Technische Universität München

Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt

Fachgebiet Biostatistik

Statistical modeling with finite mixtures of skew-t distributions and their application in the life sciences

Josef Martin Franz Höfler

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzende:		Prof. Dr. Annette Menzel
Prüfer der Dissertation:		
	1.	Prof. Donna P. Ankerst, Ph.D.
	2.	Associate Prof. Anastasios Panagiotelis, Ph.D.

Die Dissertation wurde am 20.11.2019 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 31.01.2020 angenommen.

Abstract

There is a need for flexible distribution and error models considering multivariate data in the life sciences. This thesis explores a flexible class of multivariate models, especially suited for data encountered in real-life situations, where outliers and skewness are prevalent and provides an algorithm for fitting such models to use them efficiently in practice.

We introduce univariate and multivariate mixtures of skew-t distributions, beginning with their mathematical definitions and followed by their hierarchical representation, which facilitates the implementation of the Expectation-Maximization (EM) algorithm for parameter estimation. A new proposal for classifying multivariate clinical data is proposed, based on fitting multivariate skew-t mixtures separately to patients from diseased case and non-diseased control groups. The ratio of multivariate densities for cases to controls forms a likelihood ratio, which multiplied by the ratio of prior probabilities of being a case versus control, leads to the posterior odds of being a case. The posterior odds of cases are back-transformed to the probability scale, resulting in individual predictions. The form of the density ratio is discussed for different situations, such as for the constraint of equal variance.

We construct efficient EM algorithms that can accommodate collapsed clusters. A collapsed cluster can be viewed as a distribution that places all of its mass on a lower-dimensional space with no variance. Our approach for the applications targeted in this thesis is that the same underlying data generating process applies to collapsed and non-collapsed clusters. We develop and publish a novel R package `fitmixst4`, which implements the EM algorithm for fitting multivariate mixtures of skew-t distributions and differentiates collapsed clusters from non-collapsed ones by restricting the variance of the latter to be above a specific bound.

We implement the algorithm in two applications. The first application concerns the update of a leading online clinical risk prediction model for prostate cancer on biopsy to incorporate two novel serum markers. We fit the multivariate skew-t mixtures to the bivariate distribution of the two markers in a sample of cancer cases and controls, respectively, thus forming the likelihood ratio. Prior odds of prostate cancer for individual patients are formed based on their standard clinical risk factor profiles from an existing online risk prediction tool. Multiplication by the likelihood ratio leads to updated individualized posterior probabilities of prostate cancer that combines information from standard risk factors with the new markers. We implement the resulting risk tool with the R package `shiny` and post it online at the Cleveland Clinic Risk Library to make it accessible to patients and clinicians worldwide.

For the second application, we fit mixtures of multivariate skew-t distributions with collapsed clusters to describe and classify trees that experience mortality versus not from a network of European Beech trees. We model up to five individual tree characteristics and competition indices to form risk prediction models for tree mortality. We visualize two-dimensional contour plots of predictive characteristics for trees experiencing versus not experiencing mortality in order to facilitate communication with forest researchers concerning indicators for mortality.

Using separate training and test sets, we show that skew-t based methods slightly outperform traditional logistic regression.

This thesis provides means for life science researchers to implement an intricate modeling framework in order to maximize prediction of outcomes, as well as an understanding of underlying complex nonlinear associations among risk factors. The published R package facilitates implementation, bringing the impact of the models to applications in many fields beyond those shown in this thesis.

Zusammenfassung

Es gibt erheblichen Bedarf an flexiblen Verteilungen und Fehlermodellen, um multivariate Daten in Life Science zu untersuchen. Die vorliegende Arbeit untersucht eine flexible Klasse von multivariaten Modellen, die sich besonders gut für reale Daten eignet, in welchen Ausreißer und Schiefen vorhanden sind. Dazu stellt sie einen Algorithmus für die Anpassung solcher Modelle zur Verfügung.

Wir führen zunächst univariate und multivariate Mixture Skew-t Verteilungen ein, beginnend mit deren mathematischer Definition, gefolgt von deren hierarchischer Darstellung, die die Implementation des Expectation-Maximisation (EM) Algorithmus für die Parameterschätzung ermöglicht. Eine neue Interpretation für die Klassifizierung von multivariaten Daten, basierend auf der Anpassung separater multivariater Skew-t Mixtures jeweils für Fall- und Kontrollgruppe, wird vorgeschlagen. Der Quotient von multivariaten Dichten für Fall- und Kontrollgruppe formt eine Likelihood Ratio, die multipliziert mit der priori Wahrscheinlichkeit zu posteriori Odds führt. Die posteriori Odds werden auf die Wahrscheinlichkeitsskala zurücktransformiert. Die Form der Dichtequotienten wird für unterschiedliche Situationen, wie etwa für den Fall von gleichen Varianzen, diskutiert.

Wir konstruieren einen effizienten EM Algorithmus, der mit Collapsed Cluster umgehen kann. Ein Collapsed Cluster kann man als eine Verteilung betrachten, das alle Masse in einem niedrigeren dimensionalen Raum ohne Varianz hat. Unser Ansatz für die Applikationen in dieser Arbeit ist die Annahme, dass der Prozess, der die Daten generiert, für Collapsed und Non-Collapsed Clusters derselbe ist. Des weiteren entwickeln und veröffentlichen wir ein neues R Paket `fitmixst4`, das den EM Algorithmus für die Anpassung von multivariaten Skew-t Mixtures Verteilungen implementiert und Collapsed Clusters von normalen Gruppen differenziert.

