j}\right)\right] =$$
$$= -\sum_{i=1}^{n}\hat{w}_{ik}^{(p)}\left(\frac{\alpha_k}{\theta_k} + y_i^{\alpha_k}\frac{\partial \theta_k^{-\alpha_k}}{\partial \theta_k}\right) = -\sum_{i=1}^{n}\hat{w}_{ik}^{(p)}\left(\frac{\alpha_k}{\theta_k} - y_i^{\alpha_k}\alpha_k\theta_k^{-\alpha_k - 1}\right) =$$
$$= -\frac{\alpha_k}{\theta_k}n\hat{\pi}_k^{(p)} + \frac{\alpha_k}{\theta_k^{\alpha_k+1}}\sum_{i=1}^{n}\hat{w}_{ik}^{(p)}y_i^{\alpha_k}.$$

Rayleigh Mixtures

The Q-function (17) that is maximized in the M-step of the EM algorithm used for estimating the parameters of a Rayleigh mixture (see Definition 10) takes the form

$$Q := Q(\mathbf{\Phi}|\mathbf{\Phi}^{(p-1)}) = \sum_{i=1}^{n}\sum_{k=1}^{K}\hat{w}_{ik}^{(p)}\left[\ln \pi_k + \ln\left(\frac{y_i}{\sigma_k^2}e^{-\frac{y_i^2}{2\sigma_k^2}}\right)\right] =$$
$$= \sum_{i=1}^{n}\sum_{k=1}^{K}\hat{w}_{ik}^{(p)}\left(A_{ik} - \ln \sigma_k^2 - \frac{y_i^2}{2\sigma_k^2}\right),$$

where $A_{ik} = \ln \pi_k + \ln y_i$ ($i = 1, \ldots, n$; $k = 1, \ldots, K$) does not contain parameter σ_k. The component index k is fixed in the following as the estimate of parameter σ_k can be found independently for each k.

Taking the derivative of the Q-function with respect to σ_k^2 resolves in

$$\frac{\partial Q}{\partial \sigma_k^2} = \frac{\partial}{\partial \sigma_k^2}\left[\sum_{i=1}^{n}\sum_{j=1}^{K}\hat{w}_{ij}^{(p)}\left(A_{ij} - \ln \sigma_j^2 - \frac{y_i^2}{2\sigma_j^2}\right)\right] = \sum_{i=1}^{n}\hat{w}_{ik}^{(p)}\left(\frac{y_i^2}{2\sigma_k^4} - \frac{1}{\sigma_k^2}\right) =$$
$$= \frac{1}{2\sigma_k^4}\sum_{i=1}^{n}\hat{w}_{ik}^{(p)}y_i^2 - \frac{n\hat{\pi}_k^{(p)}}{\sigma_k^2}.$$

After equalizing the derivative with zero and multiplying it by $2\sigma_k^4 > 0$ it is clear that

$$\sum_{i=1}^{n}\hat{w}_{ik}^{(p)}y_i^2 = 2n\hat{\pi}_k^{(p)}\sigma_k^2$$

and thus the estimate of the parameter σ_k^2 on step p is

$$(\hat{\sigma}_k^2)^{(p)} = \frac{1}{2n\hat{\pi}_k^{(p)}} \sum_{i=1}^{n} \hat{w}_{ik}^{(p)} y_i^2.$$

Truncated Normal Mixtures

The Q-function (17) that is maximized in the M-step of the EM algorithm used for estimating the parameters of a truncated normal mixture (see Definition 3) takes the form

$$Q := Q(\boldsymbol{\Phi}|\boldsymbol{\Phi}^{(p-1)}) = \sum_{i=1}^{n}\sum_{k=1}^{K} \hat{w}_{ik}^{(p)} \left[\ln \pi_k + \ln \left(\frac{1}{\sqrt{2\pi}\sigma_k \Phi\left(\frac{\mu_k}{\sigma_k}\right)} e^{-\frac{(y-\mu_k)^2}{2\sigma_k^2}} \right) \right] =$$
$$= \sum_{i=1}^{n}\sum_{k=1}^{K} \hat{w}_{ik}^{(p)} \left(A_{ik} - \ln \sigma_k - \ln \left(\Phi \left(\frac{\mu_k}{\sigma_k} \right) \right) - \frac{(y_i - \mu_k)^2}{2\sigma_k^2} \right),$$

where Φ is the standard normal distribution function (11) and $A_{ik} = \ln \pi_k - \ln \sqrt{2\pi}$ ($i = 1, \ldots, n$; $k = 1, \ldots, K$) does not contain parameters μ_k and σ_k. The component index k is fixed in the following as the estimates of parameters μ_k and σ_k can be found independently for each k.

Estimates of parameters μ_k and σ_k can be found using numerical optimization as no closed-form solution exists to the system of two equations $\partial Q/\partial \mu_k = 0$ and $\partial Q/\partial \sigma_k = 0$, because

$$\frac{\partial Q}{\partial \mu_k} = \frac{\partial}{\partial \mu_k} \left[\sum_{i=1}^{n}\sum_{j=1}^{K} \hat{w}_{ij}^{(p)} \left(A_{ij} - \ln \sigma_j - \ln \left(\Phi \left(\frac{\mu_j}{\sigma_j} \right) \right) - \frac{(y_i - \mu_j)^2}{2\sigma_j^2} \right) \right] =$$
$$= \sum_{i=1}^{n} \hat{w}_{ik}^{(p)} \left(\frac{y_i - \mu_k}{\sigma_k^2} - \frac{\phi\left(\frac{\mu_k}{\sigma_k}\right)}{\sigma_k \Phi\left(\frac{\mu_k}{\sigma_k}\right)} \right) = \frac{1}{\sigma_k^2} \sum_{i=1}^{n} \hat{w}_{ik}^{(p)} y_i - \frac{\mu_k n \hat{\pi}_k^{(p)}}{\sigma_k^2} - \frac{n\hat{\pi}_k^{(p)} \phi\left(\frac{\mu_k}{\sigma_k}\right)}{\sigma_k \Phi\left(\frac{\mu_k}{\sigma_k}\right)}$$

and

$$\frac{\partial Q}{\partial \sigma_k} = \frac{\partial}{\partial \sigma_k} \left[\sum_{i=1}^{n}\sum_{j=1}^{K} \hat{w}_{ij}^{(p)} \left(A_{ij} - \ln \sigma_j - \ln \left(\Phi \left(\frac{\mu_j}{\sigma_j} \right) \right) - \frac{(y_i - \mu_j)^2}{2\sigma_j^2} \right) \right] =$$
$$= \sum_{i=1}^{n} \hat{w}_{ik}^{(p)} \left(\frac{\mu_k \phi\left(\frac{\mu_k}{\sigma_k}\right)}{\sigma_k^2 \Phi\left(\frac{\mu_k}{\sigma_k}\right)} + \frac{(y_i - \mu_k)^2}{\sigma_k^3} - \frac{1}{\sigma_k} \right) =$$
$$= \frac{n\hat{\pi}_k^{(p)} \mu_k \phi\left(\frac{\mu_k}{\sigma_k}\right)}{\sigma_k^2 \Phi\left(\frac{\mu_k}{\sigma_k}\right)} + \frac{1}{\sigma_k^3} \sum_{i=1}^{n} \hat{w}_{ik}^{(p)} y_i^2 - \frac{2\mu_k}{\sigma_k^3} \sum_{i=1}^{n} \hat{w}_{ik}^{(p)} y_i + \frac{n\hat{\pi}_k^{(p)} \mu_k^2}{\sigma_k^2} - \frac{n\hat{\pi}_k^{(p)}}{\sigma_k},$$

where ϕ is the standard normal density (e.g., Casella and Berger, 2002, p. 53)

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad x \in \mathbb{R}.$$

2.3 Model Selection and Goodness of Fit

After fitting multiple standard mixture models a natural question that arises is which is the best model, that is, which mixture fits the data best. Fitted mixtures with component densities from the same parametric family and the same number of components, as mixtures obtained after multiple runs of the EM algorithm with different initial values, can be compared based on the value of the model likelihood $L(\hat{\mathbf{\Phi}})$ in (16), where $\hat{\mathbf{\Phi}}$ is the maximum likelihood estimator of the parameters $\mathbf{\Phi}$ for the model (obtained for example using the EM algorithm), or based on the value of the model log-likelihood

$$l(\hat{\mathbf{\Phi}}) = \ln L(\hat{\mathbf{\Phi}}) = \sum_{i=1}^{n} \ln \left(\sum_{k=1}^{K} \hat{\pi}_k g(y_i|\hat{\boldsymbol{\theta}}_k) \right). \tag{21}$$

Additionally, following criteria are commonly in use for comparing the models,

- negative log-likelihood (NLL) (e.g., Casella and Berger, 2002, p. 485),

$$\mathrm{NLL} = -l(\hat{\mathbf{\Phi}}), \tag{22}$$

- Akaike information criterion (AIC) (e.g., Frühwirth-Schnatter, 2006, p. 116),

$$\mathrm{AIC} = -2l(\hat{\mathbf{\Phi}}) + 2P = -2l(\hat{\mathbf{\Phi}}) + 2(d+1)K - 2, \tag{23}$$

  where $P = (d+1)K - 1$ is the number of free parameters,
- Bayesian information criterion (BIC) (e.g., Frühwirth-Schnatter, 2006, p. 116),

$$\mathrm{BIC} = -2l(\hat{\mathbf{\Phi}}) + P\ln n = -2l(\hat{\mathbf{\Phi}}) + (d+1)K\ln n - \ln n. \tag{24}$$

For comparing the fitted models with different number of components, it is recommended to use BIC which also takes into account the number of the free parameters and the sample size n (see also Section 2.1 and Frühwirth-Schnatter (2006, pp. 116–117)).

For checking the goodness of fit of the model on the data in time numerous methods have been developed. In this thesis two graphical methods, Q-Q-plot and P-P-plot, are used to evaluate the goodness of fit of the K-component mixtures on the data. The Q-Q-plot and P-P-plot are described, for example, in Gibbons and Chakraborti (2003, pp. 143–145). For the goodness of fit problem on the Q-Q-plot or the quantile versus quantile plot the theoretical quantiles (i.e. the quantiles of the fitted distribution) are plotted against the empirical quantiles (i.e. the ordered sample). The P-P-plot or the probability versus probability plot is constructed by plotting the fitted distribution function F versus the empirical distribution function F_n. For both methods a resulting exact 45° line on the plot suggests a perfect fit. The mixture quantiles can be found from the Equation (10).

For more formal approach, for example, the Kolmogorov-Smirnov or other Cramer-von Mises type of goodness of fit tests can be used (e.g., Gibbons and Chakraborti, 2003, p. 111–112, 152). Note that to find the p-values of the tests bootstrap methods should be applied when the mixture parameters are estimated from the same sample (e.g., Babu and Rao, 2004).

3 Application on Estonian Meteorological Data

In this section K-component mixtures are applied on Estonian meteorological data, more precisely on daily maximal hourly mean wind speed and daily mean temperature data.

The data that support the findings of this thesis are available from Estonian Weather Service[2] of Estonian Environment Agency[3]. Restrictions apply to the availability of these data, which were used under license for the current thesis, and so are not publicly available. Data are however available for the author upon reasonable request and with permission of Estonian Weather Service of Estonian Environment Agency.

The data provided by the Estonian Weather Service contain hourly average and maximum wind speeds (m/s) and hourly average temperatures (°C). The data values are measured with accuracy of one decimal place (Estonian Weather Service, 2019a). There are number of meteorological stations in Estonia. For this thesis Narva, Ristna, Tallinn-Harku, Tõravere, Vilsandi and Võru stations were selected. For Ristna, Tallinn-Harku, Tõravere, Vilsandi and Võru stations the meteorological variables were provided describing 15 years (from 01.01.2004 to 31.12.2018, total of 5479 days or 131496 hours). For Narva the data was provided for two weather stations, because the station was relocated from Narva-Jõesuu to Narva. For Narva-Jõesuu the data was available from 01.01.2004 to 19.12.2013 and for Narva from 19.12.2013 to 31.12.2018. All data was provided in a XLSX file containing 122716 rows and 22 columns in one sheet and 17546 rows and 22 columns in the second sheet. For further analysis the data was exported to R (R Core Team, 2018) and all data analysis and numerical results presented in this thesis are obtained using R.

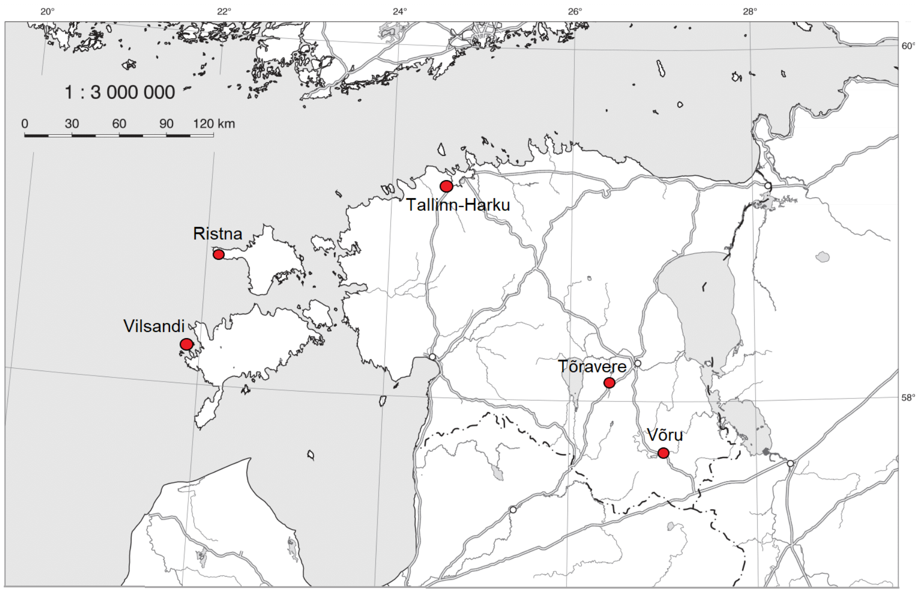


Figure 8: Approximate locations of the weather stations (base map ©Regio, 2005)

[2]https://www.ilmateenistus.ee/?lang=en

[3]https://www.keskkonnaagentuur.ee/en

From the available weather stations five stations Ristna, Tallinn-Harku, Tõravere, Vilsandi and Võru were chosen to carry out the further analysis (see Figure 8), because of the relocation of Narva station. The choice of the stations is also reasoned by their diverse locations. Two of the stations, Ristna (58°55′15″N 22°03′59″E) and Vilsandi (58°22′58″N 21°48′51″E), are located on islands in the Baltic Sea in Western Estonia, Tallinn-Harku (59°23′53″N 24°36′10″E) weather station lies in the Northern Estonia near the coastline, Tõravere (58°15′51″N 26°27′41″E) and Võru (57°50′47″N 27°01′10″E) stations are inlands in the southern part of Estonia (Estonian Weather Service, 2019b).

3.1 Estonian Wind Data

The summary statistics of the hourly mean wind speed and hourly maximal wind speed for Ristna, Tallinn-Harku, Tõravere, Vilsandi and Võru stations are presented in Appendix 1. Histograms of the same hourly wind characteristics are given in Appendix 2. Comparing the summary tables and histograms for the stations it can be said that during 2004–2018 greater (over 12 m/s) wind speeds were more frequent in the Ristna and Vilsandi stations which are located on the islands than in the other three weather stations. It can be seen from Table A1 and Table A2 that there are missing values for both raw variables for some hours for all weather stations, but the percentage of missing values does not exceed 0.30%.

For this thesis a new daily characteristic, daily maximal hourly mean wind speed, was created from the hourly mean wind speed. It was found for days for which at least one hourly mean wind speed was available. Daily maximal hourly mean wind speed may be described as the variable showing presence (or no presence) of strong winds lasting for a longer period (one hour) during the day. The daily variable was chosen (over the hourly variable) to speed up the analysis.

Presence of an overall trend during the 15 years in the daily maximal hourly mean wind speed was checked graphically for all five stations and visually none was detected. The summary statistics of daily maximal hourly mean wind speed are presented in Table 1 by stations. For illustration the histograms of the characteristic by stations are given in Figure 9. Based on Table 1 the daily variable is available for all 5479 days for all stations except for Ristna station, where for days 13.09.2015 and 27.09.2015 no wind data are present. For more insight on variation, the boxplots for all stations are presented in Figure 10, where symbol $\diamond$ denotes the data mean.