Wir verwenden den Algorithmus in zwei Anwendungen. Die erste Anwendung behandelt das Update eines führenden klinischen Online-Risikoprädiktionsmodel für Prostatakrebs mit Biopsien, in welches zwei neue Serummarker eingebaut werden. Wir schätzen multivariate Skew-t Mixtures für die bivariate Verteilung der beiden Marker für Krebs- und Kontrollfälle, um eine Likelihood Ratio zu bekommen. Die priori Odds für Prostatakrebs für individuelle Patienten werden basierend auf klinischen Standardrisikofaktorprofilen mit dem existierenden Online-Risikoprädiktionstool berechnet. Die Multiplikation mit der Likelihood Ratio führt zu angepassten individualisierten posteriori Wahrscheinlichkeiten für Prostatakrebs, die die Information der Standardrisikofaktoren mit den neuen Markern kombiniert. Wir implementieren das resultierende Risikotool mit dem R Paket `shiny` und stellen es online auf der Cleveland Clinic Risk Library zur Verfügung, um es weltweit für Patienten und Kliniker zugänglich zu machen.

Für die zweite Anwendung haben wir Multivariate Skew-t Mixtures mit Collapsed Clustern verwendet, die die Sterbewahrscheinlichkeit von Bäumen in einem europäischen Netzwerk für

Buchen beschreiben und klassifizieren. Wir modellieren bis zu fünf individuelle Baumcharakteristiken und Wettbewerbsindizes, um ein Risikoprädiktionsmodel für die Sterblichkeit der Bäume zu entwickeln. Zusätzlich haben wir zweidimensionale Konturdiagramme der prädiktiven Charakteristiken visualisiert, um eine Grundlage für die Kommunikation mit Forstwissenschaftlern zu schaffen. Mit Hilfe von separaten Trainings- und Validierungssets, kann gezeigt werden, dass der Ansatz mit den Skew-t Verteilungen die traditionelle logistische Regression übertrifft.

Die vorliegende Arbeit stellt Forschern in den Life Sciences ein komplexes Modeling Framework zur Verfügung, welches die Prädiktionsresultate maximiert und das Verständnis der zu Grunde liegenden nicht-linearen Assoziationen veranschaulicht. Das publizierte R Paket erleichtert die Implementation, um die Anwendbarkeit dieser Modelle auf andere Sachgebiete zu übertragen.

Contents

List of figures

List of tables

1 Introduction

There is a need for flexible distribution and error models considering multivariate data in the life sciences. The multivariate normal distribution is the most commonly used model due to its transparent analytic properties. Multivariate regression can be used to accommodate covariates in the prediction of the error mean. However, if the data or error residuals do not follow multivariate normal distributions, then these models yield incorrect inference. In some cases, sought transformations may be available so that the assumptions hold, but in many cases, there are no suitable transformations. Skewness and clusters not explainable by covariates frequently arise in data in the life sciences. This thesis ultimately explores a flexible class of multivariate models tailored for data encountered in real-life situations, where outliers and skewness are prevalent and provides an algorithm for fitting such models to use them efficiently in practice.

This work shows two applications to be analyzed, the first in forestry and the second in prostate cancer early detection research. Figure 1 shows a bivariate scatterplot of two competition indices widely used in forestry that are thought to be predictive of tree mortality. Bivariate scatterplots of the two indices are shown separately for the subpopulations of trees that experienced mortality in the subsequent five years versus those that did not. For both groups, the overlaid empirical smoothed contours are not the typical ellipses of a multivariate normal distribution.

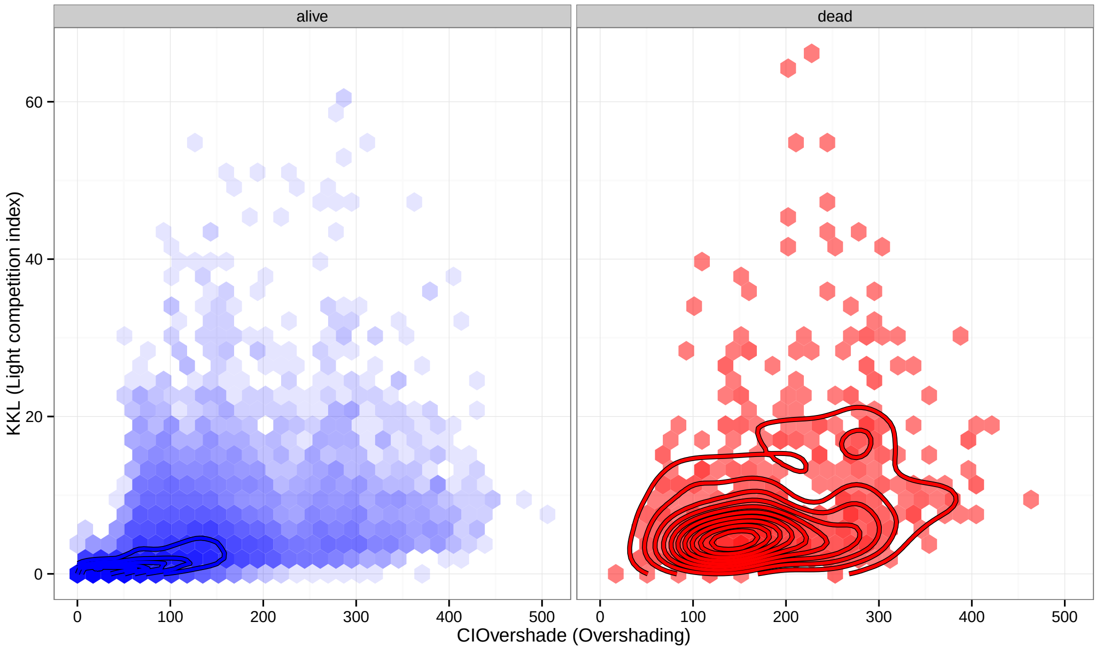


Figure 1 Distributions of two competition indices, CIOvershade (quantifies overshading) and KKL (quantifies light competition by neighboring trees), among 585 trees that experienced mortality over the next five years (red right) and $14,239$ trees that did not (blue left) in a Bavarian forest. Higher values of both indices mean more competition and less light for the tree. Contour lines of the kernel density estimates of the two groups are overlaid. The hexagons are darker at regions of higher density.

The other example deals with two blood markers, percent free prostate-specific antigen (PSA), and [-2]proPSA, which serve as early diagnostic indicators for high-grade prostate cancer (red), low-grade prostate cancer (yellow), or no prostate cancer (green) on biopsy. Bivariate normal ellipses are overlaid on data from the three groups. It is typical for biomarkers from patients with no cancer to be better described by normal distributions than those from patients with low- or high-grade cancer. The latter groups often display clusters due to the variable cancer growth and stage among them.

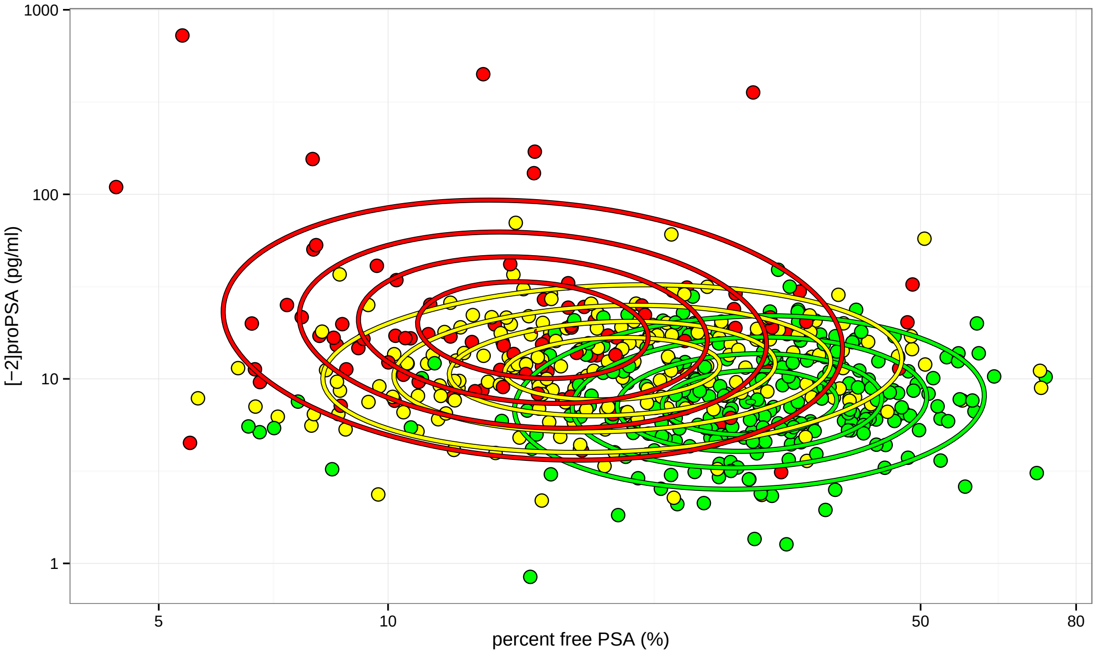


Figure 2 Distributions of two blood markers, percent free PSA and [-2]proPSA, among patients who would be diagnosed with high-grade prostate cancer (red), low-grade prostate cancer (yellow), and no prostate cancer (green) on biopsy. Best-fitting bivariate normal contours are overlaid on the groups.

Skew-t mixtures have emerged as flexible distributions for modeling non-normal multivariate data. Finite mixtures of skew-t distributions offer a flexible framework for data that possess skewness, multiple clusters, and outliers. Skew-t distributions are an extension of skew-normal distributions. The skew-t and skew-normal distributions to be used in this work belong to the class of skew-elliptical distributions described by Sahu et al. (2003). The class of elliptical distributions was first introduced by Kelker (1970), with properties derived by Cambanis et al. (1981) and Fang (2018). Elliptical distributions have been extensively studied and widely applied for characterizing multivariate data (Owen and Rabinovitch 1983; Van Praag and Wesselman 1989; Genton 2004; Vilca-Labra and Leiva-Sánchez 2006). Many popular distributions, including multivariate normal, multivariate t, and Pearson type II distributions, belong to this symmetric class. However, symmetric distributions are not suitable for all kinds of data, and it was the need for flexibility that first led to the increased interest in other alternatives.

The first formulation for skew-normal distributions occurred in a paper by Birnbaum (1950), who studied the effect of linear truncation on a multinomial population. Azzalini (1986) later derived some properties for the univariate case. Since many different manipulations of normal variables can generate the skew-normal distribution, it is not surprising that many authors arrived at the same or similar expressions (Nelson 1964; Roberts 1966; O'hagan and Leonard 1976; Aigner et al. 1977). Around ten years later, the distribution was generalized to the multivariate skew-normal distribution by Azzalini and Valle (1996), with applications in Azzalini and Capitanio (1999). The skew-normal distribution was embedded in a more general skew elliptical distribution by Branco and Dey (2001).

The skew-normal distributions can be classified into two subcategories, restricted and unrestricted, as described in Lee and Mclachlan (2013). In the restricted case, skewness arises from conditioning on one suitable random variable conditioned to be higher than zero; in the unrestricted case, the random variable is conditioned on as many random variables as dimensions. Only in the univariate case, both definitions coincide. The skew-normal distributions defined by Azzalini and Valle (1996) belong to the restricted family, whereas Sahu et al. (2003) introduced the unrestricted skew-normal distribution.

Since its introduction, there have been multiple variations on the skew-normal distribution (Azzalini 2005; Arellano-Valle and Azzalini 2006), including the generalized and extended skew-normal distributions. Generalized skew-normal distributions relax the condition of classic skew-normal distributions by not requiring the skewing function to be normal. Alternatively, extended skew-normal distributions, such as those by Arnold et al. (2002) or Liseo and Loperfido (2003), incorporate a separate extension parameter. An exciting feature of extended skew-normal distributions is that they are closed under conditioning; in other words, they have not only marginal but also conditional distributions of the same type. In the discussion to the paper on extended skew-t distributions by Arnold et al. (2002), Azzalini wrote that in the last five years, while there was a rapid development of theoretical results, they were not widely applied to real problems. He remarked that the real benchmark for stochastic models is their practical usefulness, ending with a plea for more work on applications to make use of these models.