Table 1: Summary statistics for daily maximal hourly mean wind speeds (m/s)

Station	n	nmiss	Minimum	Q1	Median	Mean	Q3	Maximum	SD
Ristna	5477	2	0.7	3.1	4.3	5.16	6.5	18.9	2.794
Tallinn-Harku	5479	0	0.0	3.7	4.7	4.87	5.9	11.9	1.618
Tõravere	5479	0	0.1	3.0	3.8	3.92	4.8	10.9	1.348
Vilsandi	5479	0	2.2	5.8	7.7	8.22	10.3	22.6	3.262
Võru	5479	0	0.1	2.9	3.8	3.93	4.8	10.1	1.376

n: number of values, nmiss: number of missing values, Q1: lower quartile, Q3: upper quartile, SD: standard deviation

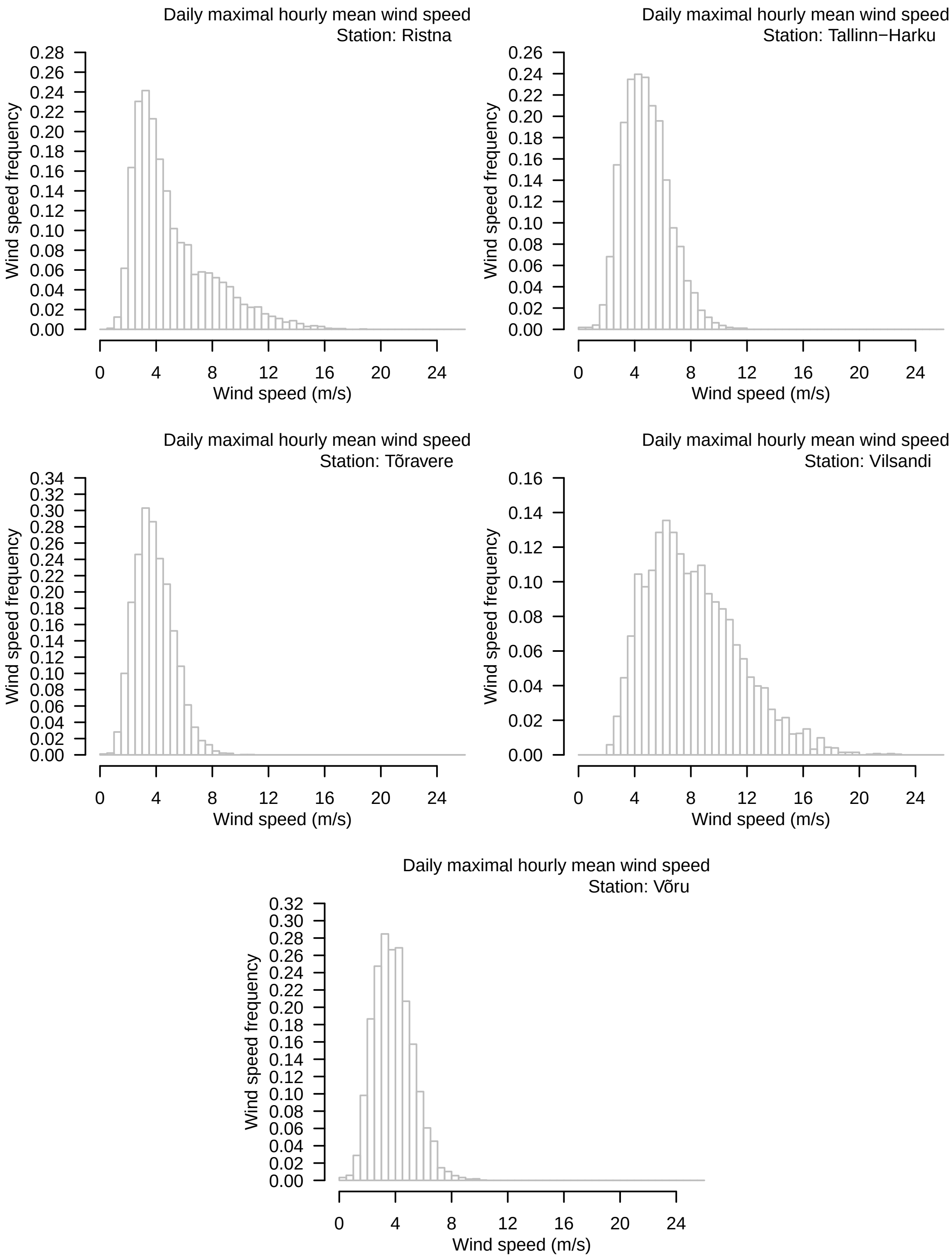


Figure 9: Histograms of daily maximal hourly mean wind speeds (m/s) by stations

The daily maximal hourly mean wind speed is greater for Vilsandi station compared to the other stations. From Table 1 it can be seen that during 2004–2018 for at least half of the days it was greater than 7.7 m/s in Vilsandi station while for other stations it stayed below 5 m/s for half of the days. In Figure 9 the distributions of daily maximal hourly mean wind speeds for Tõravere and Võru stations are rather similar to each other, but the other distributions

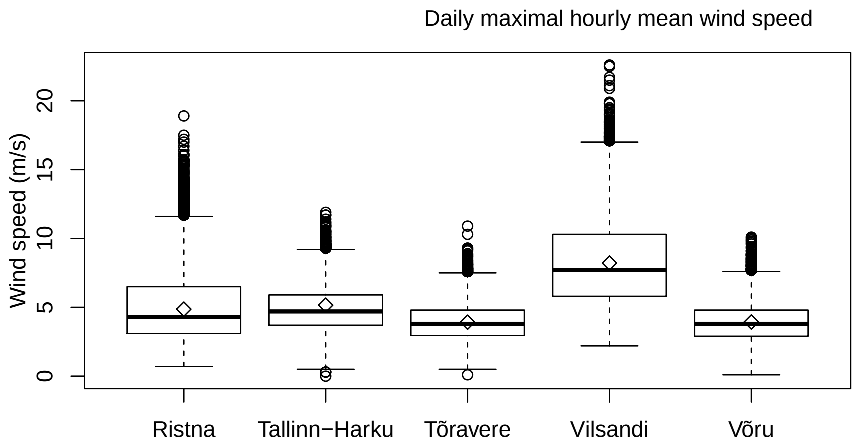


Figure 10: Boxplots of daily maximal hourly mean wind speeds (m/s) by stations

seem to be more or less different, for example, by comparing the quartiles and mean in Table 1 and the boxplots in Figure 10.

Based on Table 1 the maximum of daily maximal hourly mean wind speeds for Vilsandi station is 22.6 m/s from 22.12.2004, for Võru station it is 10.1 m/s and for Ristna station 18.9 m/s from 09.01.2005, for Tallinn-Harku station the maximum is 11.9 m/s from 13.12.2012 and for Tõravere station 10.9 m/s from 13.02.2005. The daily maximal of hourly mean wind speed variates the most in Vilsandi and Ristna stations.

3.2 Estonian Temperature Data

The summary statistics of the hourly mean temperature for Ristna, Tallinn-Harku, Tõravere, Vilsandi and Võru stations are presented in Appendix 1 in Table A3. Histograms of the hourly mean temperatures are given in Appendix 2 in Figure A3. It can be seen from Table A3 that there are missing values for the raw variable for some hours for all weather stations, but the percentage of missing values does not exceed 2.2%.

For this thesis a new daily variable, daily mean temperature, was created using the hourly mean temperature. It was found for days for which at least one hourly mean temperature was available. Only the daily variable is used in the further analysis.

Presence of an overall trend during the 15 years in the daily mean temperature was checked

Table 2: Summary statistics for daily mean temperatures (°C)

Station	n	nmiss	Minimum	Q1	Median	Mean	Q3	Maximum	SD
Ristna	5368	111	-19.8	1.8	6.6	7.27	13.7	27.3	7.564
Tallinn-Harku	5478	1	-23.7	0.6	6.4	6.55	13.7	27.4	8.500
Tõravere	5476	3	-27.0	0.1	6.4	6.46	14.2	26.6	9.239
Vilsandi	5479	0	-19.4	2.2	7.3	7.83	14.5	27.4	7.640
Võru	5479	0	-28.0	0.3	6.5	6.63	14.5	28.7	9.437

n: number of values, nmiss: number of missing values, Q1: lower quartile, Q3: upper quartile, SD: standard deviation

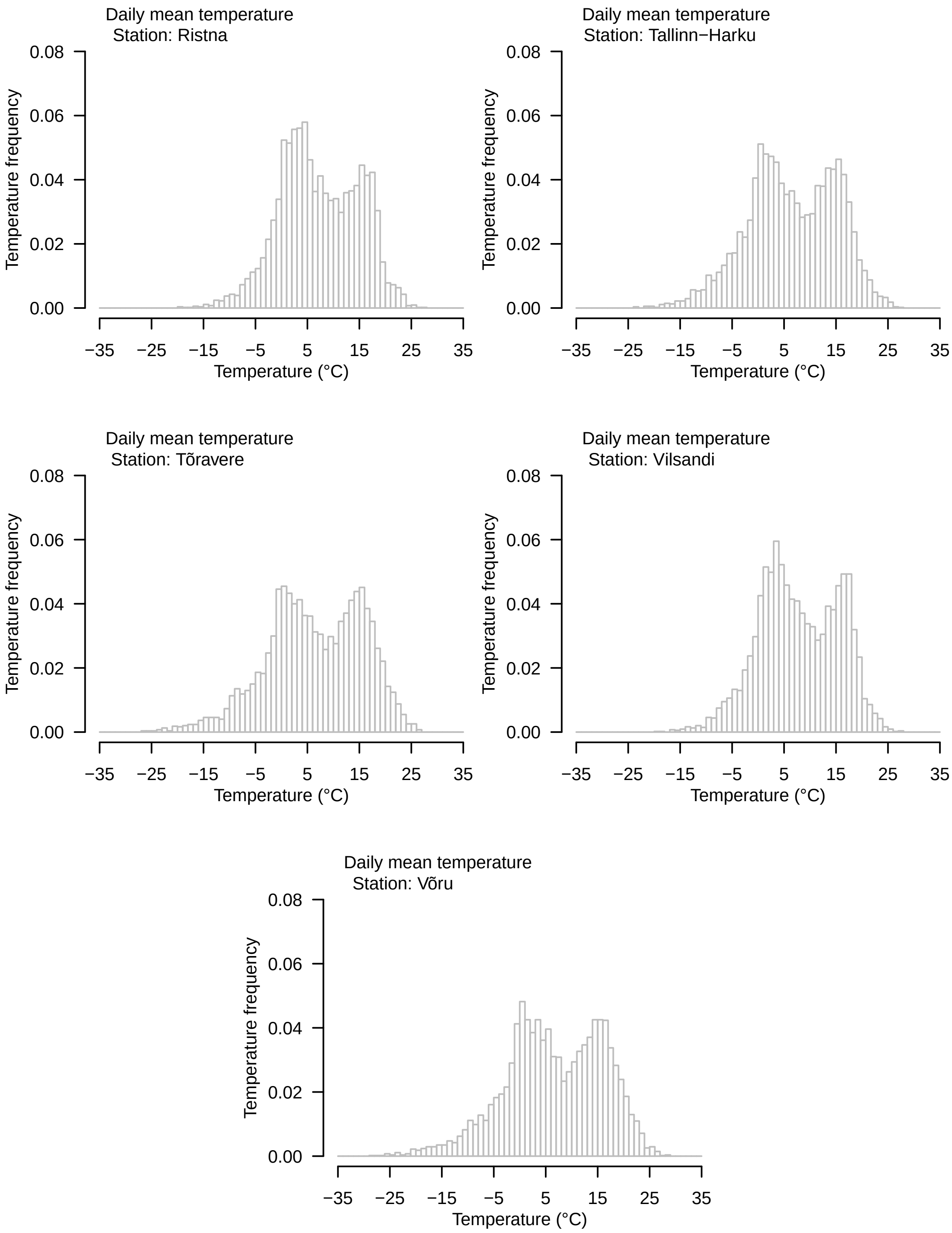


Figure 11: Histograms of daily mean temperatures (°C) by stations

graphically for all five stations and visually non was detected. The summary statistics of daily mean temperature are presented in Table 2 by stations. For illustration the histograms of the characteristic by stations are given in Figure 11. According to Table 2 the daily mean temperature is available for all 5479 days for Vilsandi and Võru stations, it is missing for 18.11.2010 for Tallinn-Harku station and for 14.03.2015, 15.03.2015 and 27.12.2016 for

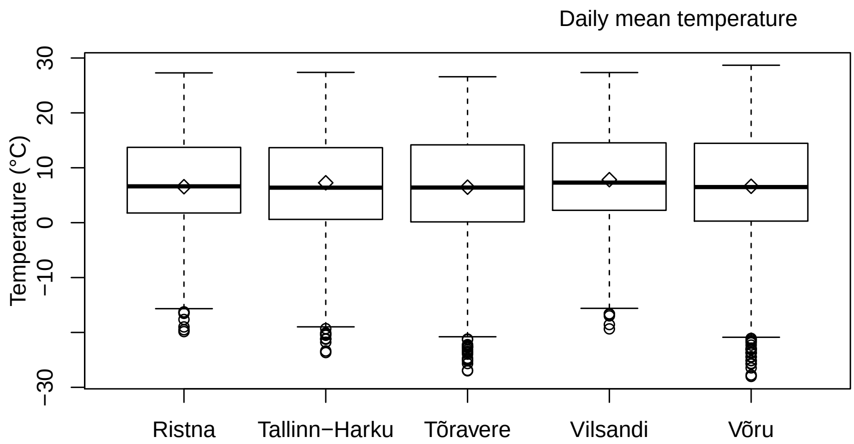


Figure 12: Boxplots of daily mean temperatures (°C) by stations

Tõravere station. For Ristna station the daily mean temperature is missing for 111 days with the longest consecutive missing period being from 14.06.2015 to 13.09.2015. For more insight on variation, the boxplots for all stations are presented in Figure 12.

At first glance the daily mean temperatures look quite similar across the five stations during 2004–2018 based on Figure 11. From the boxplots in Figure 12 it can be seen that the daily mean temperature distributions for the stations resemble each other. The seemingly most differencing aspect for daily mean temperatures for the stations is the minimum of the daily mean temperatures in Table 2. During 2004–2018 it was −28.0°C for Võru weather station and −27.0°C for Tõravere station from 20.01.2006, but −19.4°C for Vilsandi from 19.01.2006, −19.8°C for Ristna from 03.02.2012 and −23.7°C for Tallinn-Harku station from 04.02.2012. The maximums of the daily mean temperatures from 2004–2018 for the stations are more similar, 27.3°C for Ristna station and 27.4°C for Vilsandi station from 04.08.2014, 27.4°C for Tallinn-Harku station from 29.07.2018, 26.6°C for Tõravere station from 14.07.2010 and 28.7°C for Võru station from 29.07.2012.

3.3 Modeling via Finite Mixtures Using EM Algorithm

In this thesis finite mixtures up to $K = 8$ components are fitted on data until the value of BIC, given by (24), of mixture of $K+1$ components is less than the value of BIC of K-component mixture (see also Section 2.1). Thus, the K is fixed when estimating the mixture parameters.

Note that estimation of parameters of finite mixtures using the EM algorithm is implemented in multiple packages in the R environment (R Core Team, 2018). For example, package `mixtools` can be used for example to estimate parameters of normal and gamma mixtures using the EM algorithm (Benaglia et al., 2009), package `rebmix` allows besides other applications to estimate parameters of normal, lognormal, Weibull, gamma, binomial, Poisson, Dirac or circular von Mises mixtures (Nagode, 2015), package `mclust` focuses on mixtures of multivariate normal distributions (Scrucca et al., 2016), the EM algorithm can also be applied to estimate parameters of finite mixtures in packages `mixdist` (Macdonald and Du, 2018) and `mixR` (Yu, 2018).

However, it was chosen in this thesis to estimate the mixture parameters using the EM algorithm described in Section 2.2 using the R package `flexmix` (Leisch, 2004). That is because the package `flexmix` includes opportunity to estimate parameters of the needed lognormal, inverse Gaussian, gamma, Burr, inverse Burr, Weibull (in driver `FLXMCdist1`) and normal (in driver `FLXMCnorm1`) mixtures. Package `flexmix` is one of the most flexible packages for estimating parameters of finite mixtures as it allows also the user to easily write additional drives for the M-step of the expectation-maximization algorithm when needed (Leisch, 2004; Grün and Leisch, 2008; Grün et al., 2019). In addition, the package is applied in the article Miljkovic and Grün (2016), where a driver `FLXMCdist1` for the M-step is described. However, for estimating parameters of Rayleigh and truncated normal mixtures the author of this thesis wrote an additional driver `FLXMCdist2` for the M-step of the EM algorithm taking the existing driver `FLXMCdist1` as base.

The derivations of the formulas of the closed-form parameter estimates (on step p in the EM algorithm) used in the driver `FLXMCdist1` of the package `flexmix` and in the new driver `FLXMCdist2` written by the author for estimating Rayleigh and truncated normal mixtures are given in Section 2.2.2. For parameter estimates that require numerical optimization the iterative Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm (e.g., Fletcher, 1987, p. 55) is applied. The EM algorithm is run for each value of K repeatedly 10 times each time until the relative difference between the log-likelihood values found on consecutive steps is smaller than 10^{-6} or maximum number of iterations, 1000, is reached, the solution corresponding to the highest log-likelihood value is retained. The EM algorithm is initialized using random initialization, this means that the observations are split randomly to K groups and the starting values of the algorithm (parameter estimates) are found based on the partitions. In addition to following Miljkovic and Grün (2016): "any initial partition that contains less than 1% of the data is disregarded and only partitions meeting this size criterion on the components are used to initialize the EM algorithm", to avoid the potential overfitting the components are collapsed automatically when the estimate of a component weight π_k ($k = 1,\dots,K$) falls below 0.01 (Leisch, 2004). This means the EM algorithm is allowed to return estimates of a mixture of $K < K^*$ components while having started with fitting a K^*-component mixture.

From all fitted models the best fitting mixture is found by comparing the BIC (24) values, additionally NLL (22) and AIC (23) values for the models are given. The goodness of fit is illustrated for the best fitting models for each station using Q-Q-plot and P-P-plot; using goodness of fit tests stays out of scope of this thesis (see also Section 2.3).

3.3.1 Daily Maximal Hourly Mean Wind Speed

The distribution of wind speed can be relatively different even across short distances, for example Carta and Ramírez (2007) state that there are different wind speed distributions on the Gran Canaria island in Spain. Based on Table 1 and Figure 10 the distributions of daily maximal hourly mean wind speed varies among Ristna, Tallinn-Harku, Tõravere, Vilsandi and Võru weather stations.

The aim of this section is to find the best fitting (in the sense of the criterion given in Section 3.3) K-component mixture to model daily maximal hourly mean wind speed data measured in various weather stations in Estonia. Lognormal, inverse Gaussian, Gamma, Burr, inverse Burr, Weibull and truncated normal mixtures are applied, because they have been used for modeling positive variables in multiple articles (e.g., Carta and Ramírez, 2007; Akpinar and Akpinar, 2009; Chang, 2011; Kollu et al., 2012; Miljkovic and Grün, 2016; Ouarda and Charron, 2018). Additionally, a special case of Weibull mixture, Rayleigh mixture, is applied. In addition, the interest is to see whether same K-component mixtures (i.e. components are from the same distribution family and number of components is the same) fit for the stations.

Parameter estimates of a lognormal mixture (see Definition 4) on step p of the EM algorithm are derived in Section 2.2.2 on page 21, of an inverse Gaussian mixture (see Definition 5) on page 22, of a Gamma mixture (see Definition 6) on page 24, of a Burr mixture (see Definition 7) on page 25, of an inverse Burr mixture (see Definition 8) on page 26, of a Weibull mixture (see Definition 9) on page 27, of a Rayleigh mixture (see Definition 10) on page 28 and of a truncated normal mixture (see Definition 3) on page 29.

Ristna station

The values of the goodness of fit measures of mixture models fitted on daily maximal hourly mean wind speeds for Ristna weather station are presented in Table 3. The best mixtures brought out in bold in Table 3 are displayed in Figure 13 with the sample histogram.

The best fitting model for Ristna station is the two-component gamma mixture (BIC = 24268.88) followed by the two-component lognormal mixture (BIC = 24272.01). When comparing the best mixtures of different component distributions with each other the

Table 3: Fitted mixtures for daily maximal hourly mean wind speed for Ristna station

Mixture	K	NLL	AIC	BIC	Mixture	K	NLL	AIC	BIC
Lognormal	1	12278.18	24560.37	24573.58	Inverse	1	12293.99	24593.99	24613.81
	2	**12114.48**	**24238.97**	**24272.01**	Burr	**2**	**12122.15**	**24258.30**	**24304.56**
	3	12110.99	24237.99	24290.85		3	12115.58	24253.15	24325.85
Inverse	1	12244.87	24493.73	24506.95	Weibull	1	12785.90	25575.81	25589.02
Gaussian	2	12244.85	24499.71	24532.75		2	12200.43	24410.87	24443.91
	3	12244.81	24505.63	24558.50		3	12135.84	24287.69	24340.56
	4	**12108.37**	**24238.74**	**24311.43**		**4**	**12114.83**	**24251.66**	**24324.35**
	5	12108.28	24244.55	24337.07		5	12102.10	24232.19	24324.71
Gamma	1	12468.82	24941.65	24954.86	Rayleigh	1	12786.13	25574.27	25580.87
	2	**12112.92**	**24235.84**	**24268.88**		**2**	**12701.27**	**25408.54**	**25428.37**
	3	12109.24	24234.47	24287.34		3	12701.22	25412.44	25445.48
Burr	1	12356.22	24718.45	24738.27	Truncated	1	13144.35	26292.70	26305.92
	2	**12119.71**	**24253.42**	**24299.68**	normal	2	12223.09	24456.19	24489.23
	3	12106.96	24235.91	24308.60		3	12135.18	24286.36	24339.23
						4	**12107.30**	**24236.61**	**24309.30**
						5	12100.32	24228.64	24321.16

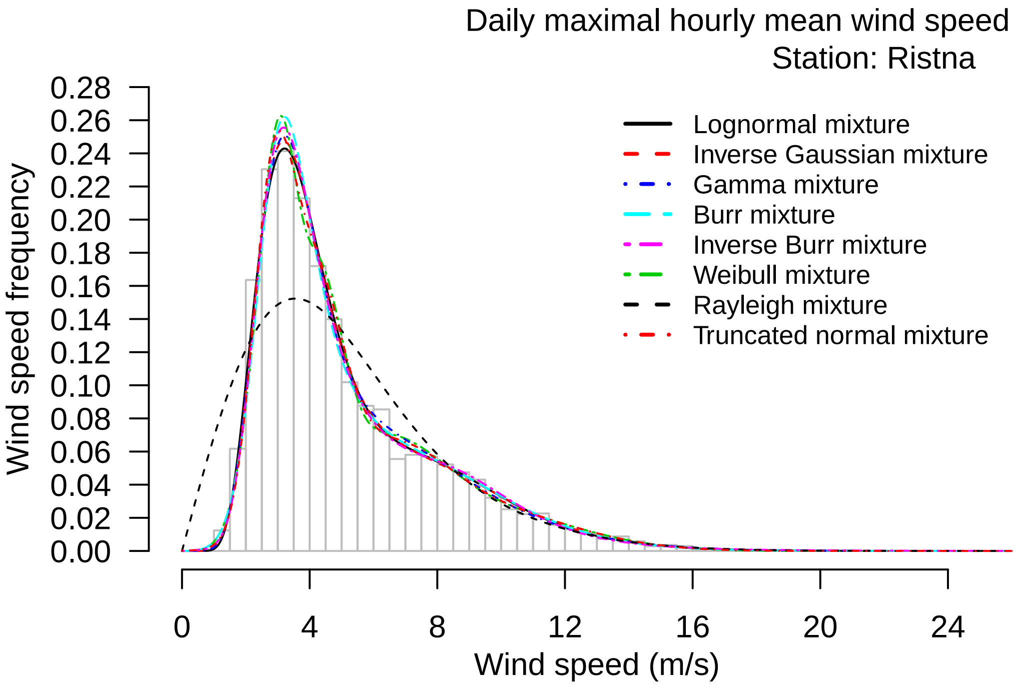


Figure 13: Best fitting mixtures from Table 3 (marked in bold)

two-component Rayleigh mixture has the lowest BIC value and its density has a noticeably different shape from the other quite close densities displayed in Figure 13.

The parameter estimates of the two-component gamma mixture are given in Table 8 on page 45 and the mixture density (the unbroken line) is displayed with the weighted component densities (dashed lines) in Figure A4. It can be seen that the two-component mixture density is a mixture of two gamma densities with nearly equal component weights with one of the components (green) giving the mixture density the relatively long right tail. In Figure A4 the Q-Q-plot for the mixture model suggests a relatively good fit on most of the data, but there are deviations from the 45° line for the larger values, the P-P-plot in the figure also shows a suitable fit as there are only minor deviations from the 45° line. Based on the Q-Q-plot and P-P-plot in Figure A4 the second best mixture model, two-component lognormal mixture, has also relatively suitable fit on the data.

Tallinn-Harku station

For fitting mixtures one value of daily maximal hourly mean wind speed for Tallinn-Harku station that was equal to zero (for 24.02.2011, see also Table 1) was replaced with 0.05 as some of the used mixture distributions are only defined on positive values (e.g., lognormal mixtures). In addition, the raw variable values are given with accuracy of one decimal place, thus $0.1/2 = 0.05$ was used to replace the zero-value.

The values of the goodness of fit measures of mixture models fitted on daily maximal hourly mean wind speeds for Tallinn-Harku weather station are presented in Table 4. The best mixtures brought out in bold in Table 4 are displayed in Figure 14 with the sample histogram. Note that only one-component mixtures of all K-component ($K = 1, \ldots, 8$)

Table 4: Fitted mixtures for daily maximal hourly mean wind speed for Tallinn-Harku station

Mixture	K	NLL	AIC	BIC	Mixture	K	NLL	AIC	BIC
Lognormal	**1**	**10471.14**	**20946.28**	**20959.50**	Inverse Gaussian	**1**	**10815.24**	**21634.48**	**21647.70**
Gamma	**1**	**10299.03**	**20602.07**	**20615.28**	Rayleigh	**1**	**11246.59**	**22495.19**	**22501.80**
	2	10298.98	20607.96	20641.00		2	11246.59	22499.19	22519.01
Burr	1	10296.96	20599.93	20619.75	Inverse	1	10338.50	20683.00	20702.82
	2	**10262.16**	**20538.31**	**20584.57**	Burr	**2**	**10266.86**	**20547.71**	**20593.97**
	3	10252.86	20527.73	20600.42		3	10249.93	20521.86	20594.56
Weibull	1	10404.24	20812.48	20825.70	Truncated	1	10403.16	20810.33	20823.55
	2	10298.40	20606.80	20639.84	normal	2	10273.97	20557.94	20590.98
	3	**10261.28**	**20538.57**	**20591.44**		**3**	**10243.94**	**20503.88**	**20556.74**
	4	10255.10	20532.20	20604.90		4	10242.75	20507.51	20580.20

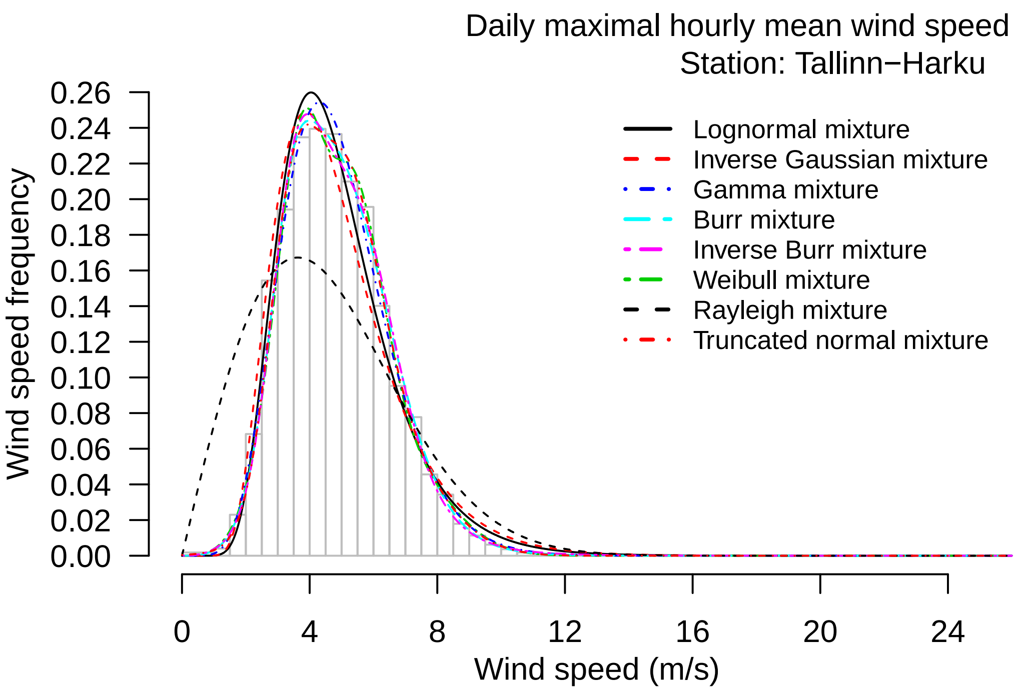


Figure 14: Best fitting mixtures from Table 4 (marked in bold)

mixtures of lognormal and inverse Gaussian distributions were fitted as for these two cases the parameter estimation of a two-component mixture reduced to estimation of an one-component mixture.

The best fitting model for Tallinn-Harku station is the three-component truncated normal mixture (BIC = 20556.74), it is followed by the two-component Burr mixture (BIC = 20584.57), the three-component Weibull mixture (BIC = 20591.44) and the two-component inverse Burr mixture (BIC = 20593.97). When comparing the best mixtures of different component distributions with each other the one-component Rayleigh mixture has the lowest BIC value and its density has, again, different shape from the densities of the other mixtures displayed in Figure 14.

The parameter estimates of the three-component truncated normal mixture are given in Table 8 on page 45 and the mixture density (the unbroken line) is displayed with the weighted component densities (dashed lines) in Figure A5. The three-component mixture density is a mixture of three truncated normal densities with two mixing weights approximately equal to 0.36 and one approximately equal to 0.28. In Figure A5 the Q-Q-plot for the mixture model shows a relatively good fit on the data as there are only few small deviations from the 45° line. Based on the P-P-plot the mixture model also has a good fit as the plot follows the 45° line quite well. The second best mixture, two-component Burr mixture, has also a good fit on the data based on the Q-Q-plot and P-P-plot in Figure A5.

Tõravere station

The values of the goodness of fit measures of mixture models fitted on daily maximal hourly mean wind speeds for Tõravere weather station are presented in Table 5. The best mixtures brought out in bold in Table 5 are displayed in Figure 15 with the sample histogram. Note that of inverse Gaussian mixtures only one- and two-component mixtures were fitted as the parameter estimation of a three-component mixture reduced to estimation of a two-component mixture.

The best model for Tõravere station is the one-component gamma mixture (BIC = 18558.49) followed by the two-component Burr mixture (BIC = 18580.16). When comparing the best mixtures of different component distributions with each other the best Rayleigh mixture differs, again, noticeably from the other mixtures displayed in Figure 15.

Table 5: Fitted mixtures for daily maximal hourly mean wind speed for Tõravere station

Mixture	K	NLL	AIC	BIC	Mixture	K	NLL	AIC	BIC
Lognormal	1	9390.674	18785.35	18798.57	Inverse	1	9493.408	18990.82	19004.03
	2	**9292.235**	**18594.47**	**18627.51**	Gaussian	**2**	**9332.357**	**18674.71**	**18707.76**
	3	9292.222	18600.44	18653.31					
Gamma	**1**	**9270.635**	**18545.27**	**18558.49**	Rayleigh	**1**	**10118.43**	**20238.85**	**20245.46**
	2	9263.073	18536.15	18569.19		2	10118.43	20242.85	20262.68
Burr	1	9282.430	18570.86	18590.69	Inverse	1	9328.849	18663.70	18683.52
	2	**9259.948**	**18533.90**	**18580.16**	Burr	**2**	**9268.059**	**18550.12**	**18596.38**
	3	9259.379	18540.76	18613.45		3	9255.863	18533.73	18606.42
Weibull	1	9374.854	18753.71	18766.93	Truncated	1	9399.224	18802.45	18815.66
	2	9288.422	18586.84	18619.89	normal	2	9280.294	18570.59	18603.63
	3	**9263.566**	**18543.13**	**18596.00**		**3**	**9260.526**	**18537.05**	**18589.92**
	4	9260.776	18543.55	18616.25		4	9250.674	18523.35	18596.04

The parameter estimates of the gamma distribution are given in Table 8 on page 45 and the density in Figure A6. The density is nearly bell-shaped. In Figure A6 the Q-Q-plot for the mixture model shows a relatively good fit on the data as there are only some minor deviations from the 45° line for the extreme values. Based on the P-P-plot the mixture model also has a good fit as there are no observable deviations from the 45° line. The second best mixture, two-component Burr mixture, has also based on Q-Q-plot and P-P-plot in Figure A6 a relatively good fit on the data.