One problem with normal and skew-normal distributions is that their tails do not tend to be large enough for many applications. Atypical observations affect the estimation of means and covariance matrices. Ensuring adequate protection against outliers becomes more difficult with the increasing dimension of the data distribution. Robust and symmetric t distributions are often applied as a first-line defense against outliers. However, as noted by Lange et al. (1989), not all forms of robustness are covered by t distributions. Skew-t distributions are an extension to skew-normal distributions that relax the symmetric constraint while still possessing thicker tails. The density function of the simple t distribution converges to the normal density as the degrees of freedom approaches infinity. Similarly, the skew-t density converges to the skew-normal density with increasing degrees of freedom. Consequently, the distinct parameterizations of skew-t distributions that are differently defined in the literature may converge to different parameterizations of the skew-normal distribution.

The various definitions of skew-t distributions can be classified as restricted versus unrestricted

(Lee and Mclachlan 2013). Analogous to the restricted skew-normal distributions, skew-t distributions defined by Azzalini and Valle (1996), and Gupta (2003) belong to the class of restricted skew-t distributions. It is interesting to note that skew-t distributions, which belong to the so-called skew-normal independent family by Cabral et al. (2012), are equivalent to skew-t distributions by Azzalini and Valle (1996) after a suitable reparameterization. One of the most notable extensions is the skew-t distribution introduced by Sahu et al. (2003), wherein the framework of elliptical distributions, skewness is induced by conditioning on a multivariate random variable and not on a univariate random variable.

As the theory on these new distributions evolved, so too came new methods for regression with non-normal distributed errors. Regression with skew-normal or skew-t distributions is helpful in cases where the assumptions of normal regressions do not hold, such as violation of symmetry or in the presence of outliers. Examples of skew-normal and skew-t regressions are presented in the publications Branco and Dey (2002), Jones and Faddy (2003), Sahu et al. (2003), Lachos et al. (2007), Ho and Lin (2010), Lachos et al. (2010), and Lin et al. (2014). Skew-normal and skew-t regression are harder to fit than using the least-squares theory of linear regression because there is no analytical solution for the regression parameters.

While skew-normal, as well as skew-t distributions, have been extensively studied, it is essential to make use of theoretical models via the development of software tools that can implement them. Hence there are now several packages available in the widely-used R open-source platform. One of the most well-known is the `sn` package (Azzalini 2019), which implements the skew-normal and skew-t distributions in the restricted family. Additionally, skew-normal distributions are implemented on the R package `VGAM` (Yee 2019), with corresponding theory (Yee and Wild 1996; Yee et al. 2010; Yee 2015), where they are used as error distributions for generalized linear and additive models.

Since their inception into the statistical community, finite mixtures of distributions of all types have gained popularity (Redner and Walker 1984; Aitkin and Rubin 1985; Titterington et al. 1985; McLachlan and Basford 1988). They are highly flexible alternatives to single distributions that can deal with skewness, thick tails, and multi-modality, which have spawned their use in biometrics (Santago and Gage 1993; Wang and Lei 1994; Caillette et al. 2005), medicine (Deb and Trivedi 1997; Thompson et al. 1998; Pham et al. 2000; Schlattmann 2009), genetics (Anderson and Thompson 2002; Pernkopf and Bouchaffra 2005; Tohka et al. 2007) and finance (Beard et al. 1991; Eberlein, Keller, et al. 1995; Wedel and DeSarbo 2002; Finkenstadt and Rootzén 2003). Along with flexibility comes new problems, such as the need for constraints for identifiability, the need for a larger sample size to estimate an increase in parameters, longer convergence times or lack of convergence, as well as the need for efficient algorithms for fitting.

The problem of identifying components and parameters of mixture distributions was first explicitly addressed by Pearson (1894) in an application characterizing the matter of forehead to body length ratios in female shore crab populations. Matching the empirical moments with the moments of the mixture distribution was one of the first methods of fitting normal mixtures. After the development of the maximum likelihood estimation by Fisher between

1912 to 1922 (Aldrich 1997), it took many years until starting in 1984 that mixture models quickly became more widely used.

The famous Expectation-Maximization (EM) algorithm introduced by Dempster et al. (1977) is still the most ubiquitous likelihood-based method for fitting mixtures in use today. It has been employed in countless applications in the life sciences and extended for a multitude of purposes. McLachlan and Peel (2004) and McLachlan and Krishnan (2007) provided a comprehensive overview of EM algorithms fitting mixture models. The first paper using mixtures of t distributions was published by Peel and McLachlan (2000). Univariate and multivariate mixtures of skew-t distributions were later described by Lin et al. (2007) and Lin (2010) using the representations by Azzalini (1986) and Sahu et al. (2003), respectively. There are several R implementations of finite mixture models, including the R package `mixsmsn` by Prates et al. (2013), which uses restricted skew-normal and skew-t mixtures and the R package `EMMIXuskewt` (Lee and McLachlan 2011, 2014; Lee et al. 2014) for mixtures of skew-t distributions in the unrestricted case by Lee and McLachlan (2012). The various packages have advantages and disadvantages, and undoubtedly more will appear as specific applications, such as those to be introduced in this thesis, warrant.

A not uncommon problem that arises in data from the life sciences, and encountered by us, is the existence of collapsed clusters, which reside on lower-dimensional planes than the remaining clusters. For example, in our forestry example shown in Figure 1, it is not unusual for many trees to have no occurrences of a specific type of competition. Having no occurrence can be seen to a small extent for CIOvershade among trees that eventually died, but there are other more pronounced examples in the data set (Wang and Lei 1994). To handle this type of cluster, we extend the Expectation Conditional Maximization Either (ECME) algorithm by Lin (2010) and Lee and McLachlan (2011).