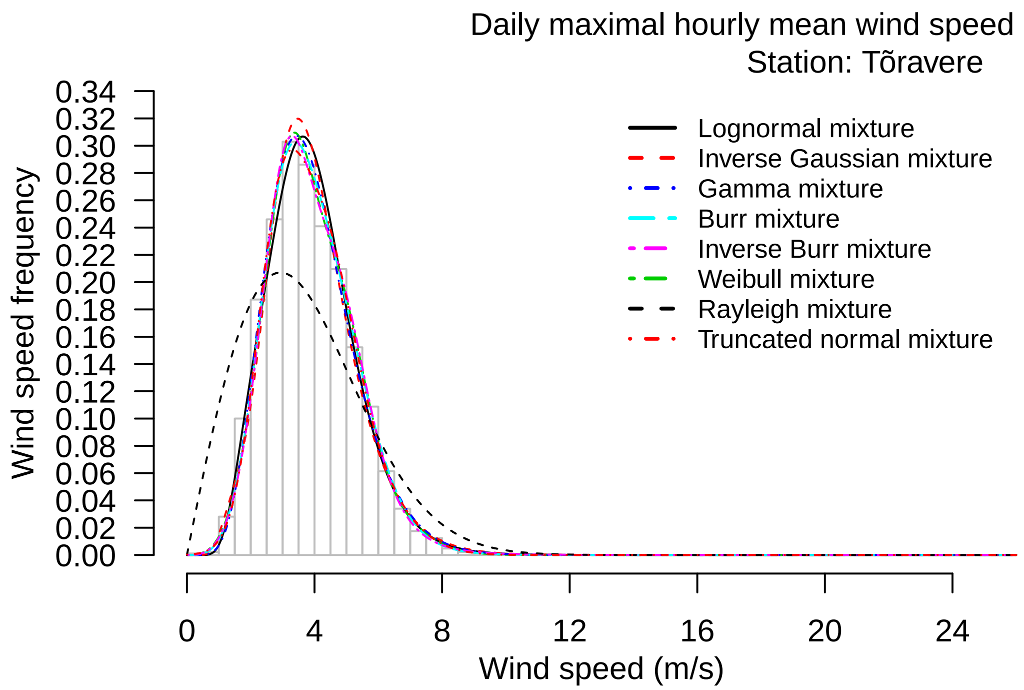


Figure 15: Best fitting mixtures from Table 5 (marked in bold)

Vilsandi station

The values of the goodness of fit measures of mixture models fitted on daily maximal hourly mean wind speeds for Vilsandi weather station are presented in Table 6. The best mixtures brought out in bold in Table 6 are displayed in Figure 16 with the sample histogram.

The best model for daily maximal hourly mean wind speed for Vilsandi station is the two-component inverse Gaussian mixture (BIC = 27800.36) followed by the two-component lognormal mixture (BIC = 27803.93). When comparing the best mixtures

Table 6: Fitted mixtures for daily maximal hourly mean wind speed for Vilsandi station

Mixture	K	NLL	AIC	BIC	Mixture	K	NLL	AIC	BIC
Lognormal	1	13945.93	27895.85	27909.07	Inverse	1	13939.28	27882.56	27895.77
	2	**13880.44**	**27770.89**	**27803.93**	Gaussian	**2**	**13878.66**	**27767.32**	**27800.36**
	3	13879.89	27775.77	27828.64		3	13877.87	27771.73	27824.60
Gamma	**1**	**13924.12**	**27852.25**	**27865.46**	Rayleigh	**1**	**14460.04**	**28922.07**	**28928.68**
	2	13924.10	27858.19	27891.24		2	14460.04	28926.08	28945.90
Burr	1	13998.59	28003.19	28023.01	Inverse	1	14050.74	28107.48	28127.30
	2	13903.14	27820.28	27866.54	Burr	2	13911.85	27837.69	27883.95
	3	**13879.63**	**27781.26**	**27853.95**		**3**	**13885.15**	**27792.31**	**27865.00**
	4	13872.31	27774.62	27873.75		4	13874.03	27778.06	27877.19
Weibull	1	14096.01	28196.03	28209.25	Truncated	1	14216.13	28436.26	28449.48
	2	13963.34	27936.69	27969.73	normal	2	13962.92	27935.83	27968.88
	3	13911.16	27838.31	27891.18		3	13905.03	27826.06	27878.93
	4	13905.93	27833.86	27906.55		**4**	**13873.23**	**27768.46**	**27841.15**
	5	**13875.56**	**27779.12**	**27871.65**		5	13871.59	27771.18	27863.70
	6	13873.72	27781.44	27893.79					

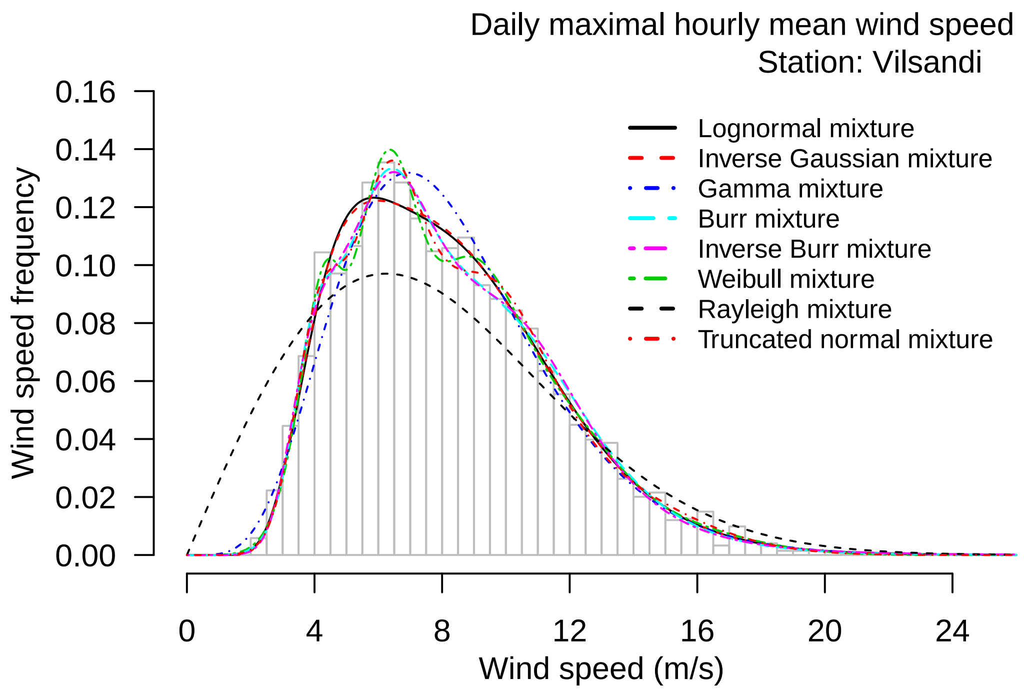


Figure 16: Best fitting mixtures from Table 6 (marked in bold)

of different component distributions with each other the one-component Rayleigh mixture has the lowest BIC value and its density has, again, different shape than the other densities in Figure 16. From the rest some mixtures (e.g., Weibull mixture) follow more or less the seemingly trimodal shape of the displayed histogram, but densities of the two best mixtures (inverse Gaussian and lognormal) have a slightly different shape from all the other densities.

The parameter estimates of the two-component inverse Gaussian mixture are given in Table 8 on page 45 and the mixture density (the unbroken line) is displayed with the weighted component densities (dashed lines) in Figure A7. The two-component mixture density is a mixture of two inverse Gaussian densities with quite close component weights (approximately 0.45 and 0.55). The Q-Q-plot in Figure A7 for the mixture model suggests a relatively good fit on the data although there are some minor deviations from the 45° line for the greater values, the P-P-plot also shows a suitable fit as there are no spottable large deviations from the 45° line. The second best mixture, two-component lognormal mixture, has also based on Q-Q-plot and P-P-plot in Figure A7 a relatively suitable fit on the data.

Võru station

The values of the goodness of fit measures of mixture models fitted on daily maximal hourly mean wind speeds for Võru weather station are presented in Table 7. The best mixtures brought out in bold in Table 7 are displayed in Figure 17 with the sample histogram.

The best model for Võru station is the two-component gamma mixture (BIC = 18840.04), it is followed by the one-component Burr mixture (BIC = 18847.12) and three-component lognormal mixture (BIC = 18851.19). When comparing the best mixtures of different component distributions with each other the one-component Rayleigh mixture has the

Table 7: Fitted mixtures for daily maximal hourly mean wind speed for Võru station

Mixture	K	NLL	AIC	BIC	Mixture	K	NLL	AIC	BIC
Lognormal	1	9631.086	19266.17	19279.39	Inverse	1	9796.134	19596.27	19609.49
	2	9455.736	18921.47	18954.51	Gaussian	2	9470.457	18950.91	18983.96
	3	**9391.159**	**18798.32**	**18851.19**		**3**	**9394.434**	**18804.87**	**18857.74**
	4	9389.810	18801.62	18874.32		4	9394.547	18805.09	18857.96
Gamma	1	9440.527	18885.05	18898.27	Inverse	1	9450.806	18907.61	18927.44
	2	**9398.498**	**18807.00**	**18840.04**	Burr	**2**	**9404.009**	**18822.02**	**18868.28**
	3	9390.547	18797.09	18849.96		3	9391.788	18805.58	18878.27
Burr	**1**	**9410.649**	**18827.30**	**18847.12**	Rayleigh	**1**	**10181.07**	**20364.14**	**20370.75**
	2	9404.136	18822.27	18868.53		2	10181.07	20368.14	20387.97
Weibull	1	9489.737	18983.47	18996.69	Truncated	1	9509.118	19022.24	19035.45
	2	9419.222	18848.44	18881.49	normal	2	9415.939	18841.88	18874.92
	3	**9404.170**	**18824.34**	**18877.21**		**3**	**9393.800**	**18803.60**	**18856.47**
	4	9392.581	18807.16	18879.86		4	9388.274	18798.55	18871.24

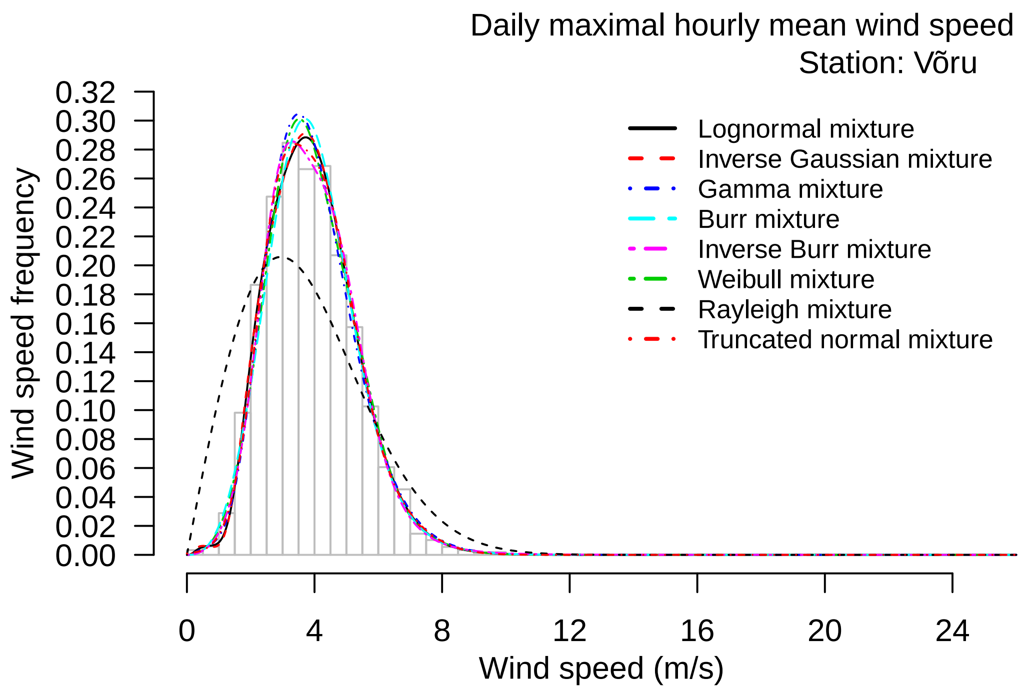


Figure 17: Best fitting mixtures from Table 7 (marked in bold)

lowest BIC value and its density has, again, different shape than the other quite similar-looking densities in Figure 17.

The parameter estimates of the two-component gamma mixture are given in Table 8 on page 45 and the mixture density (the unbroken line) is displayed with the weighted component densities (dashed lines) in Figure A8. The two-component mixture density is a mixture of two gamma densities with one component having large weight, approximately 0.97, and the other component having small weight, approximately 0.03. In Figure A8 the Q-Q-plot for the mixture model suggests a relatively good fit on the data although there are few deviations from the 45° line for the greater values, the P-P-plot also shows a suitable fit as there are no large deviations from the 45° line. The second best mixture, one-component Burr mixture might also have a suitable fit on the data based on the Q-Q-plot and P-P-plot in A8 although there are minor deviations from the 45° line in the P-P-plot.

Table 8: Parameter estimates of the best fitting mixtures for daily maximal hourly mean wind speed by stations

			Component		
Station	Mixture		1	2	3
Ristna	Gamma	$\hat{\pi}_k$	0.521540	0.478460	
		$\hat{\alpha}_k$	11.803291	5.898564	
		$\hat{\lambda}_k$	3.465999	0.834919	
Tallinn-Harku	Truncated normal	$\hat{\pi}_k$	0.355075	0.280893	0.364032
		$\hat{\mu}_k$	5.814255	3.422201	5.054864
		$\hat{\sigma}_k$	1.806142	0.752645	1.031898
Tõravere	Gamma	$\hat{\pi}_k$	1		
		$\hat{\alpha}_k$	8.175239		
		$\hat{\lambda}_k$	2.086373		
Vilsandi	Inverse Gaussian	$\hat{\pi}_k$	0.551372	0.448628	
		$\hat{\mu}_k$	6.360028	10.492660	
		$\hat{\lambda}_k$	50.813816	140.088100	
Võru	Gamma	$\hat{\pi}_k$	0.969397	0.030603	
		$\hat{\alpha}_k$	8.558858	2.585145	
		$\hat{\lambda}_k$	2.153524	0.966705	

Best Mixtures

According to Table 8 for three stations (Ristna, Tõravere and Võru) gamma mixtures had the best fit on the daily maximal hourly mean wind speed data. For Ristna and Võru stations the most well-fitting distribution for daily maximal hourly mean wind speed was a two-component gamma mixture, although the two mixture densities are not extremely similar: the two-component gamma mixture for Ristna station has nearly equal mixing weights while the two-component gamma mixture for Võru station has one weight close to 1 and the other weight close to 0 making the last's density in Figure A8 more similar-looking to the one-component gamma mixture density for wind speed for Tõravere station in Figure A6. For wind speed for Tallinn-Harku station the best mixture was the truncated normal mixture with three components. For daily maximal hourly mean wind speed in Vilsandi station the most

Table 9: Summary characteristics for the best mixtures (given in Table 8) for daily maximal hourly mean wind speed

Station	Mixture	Q1	Median	Mean	Q3	SD
Ristna	Gamma	3.1	4.2	5.16	6.5	2.811
Tallinn-Harku	Truncated normal	3.7	4.7	4.87	5.9	1.618
Tõravere	Gamma	2.9	3.8	3.92	4.7	1.370
Vilsandi	Inverse Gaussian	5.7	7.8	8.21	10.2	3.273
Võru	Gamma	2.9	3.8	3.93	4.8	1.387

Q1: lower quartile, Q3: upper quartile, SD: standard deviation suitable mixture was the two-component inverse Gaussian mixture.

Additionally, a two-component lognormal mixture provided the second best fit on daily maximal hourly mean wind speed data for Ristna and Vilsandi stations, for Tõravere and Tallinn stations the second best mixtures were two-component Burr mixtures and for Võru station the second best was an one-component Burr mixture (see also Tables 3–7).

Summary characteristics (lower quantile, median, mean, upper quantile, standard deviation) for the best mixtures for daily maximal hourly mean wind speed by weather stations are given in Table 9 (see also Section 1.1 for formulas). For all stations the empirical estimates of the characteristics in Table 1 and the theoretical estimates in Table 9 are close.

Table 10: Some theoretical (given by mixtures in Table 8) and empirical quantiles for daily maximal hourly mean wind speed

		Theoretical		Empirical	
Station	Mixture	$q_{0.95}$	$q_{0.99}$	$q_{0.95}$	$q_{0.99}$
Ristna	Gamma	10.86	14.15	11.00	13.92
Tallinn-Harku	Truncated normal	7.79	9.26	7.80	9.30
Tõravere	Gamma	6.41	7.79	6.30	7.60
Vilsandi	Inverse Gaussian	14.21	17.44	14.20	17.30
Võru	Gamma	6.43	7.80	6.30	7.70

The 0.95-quantiles and 0.99-quantiles in Table 10 give some insight to the extreme daily maximal hourly wind speeds for the stations. During 2004–2018 for Ristna station for 5% of the days the daily maximal hourly wind speed was greater than 11 m/s and for Vilsandi station greater than 14.2 m/s. For other stations daily maximal hourly wind speeds greater than 10 m/s occurred for less than 1% of the days. When the theoretical quantiles are viewed as the true population quantiles, it can be said that for 5% of days (about 18 days a year) the daily maximal hourly winds are greater than 10.86 m/s for Ristna station, 7.79 m/s for Tallinn-Harku station, 6.41 m/s for Tõravere station, 14.21 m/s for Vilsandi station and 6.43 m/s for Võru station; for 1% of days (about 3 days a year) they are greater than 14.15 m/s for Ristna station, 9.26 m/s for Tallinn-Harku station, 7.79 m/s for Tõravere station, 17.44 m/s for Vilsandi station and 7.80 m/s for Võru station. The relative differences between empirical and theoretical quantiles in Table 10 are rather small, not over 2.5%.

3.3.2 Daily Mean Temperature

The aim of this section is to find the best fitting (in the sense of the criterion given in Section 3.3) K-component normal mixture to model daily mean temperature data measured in various weather stations in Estonia. Normal mixtures are applied, because normal distributions and normal mixtures have been applied on temperature data (e.g., Grace and Curran, 1993; Harmel et al., 2002). Note that temperature distributions are, for example, helpful for modeling extreme temperatures (e.g., Kjellström et al., 2007; Osman et al., 2015). In addition, in this section the interest is to see whether same K-component normal mixtures (i.e. number of components is the same) fit for the stations.

Parameter estimates of a normal mixture (see Definition 2) on step p of the EM algorithm are given, for example, in McLachlan and Peel (2000, pp. 81–83).

Table 11: Fitted normal mixtures for daily mean temperatures by stations

Station	K	NLL	AIC	BIC	Station	K	NLL	AIC	BIC
Ristna	**1**	**18477.64**	**36959.28**	**36972.46**	Vilsandi	**1**	**18915.20**	**37834.40**	**37847.61**
	2	18477.62	36965.25	36998.19		2	18915.16	37840.32	37873.37
Tallinn-Harku	1	19495.81	38995.61	39008.83	Võru	1	20072.28	40148.55	40161.77
	2	19283.24	38576.48	38609.52		2	19851.53	39713.05	39746.10
	3	**19235.97**	**38487.95**	**38540.81**		**3**	**19775.09**	**39566.19**	**39619.06**
	4	19235.85	38493.71	38566.40		4	19774.30	39570.60	39643.29
Tõravere	1	19945.13	39894.26	39907.48					
	2	19717.16	39444.31	39477.35					
	3	**19649.38**	**39314.77**	**39367.63**					
	4	19648.68	39319.36	39392.05					

The values of goodness of fit measures of normal mixture models fitted on daily mean temperatures for the five weather stations are given in Table 11. Densities of the best normal mixtures (the unbroken line) are displayed with weighted component densities (dashed lines) in Figure A9 with the sample histograms. The best normal mixtures for Ristna and Vilsandi stations are one-component normal mixtures, but for Tallinn-Harku, Tõravere and Võru stations three-component normal mixtures. This highlights also the fact that detecting the most suitable number of components, K, by simply looking for the modes on the sample histogram might be misleading as for all stations the sample histograms in Figure 11 seem to have at first sight two modes which might lead to think of $K = 2$ as a possible candidate for the number of components.

Table 12: Parameter estimates of the best fitting normal mixtures (given in Table 11, marked bold) by stations

		Component		
Station		1	2	3
Ristna	$\hat{\pi}_k$	1		
	$\hat{\mu}_k$	7.267535		
	$\hat{\sigma}_k$	7.563532		
Tallinn-Harku	$\hat{\pi}_k$	0.245309	0.364007	0.390684
	$\hat{\mu}_k$	2.533927	0.6114085	14.621670
	$\hat{\sigma}_k$	3.264154	7.6839954	3.891526
Tõravere	$\hat{\pi}_k$	0.411211	0.300985	0.287804
	$\hat{\mu}_k$	14.995269	2.465318	-1.528327
	$\hat{\sigma}_k$	4.052954	3.754362	8.591004
Vilsandi	$\hat{\pi}_k$	1		
	$\hat{\mu}_k$	7.830170		
	$\hat{\sigma}_k$	7.640485		
Võru	$\hat{\pi}_k$	0.283546	0.408795	0.307659
	$\hat{\mu}_k$	-1.560077	15.369260	2.583253
	$\hat{\sigma}_k$	8.878689	4.146070	3.784726

The parameter estimates of the best normal mixtures are displayed in Table 12. In Figure A9 the densities of the three-component normal mixtures for daily mean temperature for Tallinn-Harku, Tõravere and Võru stations have a quite similar shape. Especially similar are mixtures for the inland stations Tõravere and Võru which also have close estimates of the mixture parameters. The Q-Q-plots in Figure A10 for the mixture models suggest a relatively good fit on most of the data for Tallinn-Harku, Tõravere and Võru stations, for the tails there are some small deviations from the 45° line, the P-P-plots in Figure A11 confirm the seemingly good fit. The Q-Q-plots and the P-P-plots for the normal mixture models show that the one-component normal mixtures do not have a very good fit on the data for Ristna and Vilsandi stations as there are some distinguishable deviations from the 45° lines for both stations for both plots.

Table 13: Summary characteristics for the best fitting normal mixtures (given in Table 12) for daily mean temperatures

Station	Q1	Median	Mean	Q3	SD
Ristna	2.2	7.3	7.27	12.4	7.564
Tallinn-Harku	0.6	6.3	6.56	13.7	8.501
Tõravere	0.2	6.4	6.47	14.1	9.239
Vilsandi	2.7	7.8	7.83	13.0	7.640
Võru	0.3	6.5	6.64	14.5	9.437

Q1: lower quartile, Q3: upper quartile, SD: standard deviation

Summary characteristics (lower quantile, median, mean, upper quantile, standard deviation) for the best normal mixtures for daily mean temperature by weather stations are given in Table 13. In general, the empirical estimates of the characteristics in Table 2 and the theoretical estimates in Table 13 are close for Tallinn-Harku, Tõravere and Võru stations, for Ristna and Vilsandi stations there are some differences.

Table 14: Some theoretical (given by mixtures in Table 12) and empirical quantiles for daily mean temperature

	Theoretical				Empirical			
Station	$q_{0.01}$	$q_{0.05}$	$q_{0.95}$	$q_{0.99}$	$q_{0.01}$	$q_{0.05}$	$q_{0.95}$	$q_{0.99}$
Ristna	-10.33	-5.17	19.71	24.86	-10.44	-4.77	18.69	22.31
Tallinn-Harku	-14.14	-7.80	19.18	22.36	-13.96	-7.80	18.98	22.83
Tõravere	-17.12	-9.62	19.82	23.09	-17.47	-9.36	19.85	23.27
Vilsandi	-9.94	-4.74	20.40	25.60	-9.64	-4.73	19.20	22.59
Võru	-17.62	-9.83	20.29	23.65	-18.01	-9.69	20.41	23.63

The 0.01-, 0.05-, 0.95- and 0.99-quantiles in Table 14 give some insight to the extreme daily mean temperatures for the stations. During 2004–2018 for Ristna station for 5% of the days the daily mean temperature was lower than -4.77°C and for 5% of the days higher than 18.69°C . For Vilsandi station the daily mean temperature was below -9.64°C for 1% of the days, but for Võru station it was lower than -9.69°C for 5% of the days. When the theoretical quantiles are viewed as the true population quantiles it can be said, for example, that for 1% of days (about 3 days a year) the daily mean temperatures are greater than 24.86°C for

Ristna station, 22.36°C for Tallinn-Harku station, 23.09°C for Tõravere station, 25.60°C for Vilsandi station, 23.65°C for Võru station (Table 14).

The relative differences between empirical and theoretical quantiles in Table 10 are quite small for Tallinn-Harku, Tõravere and Võru stations, not over 2.8%. For Ristna and Vilsandi stations the relative differences are greater, for example 13.3% for 0.99-quantile for Vilsandi station (empirical $q_{0.99} = 22.59$, theoretical $q_{0.99} = 25.60$).

Summary

In this thesis the finite mixtures were studied. Finite mixture is commonly defined by its distribution function, a weighted sum of component distribution functions; or by its density, a weighted sum of component densities. For estimating parameters of finite mixtures the expectation-maximization (EM) algorithm is often used. Additionally, finite mixtures were applied in this thesis to model Estonian wind speed and temperature data, the last has to the author's knowledge not been done before.

Overview of finite mixtures was given with special focus on the concept of estimating the parameters of finite mixtures with the EM algorithm. For lognormal, inverse Gaussian, gamma, Burr, inverse Burr, Weibull, Rayleigh and truncated normal mixtures corresponding theoretical formulas were derived. Mixtures of lognormal, inverse Gaussian, gamma, Burr, inverse Burr, Weibull, Rayleigh and truncated normal distributions were fitted on Estonian wind speed data and mixture of normal distribution on Estonian temperature data. For Rayleigh and truncated normal mixtures a new driver was provided for the R package `flexmix` (Leisch, 2004).

Among the proposed mixtures the best was found to fit daily maximal hourly mean wind speeds for Ristna, Tallinn-Harku, Tõravere, Vilsandi and Võru weather stations in Estonia based on values of Bayesian information criterion (BIC). For the stations the best fit was achieved by various mixtures: one-component gamma mixture (Tõravere), two-component gamma mixture (Ristna, Võru), three-component truncated normal mixture (Tallinn-Harku) and two-component inverse Gaussian mixture (Vilsandi). For none of the stations the Rayleigh mixtures provided a suitable fit. For all stations except one the best fitting mixture had less than three components. Note that in the articles where finite mixtures were used to model wind speed data (e.g., Chang, 2011) up to two components were used.

Additionally, normal mixtures were fitted on daily mean temperature data for Ristna, Tallinn-Harku, Tõravere, Vilsandi and Võru weather stations in Estonia. By value of BIC the best mixtures for daily mean temperature data for Tallinn-Harku, Tõravere and Võru stations were three-component (normal) mixtures which provided a suitable fit on the data. For daily mean temperature data for Ristna and Vilsandi stations located on islands in the Baltic Sea the best mixture by the Bayesian criterion value, one-component (normal) mixture, might according to the graphical goodness of fit methods not have a suitable fit.

The results of the thesis might be useful in the management of wind energy systems, for predicting extreme weather conditions, or in agriculture.

In areas of further research on application of finite mixtures on Estonian meteorological data it is worth to study the seasonal models as well as short term models (e.g., mean 10-minute wind speed, maximal hourly wind speed, maximal hourly temperature). Additionally, multivariate mixtures could be applied on the meteorological data.

References

[1] Abu Bakar, S. A., Nadarajah, S., and Adzhar, Z. A. A. K. (2018). Loss modeling using Burr mixtures. *Empirical Economics*, **54**(4), 1503–1516.

[2] Akpinar, S., and Akpinar, E. K. (2009). Estimation of wind energy potential using finite mixture distribution models. *Energy Conversion and Management*, **50**(4), 877–884.

[3] Antonio, K., Badescu, A., Lan, G., Lin, X., and Verbelen, R. (2014). Fitting mixtures of Erlangs to censored and truncated data using the EM algorithm. *SSRN Electronic Journal*, **45**(3).

[4] Babu, G. J., and Rao, C. R. (2004). Goodness-of-fit tests when parameters are estimated. *Sankhyā: The Indian Journal of Statistics*, **66**(1), 63–74.

[5] Benaglia, T., Chauveau, D., Hunter, D. R., and Young, D. (2009). mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software*, **32**(6), 1–29. `https://CRAN.R-project.org/package=mixtools`

[6] Carta, J., and Ramírez, P. (2007). Analysis of two-component mixture Weibull statistics for estimation of wind speed distributions. *Renewable Energy*, **32**(3), 518–531.

[7] Casella, G., and Berger, R. (2002). *Statistical inference* (2nd ed.). Duxbury.

[8] Chang, T. P. (2011). Estimation of wind energy potential using different probability density functions. *Applied Energy*, **88**(5), 1848–1856.

[9] Devis-Morales, A., Sanchez, R., Bernal, G., and Osorio, A. (2017). Assessment of extreme wind and waves in the Colombian Caribbean Sea for offshore applications. *Applied Ocean Research*, **69**, 10–26.

[10] Dias, S., Andreozzi, V., and Martins, M. R. (2013). Analysis of HIV/AIDS DRG in Portugal: a hierarchical finite mixture model. *The European Journal of Health Economics*, **14**(5), 715–723.

[11] Estonian Weather Service. (2019a). *Estonian wind speed and temperature data.*

[12] Estonian Weather Service. (2019b). *Vaatlusvõrk [Observation network].* Retrieved 09.04.2019, from `https://www.ilmateenistus.ee/ilmateenistus/vaatlusvork/`

[13] Everitt, B. S., and Hand, D. J. (1981). *Finite mixture distributions*. London: Chapman and Hall.

[14] Fletcher, R. (1987). *Practical methods of optimization* (2nd ed.). Wiley.

[15] Forbes, C., Evans, M., Hastings, N., and Peacock, B. (2011). *Statistical distributions* (4th ed.). New Jersey: Wiley.

[16] Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. New York: Springer.

[17] Gibbons, J. D., and Chakraborti, S. (2003). *Nonparametric statistical inference* (4th ed.). New York: Marcel Dekker.

[18] Grace, W., and Curran, E. (1993). A binormal model of frequency distributions of daily maximum temperature. *Australian Meteorological Magazine*, **42**, 151–161.

[19] Grün, B., and Leisch, F. (2008). FlexMix version 2: Finite mixtures with concomitant variables and varying and constant parameters. *Journal of Statistical Software*, **28**(4), 1–35.

[20] Grün, B., Leisch, F., Sarkar, D., Mortier, F., and Picard, N. (2019). *flexmix: Flexible mixture modeling. R package version 2.3-15.* `https://CRAN.R-project.org/package=flexmix`

[21] Harmel, R. D., Richardson, C. W., Hanson, C. L., and Johnson, G. L. (2002). Evaluating the adequacy of simulating maximum and minimum daily air temperature with the normal distribution. *Journal of Applied Meteorology*, **41**(7), 744–753.

[22] Kjellström, E., Bärring, L., Jacob, D., Jones, R., Lenderink, G., and Schär, C. (2007). Modelling daily temperature extremes: Recent climate and future changes over Europe. *Climatic Change*, **81**(1), 249–265.

[23] Kleiber, C., and Kotz, S. (2003). *Statistical size distributions in economics and actuarial sciences*. New Jersey: Wiley.

[24] Kollu, R., Rayapudi, S. R., Narasimham, S., and Pakkurthi, K. M. (2012). Mixture probability distribution functions to model wind speed distributions. *International Journal of Energy and Environmental Engineering*, **3**(1), 27.

[25] Leisch, F. (2004). FlexMix: A general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software*, **11**(8), 1–18.

[26] Macdonald, P., and Du, J. (2018). *mixdist: Finite mixture distribution models. R package version 0.5-5.* `https://CRAN.R-project.org/package=mixdist`

[27] McLachlan, G. J., and Krishnan, T. (2008). *The EM algorithm and extensions* (2nd ed.). New Jersey: Wiley.