A collapsed cluster can be viewed as a distribution that places all its mass on a single point in univariate space with no variance. Such distributions have long been studied in a variety of fields. For example, it has been theoretically proven that the likelihood ratio statistic for testing whether a variance component equals zero follows a mixture of the usual chi-square distribution and a point mass at zero (Self and Liang 1987). Zero-inflated models, such as zero-inflated Poisson regression by Lambert (1992) or hurdle models (Dalrymple et al. 2003; Cameron and Trivedi 2013), have been used modeling utilization and cost data, respectively, where there may be considerable numbers of zeros. Zero-inflated count models typically assume that the same process generates zero and non-zero values. The basic idea of hurdle models is that there are two different processes for zero and non-zero values. Zero-inflated Poisson models are a mixture of a Bernoulli and Poisson distribution.

Our approach for the applications targeted in this thesis instead follows the zero-inflated philosophy that the same underlying data generating process applies for zero and non-zero values. For the forestry example, this means that there is little difference between a tree with a lower level of competition, say 0.01% versus a tree with no competition whatsoever. For our cancer biomarker example, a measurement equal to zero instead means that the measure is below the detection limit.

A problem with all the existing mixture distribution algorithms for this kind of data is that they fail to converge when the variance of one or more components approaches zero. For this situation, we introduce an extension for the EM algorithm, which effectively constrains the variance. We develop an R package `fitmixst4` (Höfler 2019) that accommodates collapsed clusters by restricting the variance of the cluster to be above a certain bound and extend the ECME algorithm to operate more quickly and handle regression problems with skew-t distributed error terms.

The R package `fitmixst4` incorporates these extensions. The R package was developed with the latest C++ interface for R using the highly efficient linear algebra library Armadillo (Sanderson 2010) as well as the GNU scientific library (Galassi et al. 2009) for optimization. The fitting algorithm was restructured in a way that the implementation could take advantage of the OpenMP (OpenMP 2013) framework for multi-threading support. This feature helps significantly to reduce the computational time on modern multi-core machines, reducing such times by a factor almost equal to the number of cores.

The new efficient algorithms for fitting mixtures of skew-t distributions with collapsed clusters form the foundation for a new proposal for classifying multivariate non-normal data. The idea is to fit separate multivariate mixtures to multidimensional data from two (or more) groups to form a classifier that can be used to predict group membership for future observations. A weighted density ratio estimates the probability of being in a group. The flexibility of the mixture models reduces the need for transformations of the data, for example, as would be required if multivariate normal distributions were fit to data from the groups. Besides, the approach allows straightforward inspection of the data, as shown in Figure 2 for low dimensional data. Some theoretical properties, as well as the connection to logistic regression, are established, and we apply the developed methods to two examples.

For the forestry application, the introduced collapsed cluster is needed to depict the data correctly. Risk predictions can be performed with the calibrated density ratios and compared to the results of the logistic regression in the publication by Böck et al. (2014). The general outline of this thesis is as follows.

In Chapter 2, we introduce univariate and multivariate mixtures of skew-t distributions. The most important definitions, as well as the notation, will be used in the following chapters. The hierarchical representation of mixtures of multivariate skew-t distributions is presented since the implementation of the EM algorithm makes use of it. The proofs for some results are shown while identifying some errors in the existing literature. A new proposal for classifying multivariate clinical data is suggested. The idea of this method is to fit skew-t mixtures separately to cancer cases and controls. Calculating the density ratio of the fitted densities and back-transformed to the probability scale results in predictions for specific characteristics. The form of the density ratio is discussed for different situations, such as for the constraint of equal variance. Furthermore, the connection between the likelihood ratio and logistic regression is shown.

In Chapter 3, we construct efficient EM algorithms that can accommodate collapsed clusters. Furthermore, it is also shown how to include a multivariate skew-t regression in this framework. A short outline of the implementation details is given. Start and stop criteria for the algorithms are proposed. Several strategies on how to handle convergence issues of EM algorithms are discussed. The EM algorithm is implemented as an R package. The results of extensive simulation studies for the estimation of skew-t mixtures are shown.

Chapter 4 introduces the software solutions developed as part of this thesis. The R package `fitmixst4` with its essential elements, implementation details, data structure, and functionality is shown. Furthermore, for the application in prostate cancer, a web application was developed utilizing the R package `shiny` to make the risk calculator accessible for the public. The web application structure with a user interface and server component, as well as the application flow, is described.

Applications of the approach are provided in Chapter 5. The first application shows additional work on the prostate cancer example published in Ankerst et al. (2014). The theory of updating risk prediction models by utilizing likelihood ratios calculated by risk factor distributions is introduced. The Bayesian updating method is described for a multinomial model with three outcomes. The performance of different regression types is assessed, and the best fitting model is used for the update. Additionally, a short outline of the implementation of the risk calculator as a web application via the R package `shiny` is given. A second application is provided from forestry Böck (2014). Mixtures of multivariate skew-t distributions with collapsed clusters are used to form risk prediction models for tree mortality. The R package is retrospectively applied to describe clusters of trees that died and remained alive over a series of five-year follow-up periods from beech trees in a Bavarian long-term forest research plot network. Heat maps of the mixture density ratios corresponding to dead and alive trees are assessed as a potential prediction tool for forest mortality for future forest management.

In the discussion, a short outline of the work is given and set in context to the current literature. Pitfalls of finite mixtures and the algorithms are shown. Moreover, the need for efficient tools for data analysis is discussed.

2 Mixtures of skew-t distributions

Mixture distributions are convex combinations of probability density functions (pdfs) since the coefficients are non-negative, coefficients sum up to one, and linearity holds. Properties of pdfs, like that the integral over its support equals one, are preserved for mixture distributions. Mixture distributions naturally arise from applications where several underlying subpopulations induce a random variable. Normal mixtures were used early by Pearson (1894) to model the ratio of the forehead to body lengths in female shore crab populations utilizing the method of moments. Many extensions and algorithms were published to make normal mixtures more flexible and faster to estimate. Primarily skew-t distributions are used in many applications because of their flexibility. Not only because of their flexibility, mixture models are popular but also because of their interpretation. The underlying pdfs of mixtures can have, for example, a direct demographic interpretation such as identifying gender subgroups or an interpretation pointing out other specific characteristics. Otherwise, the components are a convenient way for flexible estimation. Drawbacks of mixture distributions are that sometimes mixture components are not identifiable. In this case, one limitation can be the interpretability, since the mixture components have no real-world explanation. Besides this, the estimation can be tricky because there is no closed-form solution. An essential aspect of mixture distributions is that they are still parametric but with significantly improved flexibility.

In this chapter, we provide the theoretical development of representations of skew-t mixtures. We derive specific properties for the univariate case followed by the multivariate case. In general, we follow the publications by Lin (2010) and Lee and McLachlan (2011) but elaborate the derivation in more detail, correct errors, and add additional information as needed for this thesis.

2.1 Univariate mixtures of skew-t distributions

The skew-t mixture belongs to the class of finite mixture models and specifies that each component in the mixture is a skew-t density, $f_{\mathcal{ST}}$. In general, finite mixture models are summing up different density functions with specific weights that the integral over the support of this sum equals one. Note that the sum has finite elements. The skew-t density is a t density with skewness added. The model-fitting algorithm exploits the stochastic properties of the density that will be later defined in detail. We begin here with the definition of the skew-t mixture distribution.

Definition 2.1 (Univariate mixtures of skew-t distributions)**.** *A random variable X is said to follow a skew-t mixture distribution with c components if its density is defined as*

$$f(x) = \sum_{i=1}^{c} \omega_i f_{\mathcal{ST}}(x \mid \mu_i, \sigma_i^2, \lambda_i, \nu_i),$$

where $f_{\mathcal{ST}}(\cdot)$ is the skew-t density with location μ_i, scale σ_i^2, skewness λ_i, and degrees of *freedom ν_i, and ω_i are the mixing proportions with $\sum_{i=1}^{c} \omega_i = 1$ and $\omega_i > 0$ for $i = 1, \ldots, k$ with $k < \infty$.*

We can interpret mixture distributions as being induced by an underlying set of random variables. Each of the underlying random variables follows an individual distribution. Hence, mixture models are used in the context of subpopulations. Since each subpopulation can follow a different distribution, mixture models account for the heterogeneity of the whole population. One possible application of this is that mixtures can do model-based clustering based on the underlying distribution properties. In this thesis, we choose skew-t distributions for the individual components because of their high flexibility.

The skew-t density is obtained by first extending the student t distribution by an arbitrary shift (μ) and scale (σ^2) parameter.

Definition 2.2 (Univariate t distribution)**.** *Let T be a t distributed random variable with degrees of freedom ν denoted by $T \sim \mathcal{T}(\nu)$, $\mu \in \mathbb{R}$, $\sigma \in (0, \infty)$ and set $X = \sigma T + \mu$. Then X is said to be t distributed with location μ, scale σ^2, and degrees of freedom ν, denoted by $X \sim \mathcal{T}(\mu, \sigma^2, \nu)$.*

The cumulative distribution function (cdf) of the extended t distribution can be written in terms of the student t distribution (with $\mu = 0$, $\sigma^2 = 1$). This property is useful for numerical calculation because efficient algorithms for evaluating the student t distribution exist. Let X be defined as in Definition 2.2 with location μ and scale σ^2. Then the cdf of X can be written as $F_X(x) = P(X \leq x) = F_T\left(\frac{x-\mu}{\sigma}\right)$, where $F_T(\cdot)$ is the cdf of a t distribution, omitting the degrees of freedom for simplicity. The pdf of X is derived from the cumulative distribution function $F_T(\cdot)$ as follows applying the chain rule:

$$\begin{aligned} f_X(x) &= \frac{d}{dx}\left(F_X\left(x\right)\right) = \frac{d}{dx}\left(F_T\left(\frac{x-\mu}{\sigma}\right)\right) \\ &= f_T\left(\frac{x-\mu}{\sigma}\right)\frac{d}{dx}\left(\frac{x-\mu}{\sigma}\right) \\ &= \frac{1}{\sigma} f_T\left(\frac{x-\mu}{\sigma}\right), \end{aligned}$$

where $f_T(\cdot)$ is the density of the random variable T. Inserting the definition of the t pdf yields the pdf of the extended t distribution:

$$f_X(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sigma\sqrt{\nu\pi}}\left(1 + \frac{(x-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}},$$

where $\Gamma(\cdot)$ is the gamma function (see Definition A.1).

The $\mathcal{T}(\mu, \sigma^2, \nu)$ distribution is symmetric about μ with expectation μ and variance $\frac{\nu}{\nu-2}\sigma^2$ for $\nu > 2$ (otherwise it is undefined). It is helpful to see that the student t distribution can be expressed as a scale mixture of normal distributions, with the gamma distribution as the mixing distribution. This property indicates how to generate random samples from t distributions, by first drawing random numbers from a gamma distribution, $\mathcal{G}(\alpha, \beta)$ with mean $\frac{\alpha}{\beta}$ and variance

$\frac{\alpha}{\beta^2}$, inserting them into a normal distribution and then drawing a random sample from the normal distribution.