[28] McLachlan, G. J., and Peel, D. (2000). *Finite mixture models*. New York: Wiley.

[29] Miljkovic, T., and Grün, B. (2016). Modeling loss data using mixtures of distributions. *Insurance: Mathematics and Economics*, **70**, 387–396.

[30] Nagode, M. (2015). Finite mixture modeling via REBMIX. *Journal of Algorithms and Optimization*, **3**(2), 14–28. `https://CRAN.R-project.org/package=rebmix`

[31] OÜ Regio. (2005). *Kontuurkaart 'Balti riigid' [Outline map of the Baltic states].* Retrieved 09.04.2019, from `http://www.regio.ee/kaardid`

[32] Osman, Y., Fealy, R., and Sweeney, J. (2015). Modelling extreme temperatures in Ireland under global warming using a hybrid peak-over-threshold and generalised Pareto distribution approach. *International Journal of Global Warming*, **7**(1), 21–47.

[33] Ouarda, T. B. M. J., and Charron, C. (2018). On the mixture of wind speed distribution in a Nordic region. *Energy Conversion and Management*, **174**, 33–44.

[34] Pearson, K. (1893). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, **185**, 71–110.

[35] Pelosi, M., Alfo, M., Martella, F., Pappalardo, E., and Musaro, A. (2015). Finite mixture clustering of human tissues with different levels of IGF-1 splice variants mRNA transcripts. *BMC Bioinformatics*, **16**, 289.

[36] Pittau, M. G., Zelli, R., and Massari, R. (2014). Evidence of convergence clubs using mixture models. *Econometric Reviews*, **35**(7), 1317–1342.

[37] Punzo, A., Mazza, A., and Maruotti, A. (2018). Fitting insurance and economic data with outliers: a flexible approach based on finite mixtures of contaminated gamma distributions. *Journal of Applied Statistics*, **45**(14), 2563–2584.

[38] R Core Team. (2018). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. `https://www.R-project.org/`

[39] Rohrbeck, C. (2017). *Statistical methods for weather-related insurance claims* (Unpublished doctoral dissertation). Lancaster University.

[40] Scrucca, L., Fop, M., Murphy, T. B., and Raftery, A. E. (2016). mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, **8**(1), 205–233. `https://CRAN.R-project.org/package=mclust`

[41] Steutel, F. W., and Van Harn, K. (2004). *Infinite divisibility of probability distributions on the real line*. New York: Marcel Dekker.

[42] Trabzuni, D., and Thomson, P. (2014). Analysis of gene expression data using a linear mixed model / finite mixture model approach: Application to regional differences in the human brain. *Bioinformatics*, **30**, 1555–1561.

[43] Withers, C. S., Nadarajah, S., and Shih, S. H. (2015). Moments and cumulants of a mixture. *Methodology and Computing in Applied Probability*, **17**(3), 541–564.

[44] Yu, Y. (2018). *mixR: Finite mixture modeling for raw and binned data. R package version 0.1.1.* `https://CRAN.R-project.org/package=mixR`

[45] Zimmer, W. J., Keats, J., and Wang, F.-K. (1998). The Burr XII distribution in reliability analysis. *Journal of Quality Technology*, **30**, 386–394.

Appendices

Appendix 1. Summary Statistics of Raw Data by Stations

Table A1: Summary statistics for hourly mean wind speeds (m/s)

Station	n	nmiss	Minimum	Q1	Median	Mean	Q3	Maximum	SD
Ristna	131156	340	0.0	1.7	2.7	3.23	4.0	18.9	2.251
Tallinn-Harku	131450	46	0.0	1.9	2.9	3.12	4.2	11.9	1.671
Tõravere	131439	57	0.0	1.4	2.3	2.48	3.4	10.9	1.385
Vilsandi	131443	53	0.0	3.3	5.1	5.64	7.4	22.6	3.040
Võru	131440	56	0.0	1.3	2.3	2.40	3.4	10.1	1.459

n: number of values, nmiss: number of missing values, Q1: lower quartile, Q3: upper quartile, SD: standard deviation

Table A2: Summary statistics for hourly maximal wind speeds (m/s)

Station	n	nmiss	Minimum	Q1	Median	Mean	Q3	Maximum	SD
Ristna	131195	301	0.0	4.4	6.5	7.07	9.2	29.7	3.642
Tallinn-Harku	131485	11	0.0	3.9	5.9	6.29	8.3	24.5	3.227
Tõravere	131473	23	0.0	3.2	5.1	5.50	7.4	26.0	2.919
Vilsandi	131489	7	0.0	5.5	8.1	8.80	11.4	33.4	4.294
Võru	131476	20	0.0	3.1	5.1	5.41	7.4	23.6	3.079

n: number of values, nmiss: number of missing values, Q1: lower quartile, Q3: upper quartile, SD: standard deviation

Table A3: Summary statistics for hourly mean temperatures (°C)

Station	n	nmiss	Minimum	Q1	Median	Mean	Q3	Maximum	SD
Ristna	128619	2877	-22.8	1.6	6.6	7.26	13.7	31.0	7.737
Tallinn-Harku	131373	123	-26.9	0.4	6.3	6.55	13.5	33.7	8.834
Tõravere	131301	195	-31.1	0.1	6.2	6.46	13.9	32.9	9.624
Vilsandi	131410	86	-21.0	2.2	7.4	7.83	14.4	31.8	7.765
Võru	131423	73	-32.1	0.2	6.3	6.63	14.2	34.3	9.855

n: number of values, nmiss: number of missing values, Q1: lower quartile, Q3: upper quartile, SD: standard deviation