Lemma 2.3. *Let* $(Z \mid G = g) \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{g}\right)$*, and* $G \sim \mathcal{G}\left(\frac{\nu}{2}, \frac{\nu}{2}\right)$*, with densities* $f_{Z|G=g}(\cdot)$ *and* $f_G(\cdot)$*, respectively. Then*

$$f_X(x \mid \mu, \sigma^2, \nu) = \int_0^\infty f_{Z|G=g}\left(x \,\middle|\, \mu, \frac{\sigma^2}{g}\right) \, f_G\left(g \,\middle|\, \frac{\nu}{2}, \frac{\nu}{2}\right) \, dg$$

is the pdf of a random variable $X \sim \mathcal{T}(\mu, \sigma^2, \nu)$ *(Andrews and Mallows 1974).*

The following lemma shows that for $\mu = 0$, multiplying the scale σ^2 of a t cdf by a constant $s^2 > 0$ is the same as dividing the t distributed random variable T by s^2. This result will be useful in further derivations.

Lemma 2.4. *Let* $F_T(t \mid 0, \sigma^2 s^2, \nu)$ *denote the distribution function of* $\mathcal{T}(0, \sigma^2 s^2, \nu)$ *with* $s^2 > 0$*. Then*

$$F_T(t \mid 0, \sigma^2 s^2, \nu) = F_T\left(\frac{t}{s} \,\middle|\, 0, \sigma^2, \nu\right).$$

The relationship shown in the lemma above is quite obvious from Definition 2.2 with corresponding considerations.

Often it will be necessary to consider truncated t distributions where the truncation is to values greater than or equal to some value a. The definition below provides the densities for these distributions.

Definition 2.5 (Truncated t distribution)**.** *Let* $T \sim \mathcal{T}(\mu, \sigma^2, \nu)$ *and* $X = T1_{\{T \geq a\}}$*, where* $1_{\{T \geq a\}}$ *is one if* $T \geq a$*, zero otherwise. Then* X *is said to be truncated t distributed, denoted by* $X \sim \mathcal{TT}(a, \mu, \sigma^2, \nu)$*, with density*

$$f(x \mid a, \mu, \sigma^2, \nu) = \frac{f_T(x \mid \mu, \sigma^2, \nu)}{1 - F_T(a \mid \mu, \sigma^2, \nu)} \, 1_{\{x \geq a\}}, \; x \in \mathbb{R},$$

where $f_T(\cdot)$ *is the density and* $F_T(\cdot)$ *is the cumulative distribution function of the t distribution.*

The analytical representation of the first two moments of the truncated t distribution will be helpful for the implementation of fitting algorithms in the next chapter.

Lemma 2.6. *Let* $X = 1_{\{T \geq a\}}T$ *be a truncated t distributed random variable above* a *with density* $f(x \mid a, \mu, \sigma^2, \nu)$*. Then the first two moments of* X *can be written as*

$$a)\; \mathrm{E}(X) = \mu + \frac{\nu\sigma^2 f_T\left(a \,\middle|\, \mu, \sigma^2\frac{\nu}{\nu-2}, \nu - 2\right)}{(\nu-2)(1 - F_T(a \mid \mu, \sigma^2, \nu))}, \text{ and}$$

$$b)\; \mathrm{E}(X^2) = \nu\sigma^2 \left(\frac{(\nu-1)\left(1 - F_T\left(a \,\middle|\, \mu, \sigma^2\frac{\nu}{\nu-2}, \nu - 2\right)\right)}{(\nu-2)(1 - F_T(a \mid \mu, \sigma^2, \nu))} - 1 \right) + \mu(2\mathrm{E}(X) - \mu),$$

where $f_T(\cdot)$ and $F_T(\cdot)$ denote the pdf and cdf of the student t distribution, respectively.

Proof. a) $\mathrm{E}(X)$, the expectation of the first moment is defined as

$$\int_{-\infty}^{\infty} x f_X(x \mid a, \mu, \sigma^2, \nu) dx = \int_a^{\infty} \frac{x f_T(x \mid \mu, \sigma^2, \nu)}{1 - F_T(a \mid \mu, \sigma^2, \nu)} dx$$
$$= \frac{1}{1 - F_T(a \mid \mu, \sigma^2, \nu)} \int_a^{\infty} x \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sigma\sqrt{\nu\pi}} \left(1 + \frac{(x-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}} dx.$$

To solve this integral, we insert a $(-\mu + \mu)$ term in the integral, yielding

$$= \frac{1}{1 - F_T(a \mid \mu, \sigma^2, \nu)} \int_a^{\infty} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sigma\sqrt{\nu\pi}} (x - \mu + \mu) \left(1 + \frac{(x-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}} dx.$$

Using the linearity of the integral and splitting it up yields

$$= \frac{1}{1 - F_T(a \mid \mu, \sigma^2, \nu)} \left[\mu \int_a^{\infty} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sigma\sqrt{\nu\pi}} \left(1 + \frac{(x-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}} dx\right.$$
$$\left. + \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sigma\sqrt{\nu\pi}} \int_a^{\infty} (x-\mu) \left(1 + \frac{(x-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}} dx\right]$$
$$= \frac{\mu(1 - F_T(a \mid \mu, \sigma^2, \nu)) + \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sigma\sqrt{\nu\pi}} \int_a^{\infty} (x-\mu) \left(1 + \frac{(x-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}} dx}{1 - F_T(a \mid \mu, \sigma^2, \nu)}$$
$$= \mu + \frac{1}{1 - F_T(a \mid \mu, \sigma^2, \nu)} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sigma\sqrt{\nu\pi}} \int_a^{\infty} (x-\mu) \left(1 + \frac{(x-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}} dx.$$

Using that the term in the integral can be re-written as a derivative with respect to x

$$(x-\mu) \left(1 + \frac{(x-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}} = \frac{d}{dx}\left(-\frac{\nu\sigma^2}{\nu - 1}\left(1 + \frac{(x-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu-1}{2}}\right),$$

it follows by replacing the term in the integral that $\mathrm{E}(X)$ equals

$$\mu - \frac{1}{1 - F_T(a \mid \mu, \sigma^2, \nu)} \int_a^{\infty} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sigma\sqrt{\nu\pi}} \frac{d}{dx}\left(\frac{\nu\sigma^2}{\nu - 1}\left(1 + \frac{(x-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu-1}{2}}\right) dx.$$