Appendix 2. Histograms of Raw Data by Stations

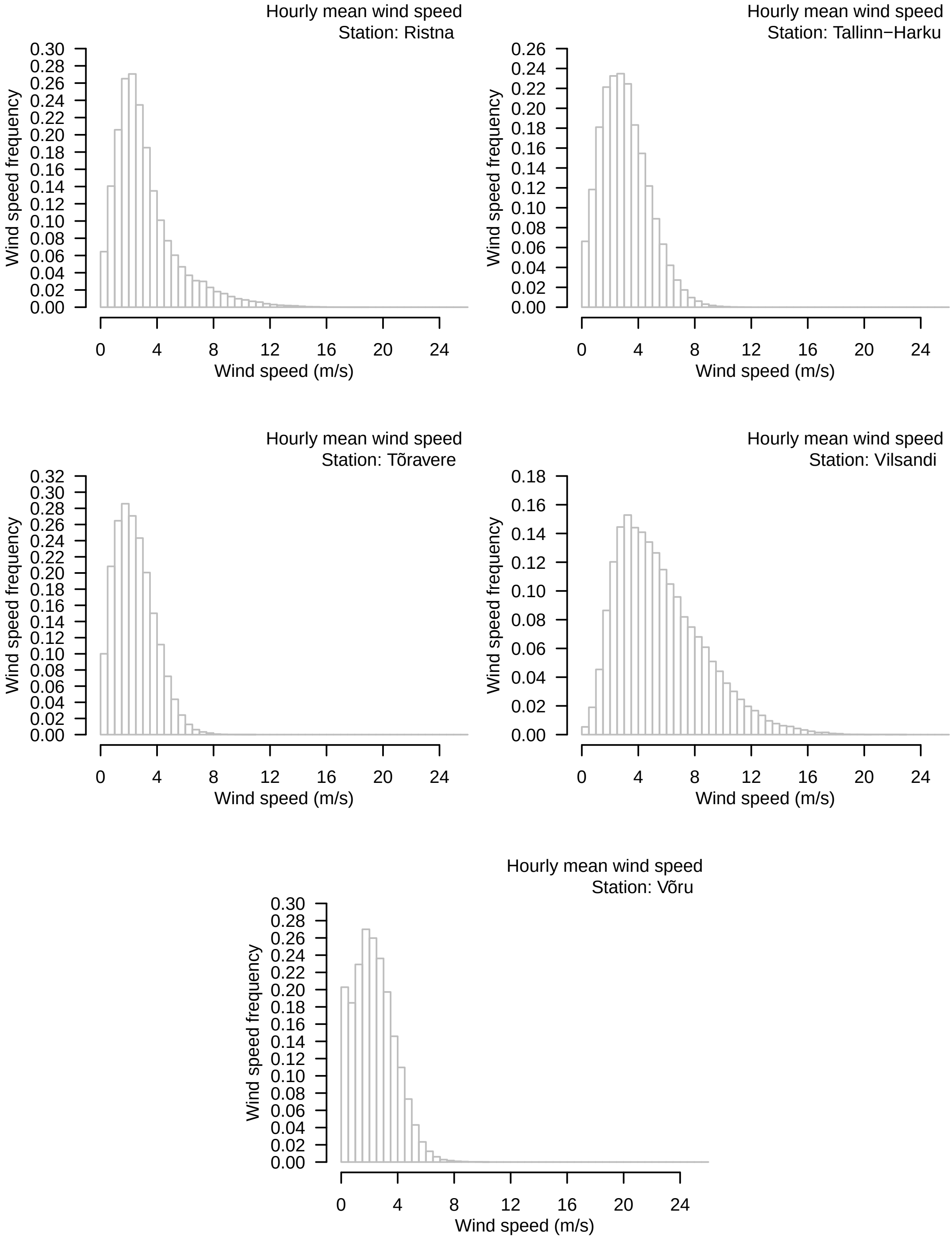


Figure A1: Histograms of hourly mean wind speeds (m/s) by stations

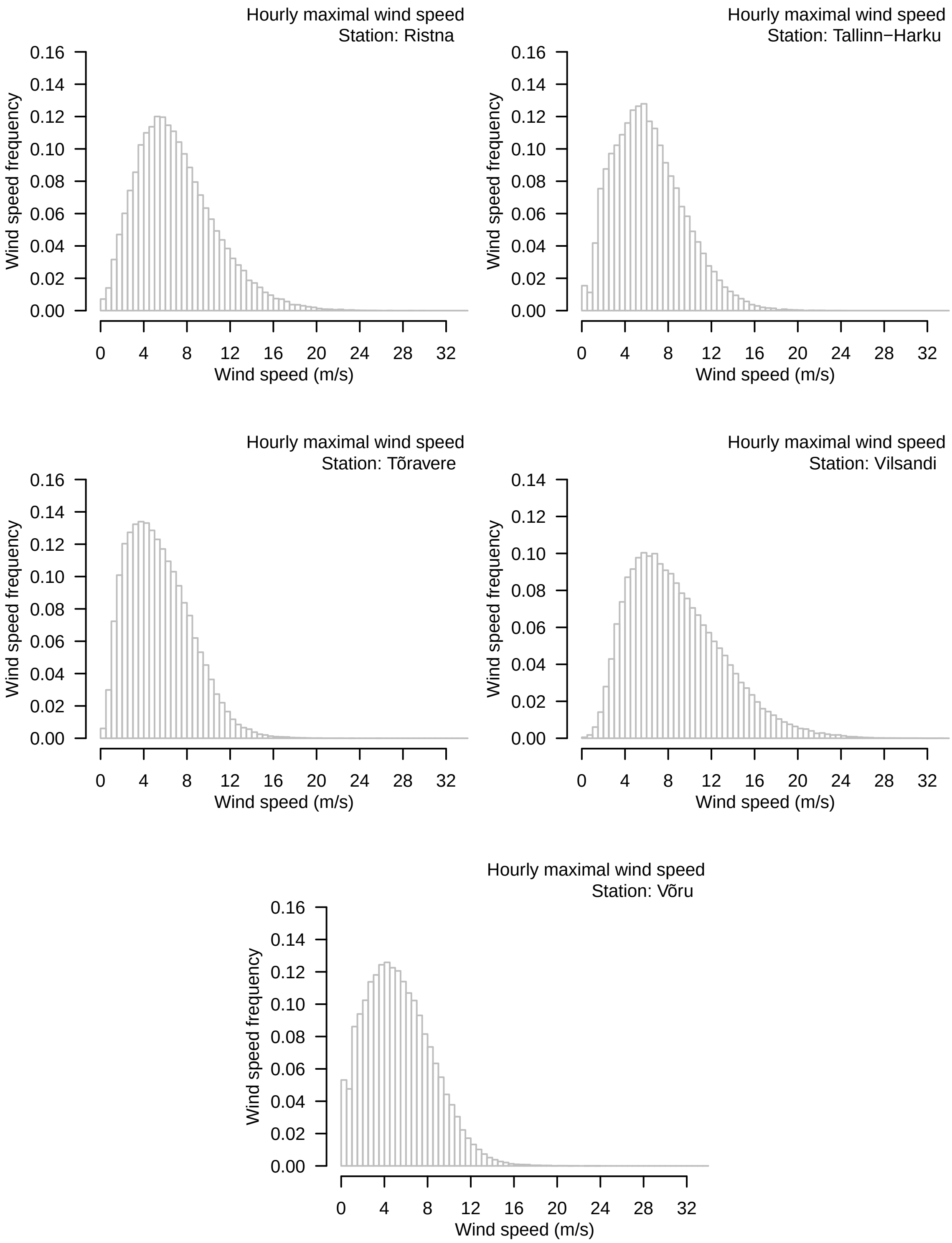


Figure A2: Histograms of hourly maximal wind speeds (m/s) by stations

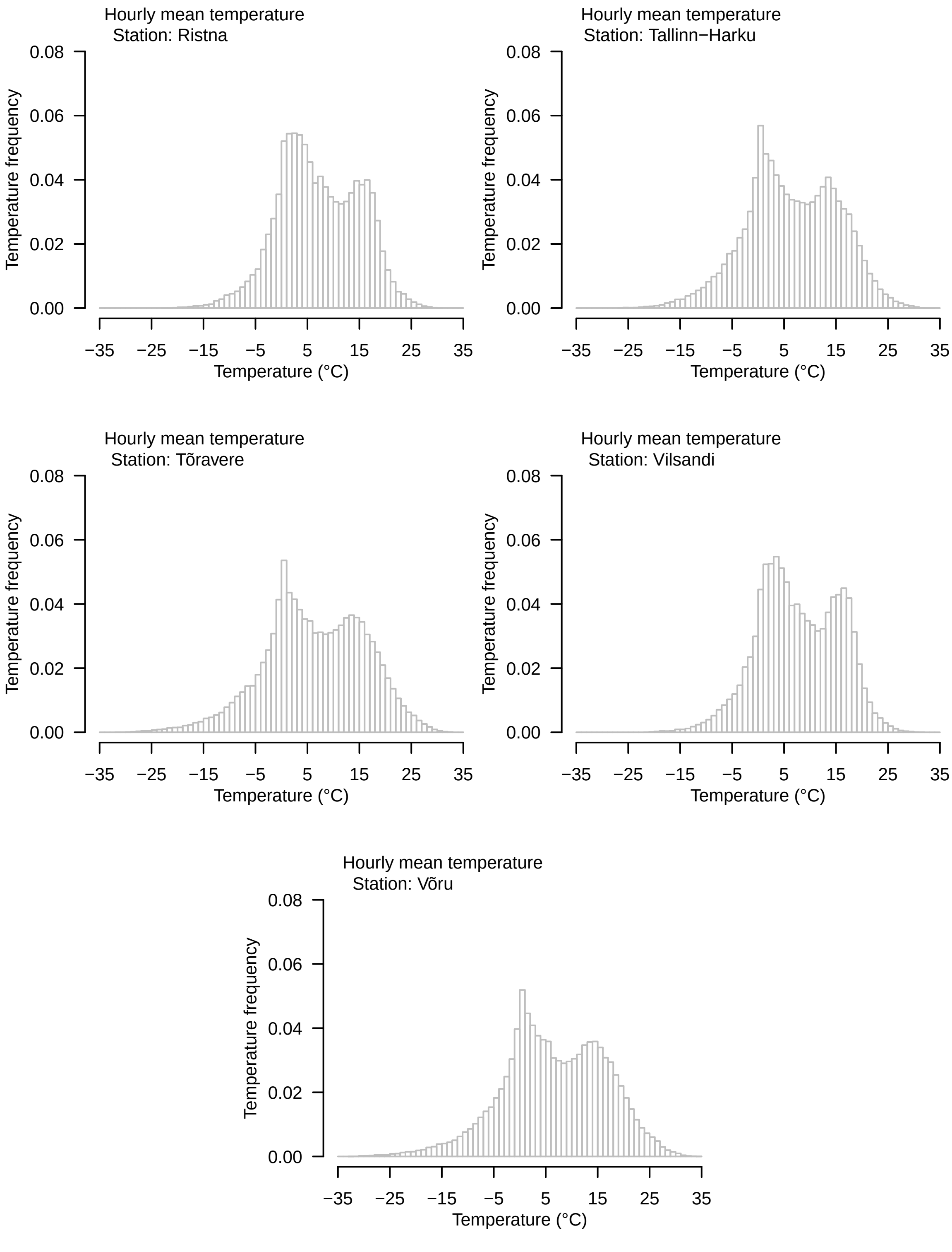


Figure A3: Histograms of hourly mean temperatures (°C) by stations

Appendix 3. Goodness of Fit Plots by Stations

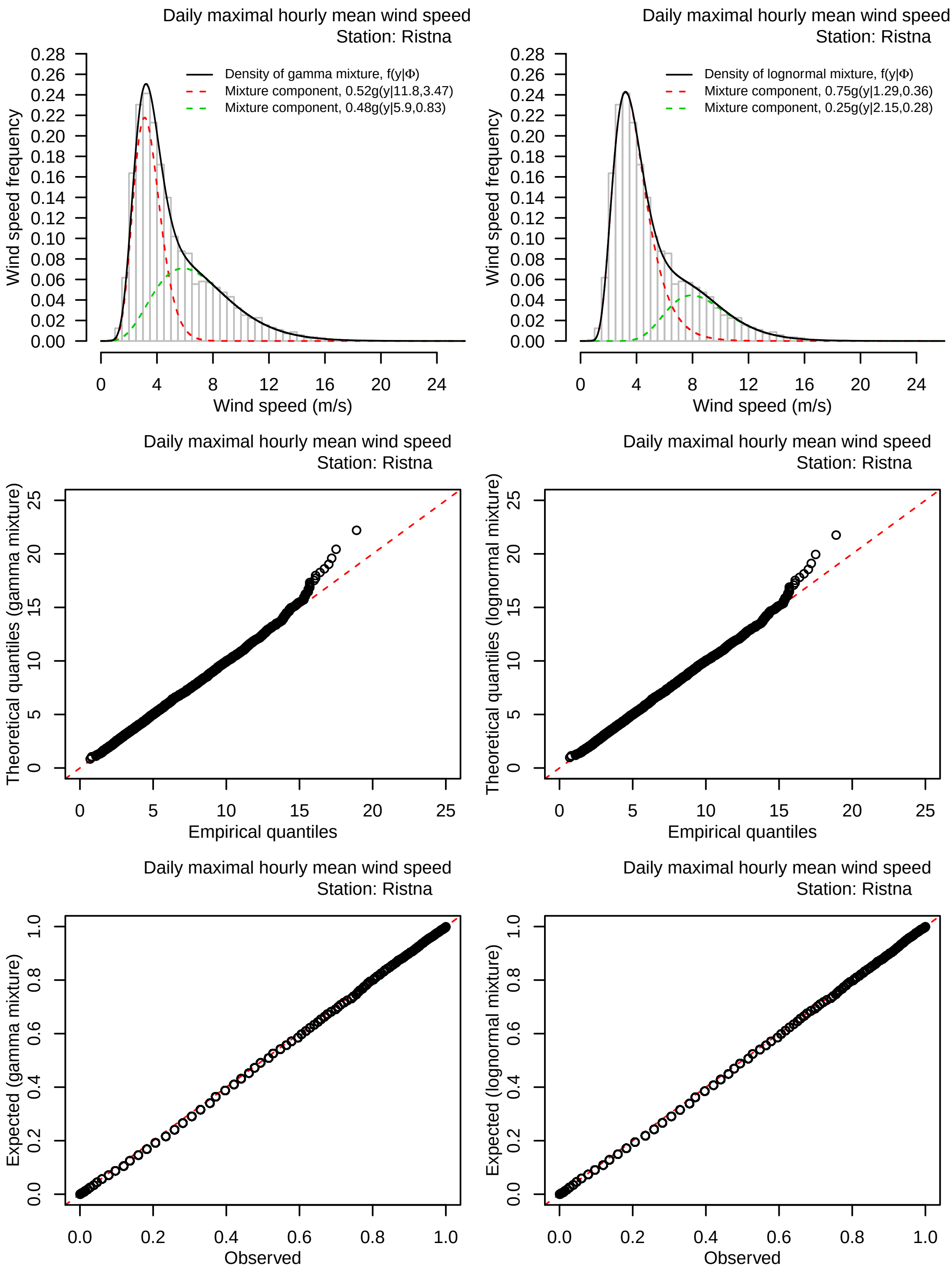


Figure A4: Two mixtures with highest BIC value in Table 3 for modeling daily maximal hourly mean wind speed, their Q-Q-plots and P-P-plots (from top to bottom: density over histogram, Q-Q-plot, P-P-plot)

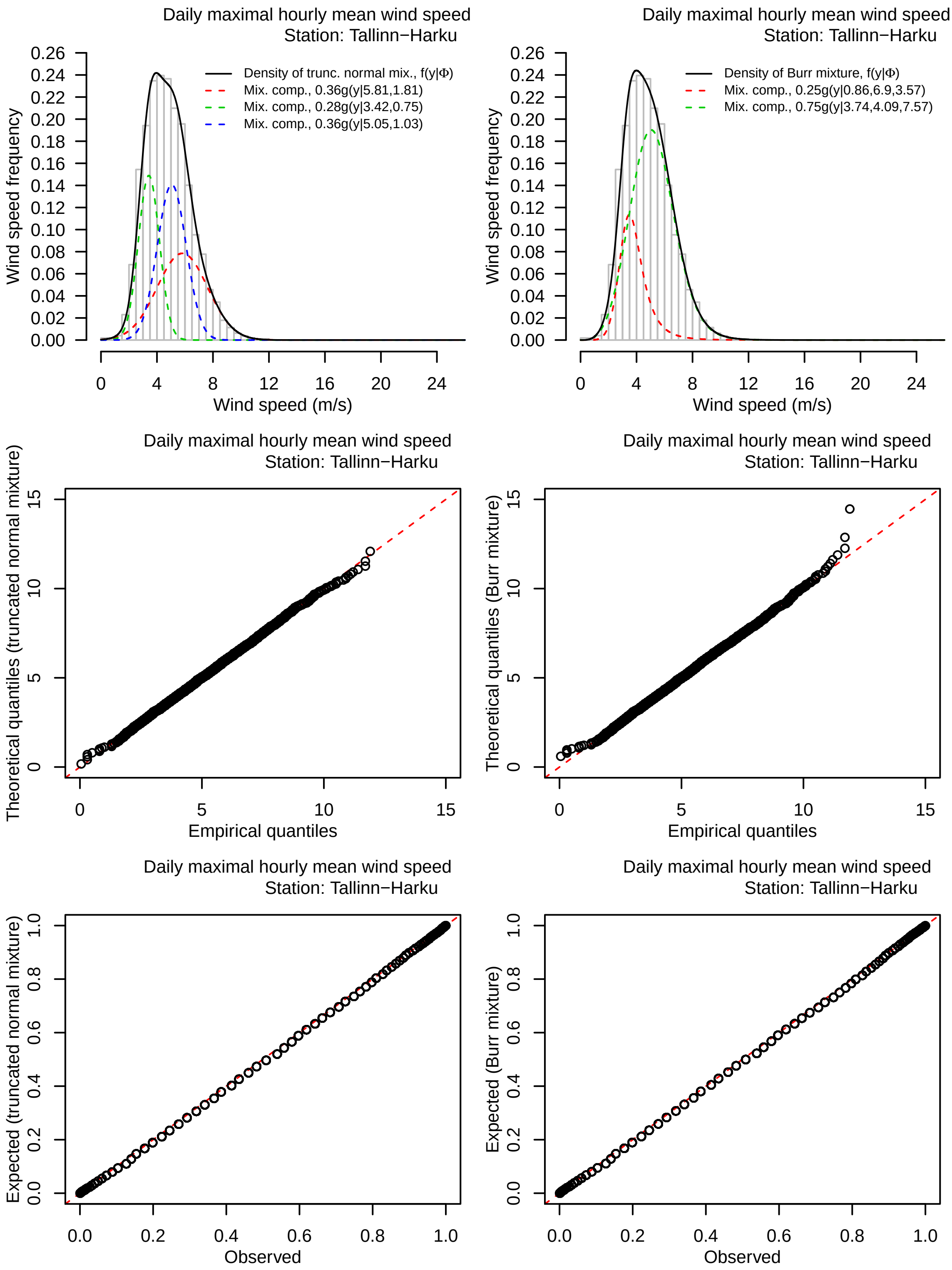


Figure A5: Two mixtures with highest BIC value in Table 4 for modeling daily maximal hourly mean wind speed, their Q-Q-plots and P-P-plots (from top to bottom: density over histogram, Q-Q-plot, P-P-plot)

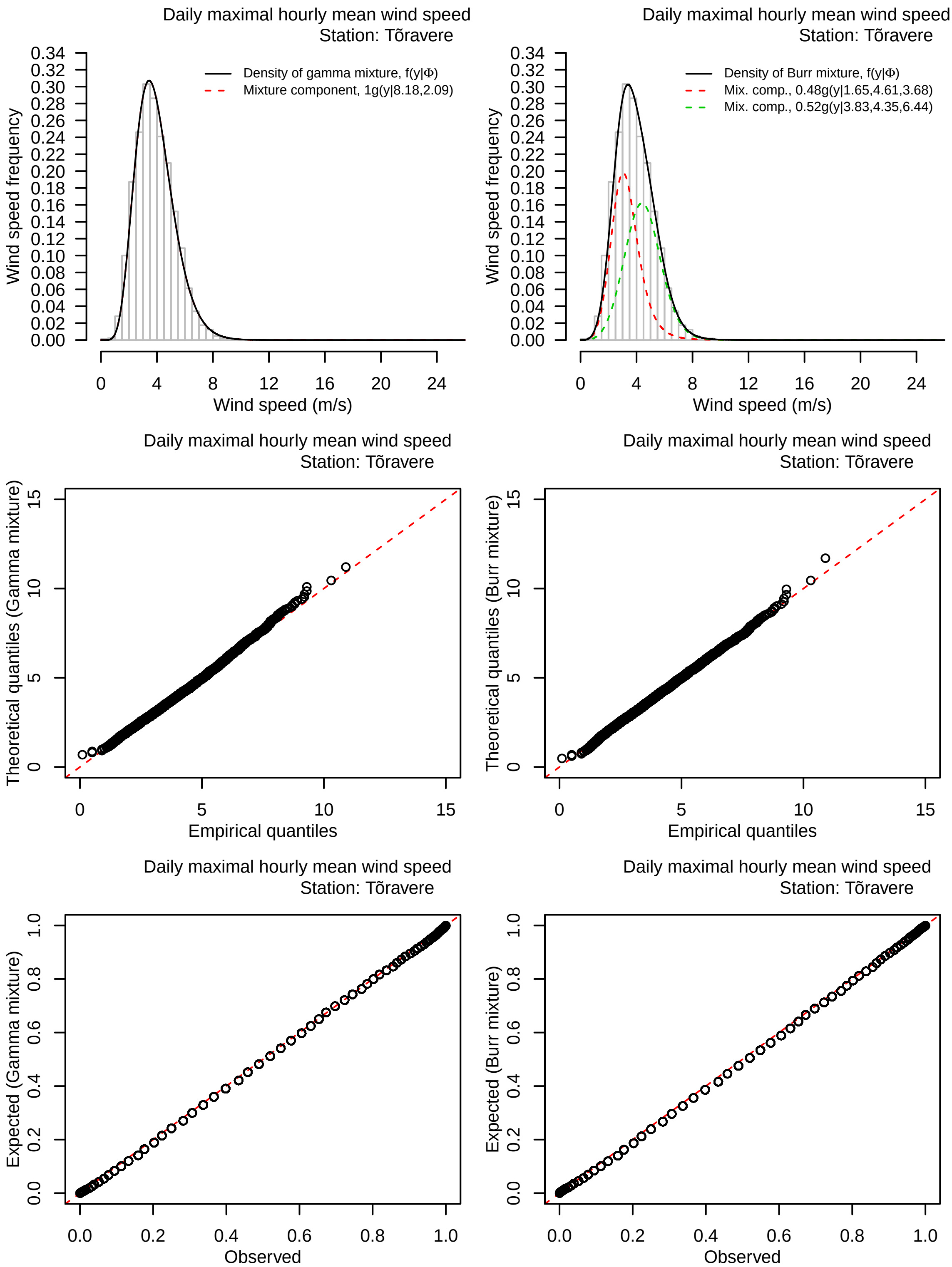


Figure A6: Two mixtures with highest BIC value in Table 5 for modeling daily maximal hourly mean wind speed, their Q-Q-plots and P-P-plots (from top to bottom: density over histogram, Q-Q-plot, P-P-plot)

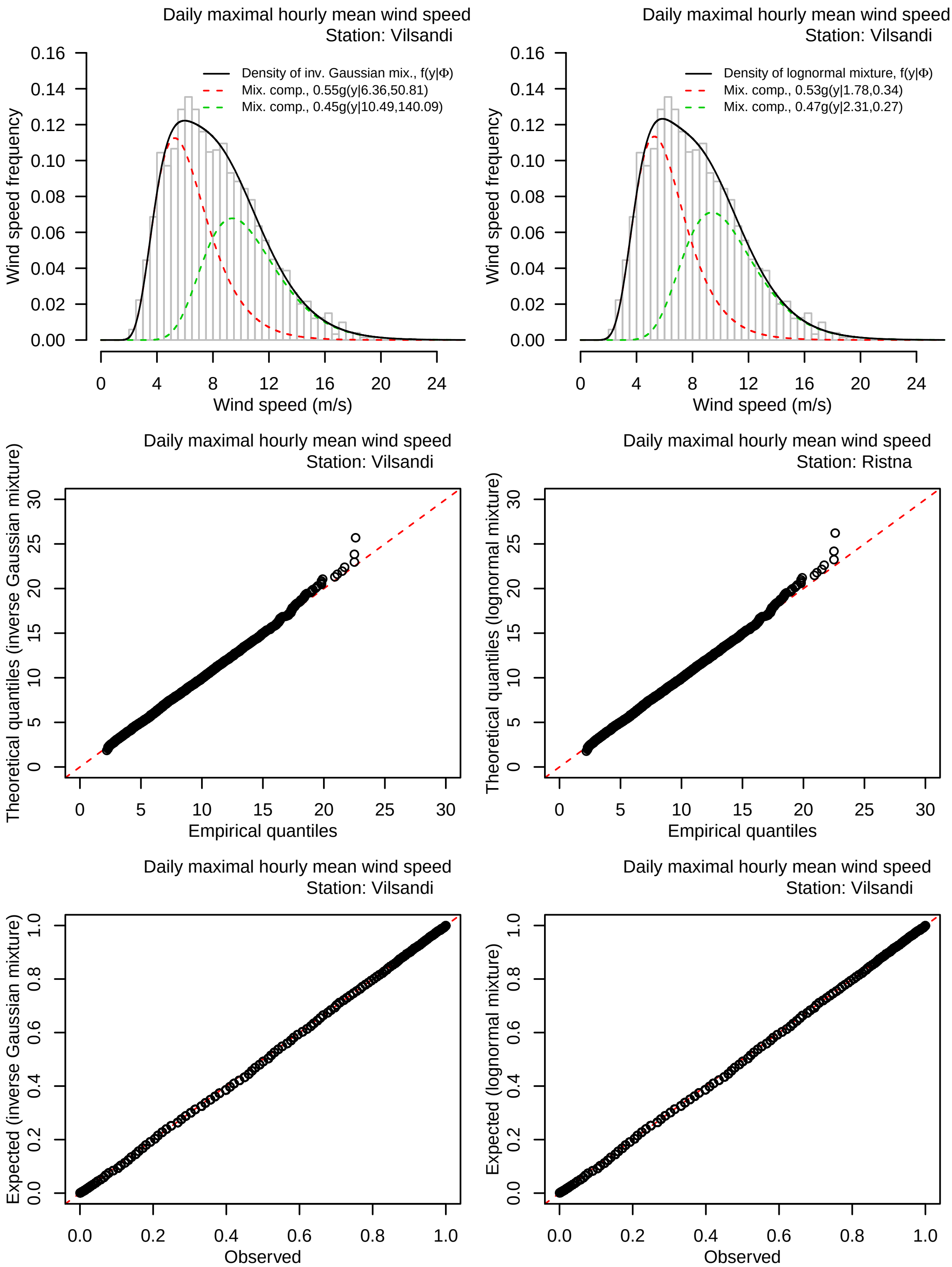


Figure A7: Two mixtures with highest BIC value in Table 6 for modeling daily maximal hourly mean wind speed, their Q-Q-plots and P-P-plots (from top to bottom: density over histogram, Q-Q-plot, P-P-plot)

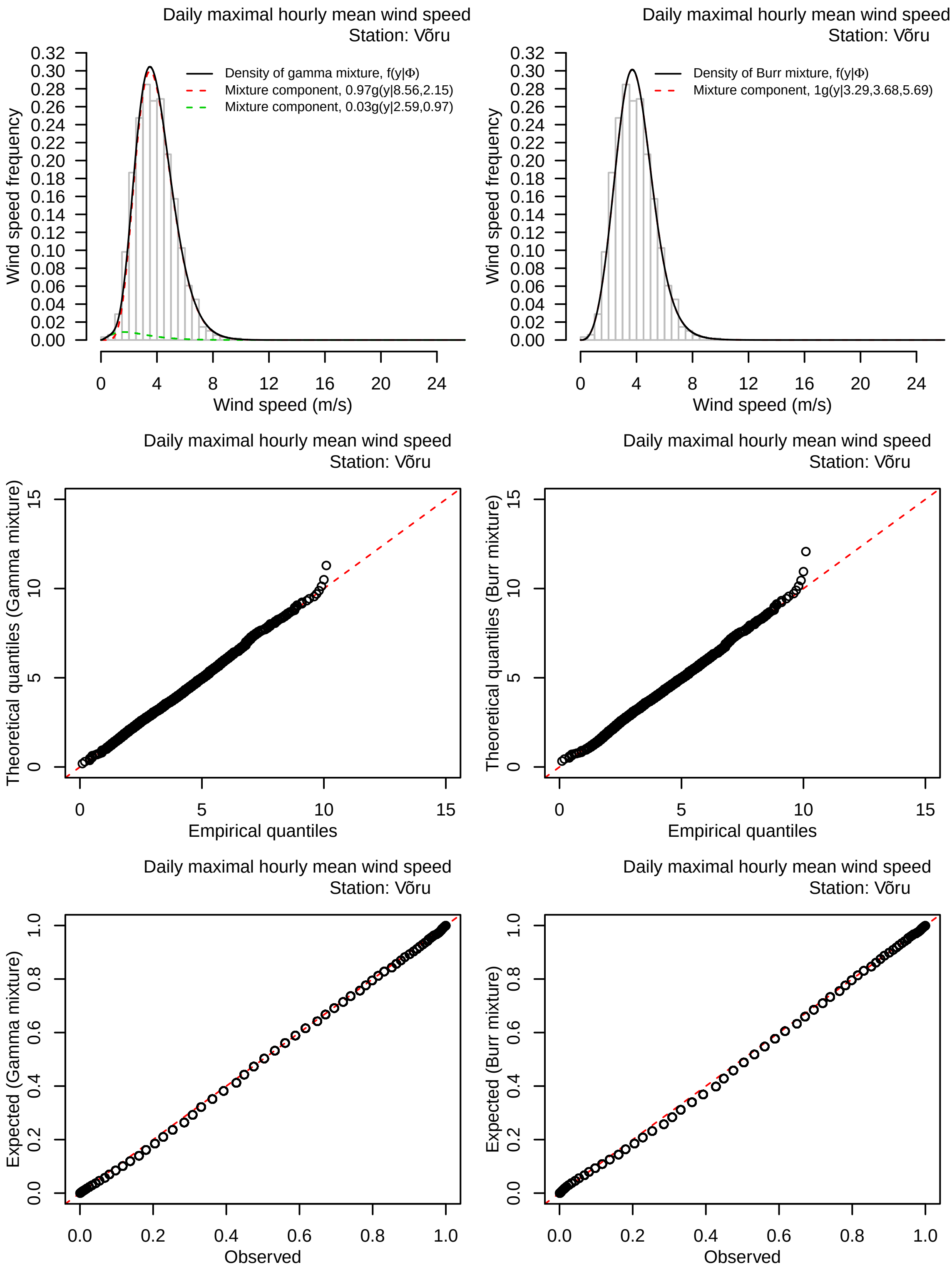


Figure A8: Two mixtures with highest BIC value in Table 7 for modeling daily maximal hourly mean wind speed, their Q-Q-plots and P-P-plots (from top to bottom: density over histogram, Q-Q-plot, P-P-plot)

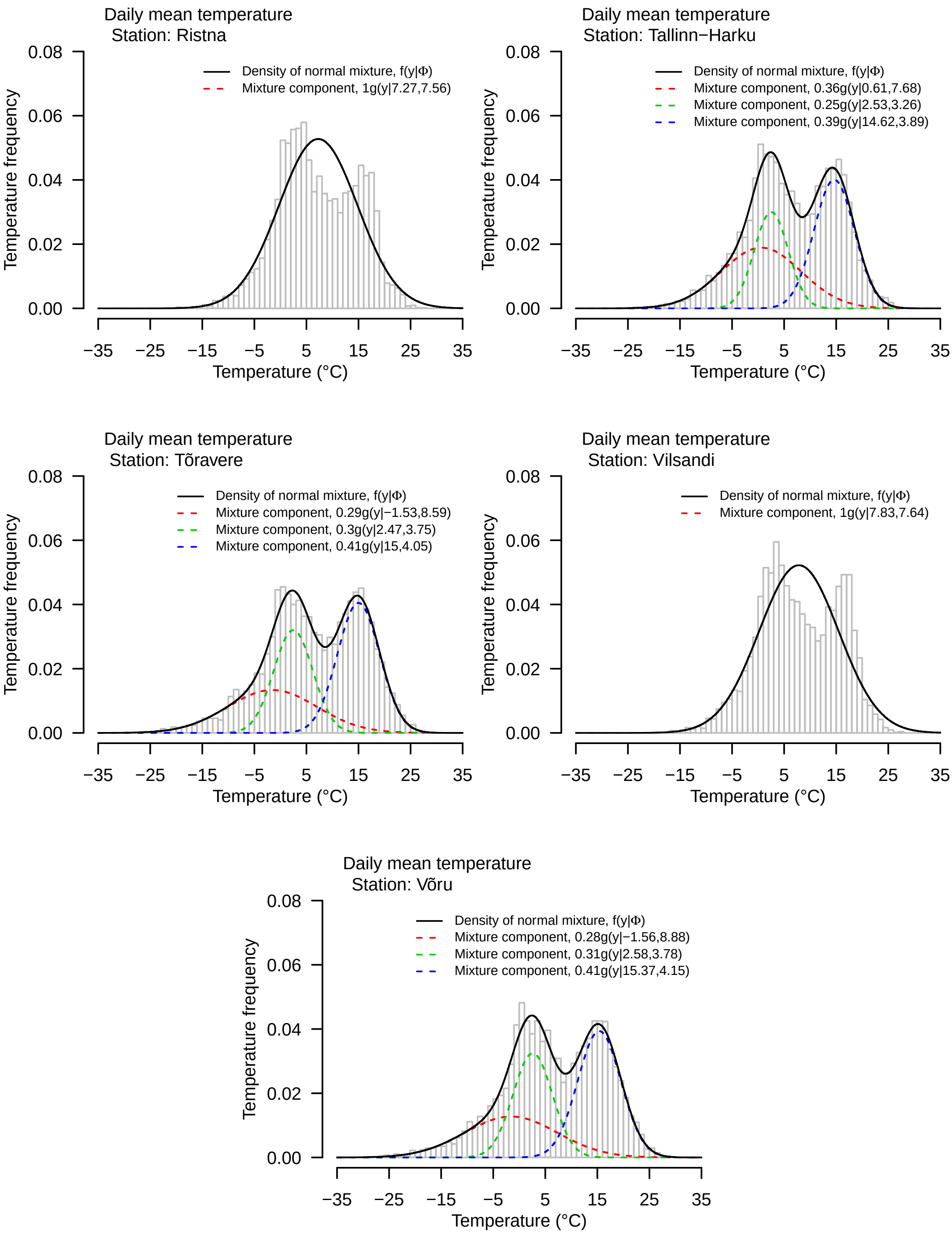


Figure A9: Mixtures with highest BIC values in Table 11 for modeling daily mean temperature over histogram by stations

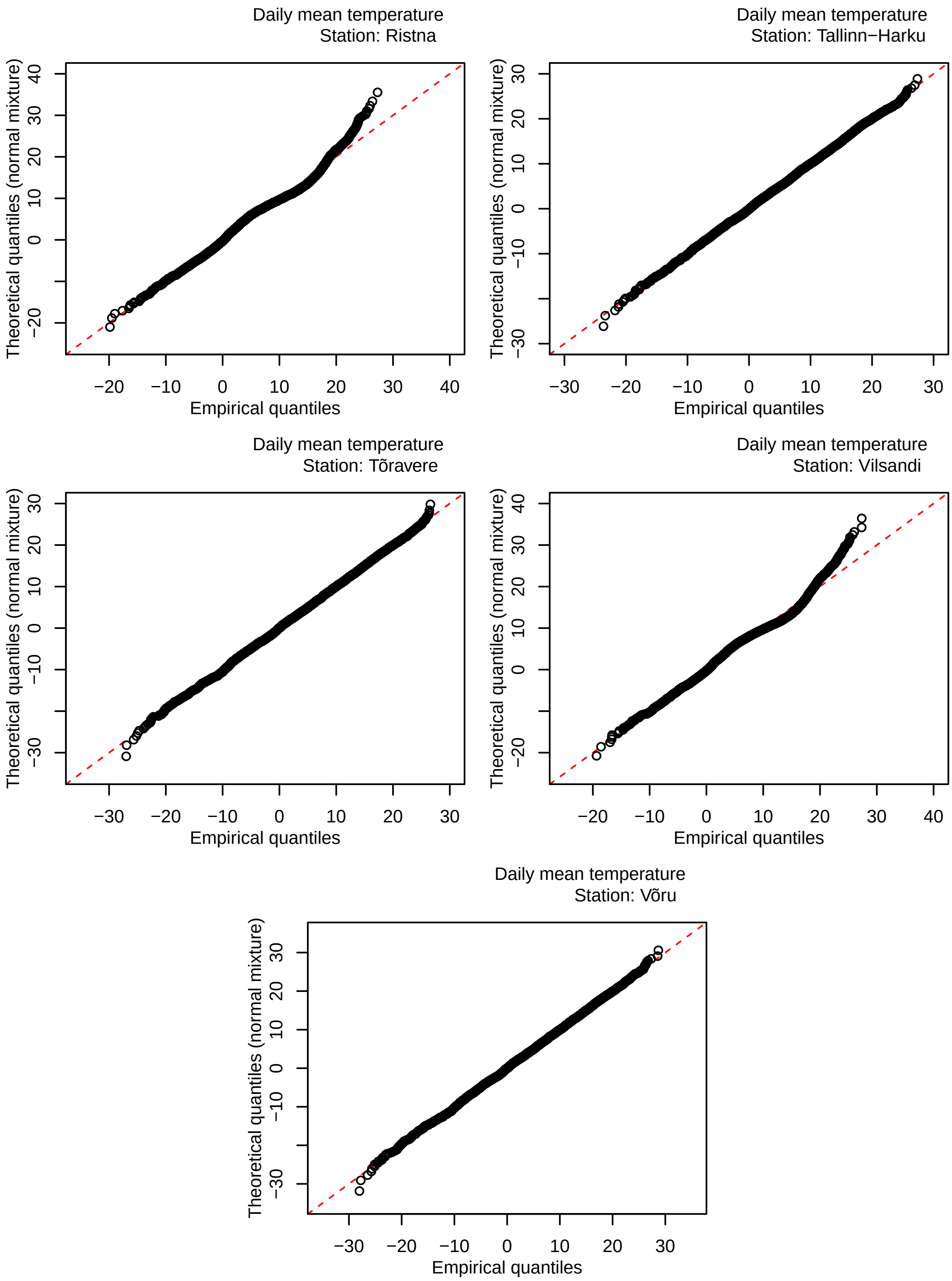


Figure A10: Q-Q-plots for mixtures with highest BIC values in Table 11 for modeling daily mean temperature by stations

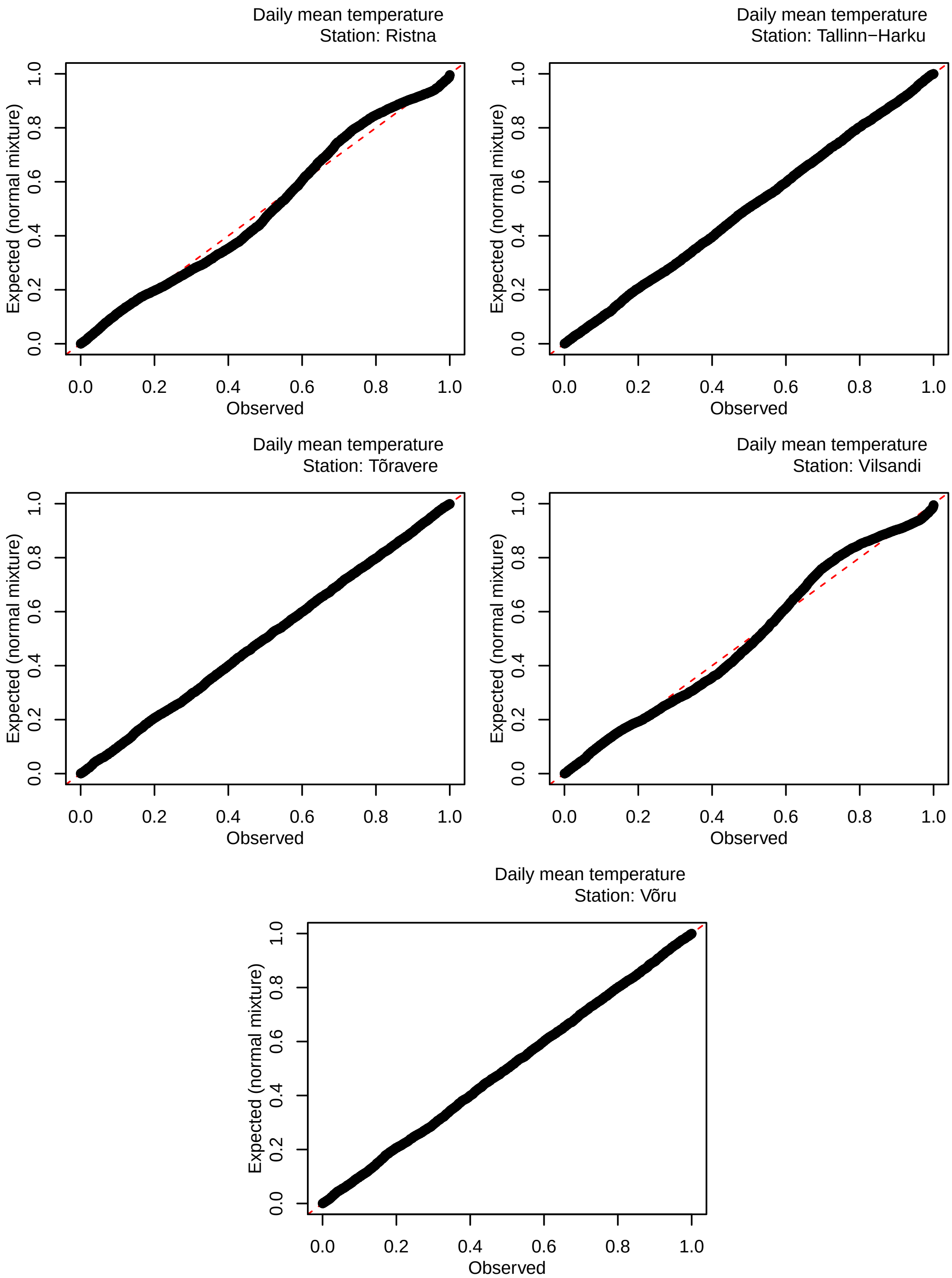


Figure A11: P-P-plots for mixtures with highest BIC values in Table 11 for modeling daily mean temperature by stations

Non-exclusive licence to reproduce thesis and make thesis public

I, Kristi Ernits,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright, Finite Mixtures with Application on Estonian Meteorological Data, supervised by Annika Krutto.

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Kristi Ernits 15/05/2019