Now solving the integral, it follows that the above expression equals

$$\mu - \frac{\nu\sigma^2}{(\nu-1)(1 - F_T(a \mid \mu, \sigma^2, \nu))} \int_a^{\infty} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sigma\sqrt{\nu\pi}} \frac{d}{dx}\left(1 + \frac{(x-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu-1}{2}} dx$$
$$= \mu - \frac{\nu\sigma^2}{(\nu-1)(1 - F_T(a \mid \mu, \sigma^2, \nu))} \left[\frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sigma\sqrt{\nu\pi}} \left(1 + \frac{(x-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu-1}{2}}\right]_a^{\infty}.$$

Simplifying the term yields

$$= \mu + \frac{\frac{\nu\sigma^2\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sigma\sqrt{\nu\pi}}\left(1+\frac{(a-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu-1}{2}}}{(\nu-1)(1-F_T(a\mid\mu,\sigma^2,\nu))} = \mu + \frac{\frac{\nu\sigma^2\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sigma\sqrt{\nu\pi}}\left(1+\frac{(a-\mu)^2}{(\nu-2)\sigma^2\frac{\nu}{\nu-2}}\right)^{-\frac{\nu-1}{2}}}{(\nu-1)(1-F_T(a\mid\mu,\sigma^2,\nu))}.$$

The numerator can be rewritten in terms of the t density $f_T\left(\cdot\;\middle|\;\mu,\sigma^2\frac{\nu}{\nu-2},\nu-2\right)$, yielding

$$\mu + \frac{\frac{\nu\sigma^2\Gamma(\frac{\nu+1}{2})\Gamma(\frac{\nu-2}{2})}{\Gamma(\frac{\nu}{2})\Gamma(\frac{\nu-1}{2})}f_T\left(a\;\middle|\;\mu,\sigma^2\frac{\nu}{\nu-2},\nu-2\right)}{(\nu-1)(1-F_T(a\mid\mu,\sigma^2,\nu))}.$$

Since for the gamma function it holds that $\Gamma(x+1)=x\Gamma(x)$ the fraction $\frac{\Gamma(\frac{\nu+1}{2})\Gamma(\frac{\nu-2}{2})}{\Gamma(\frac{\nu}{2})\Gamma(\frac{\nu-1}{2})}$ can be simplified to $\frac{\nu-1}{\nu-2}$ and it follows that the expression equals

$$\mu + \frac{\frac{\nu\sigma^2(\nu-1)}{\nu-2}f_T\left(a\;\middle|\;\mu,\sigma^2\frac{\nu}{\nu-2},\nu-2\right)}{(\nu-1)(1-F_T(a\mid\mu,\sigma^2,\nu))} = \mu + \frac{\nu\sigma^2 f_T\left(a\;\middle|\;\mu,\sigma^2\frac{\nu}{\nu-2},\nu-2\right)}{(\nu-2)(1-F_T(a\mid\mu,\sigma^2,\nu))}.$$

b) Similar to above

$$E(X^2) = \frac{1}{1-F_T(a\mid\mu,\sigma^2,\nu)}\int_a^\infty x^2\frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sigma\sqrt{\nu\pi}}\left(1+\frac{(x-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}}dx$$

and can be manipulated to

$$\begin{aligned}
&\frac{1}{1-F_T(a\mid\mu,\sigma^2,\nu)}\int_a^\infty \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sigma\sqrt{\nu\pi}}(x-\mu+\mu)^2\left(1+\frac{(x-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}}dx\\
&=\frac{1}{1-F_T(a\mid\mu,\sigma^2,\nu)}\int_a^\infty \frac{\Gamma(\frac{\nu+1}{2})((x-\mu)^2+2\mu(x-\mu)+\mu^2)}{\Gamma(\frac{\nu}{2})\sigma\sqrt{\nu\pi}}\left(1+\frac{(x-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}}dx\\
&=\frac{1}{1-F_T(a\mid\mu,\sigma^2,\nu)}\left[\int_a^\infty \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sigma\sqrt{\nu\pi}}(x-\mu)^2\left(1+\frac{(x-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}}dx\right.\\
&+2\mu\int_a^\infty \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sigma\sqrt{\nu\pi}}(x-\mu)\left(1+\frac{(x-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}}dx\\
&\left.+\mu^2\int_a^\infty \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sigma\sqrt{\nu\pi}}\left(1+\frac{(x-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}}dx\right].
\end{aligned}$$

Using results from the proof of a), it follows that the above equals

$$\begin{aligned}
&\frac{1}{1-F_T(a\mid\mu,\sigma^2,\nu)}\left[\int_a^\infty \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sigma\sqrt{\nu\pi}}(x-\mu)^2\left(1+\frac{(x-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}}dx\right]\\
&+2\mu(\mathrm{E}(X)-\mu)+\mu^2.
\end{aligned}$$

Further simplification results in

$$\frac{1}{1-F_T(a \mid \mu,\sigma^2,\nu)}\left[\frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sigma\sqrt{\nu\pi}}\nu\sigma^2\int_a^\infty\left(\frac{(x-\mu)^2}{\nu\sigma^2}+1-1\right)\left(1+\frac{(x-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}}dx\right]$$
$$+2\mu\mathrm{E}(X)-\mu^2$$

and can be manipulated to

$$\frac{1}{1-F_T(a \mid \mu,\sigma^2,\nu)}\frac{\nu\sigma^2\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sigma\sqrt{\nu\pi}}\left[\int_a^\infty\left(1+\frac{(x-\mu)^2}{\nu\sigma^2}\right)\left(1+\frac{(x-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}}dx\right.$$
$$\left.-\int_a^\infty\left(1+\frac{(x-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}}dx\right]+2\mu\mathrm{E}(X)-\mu^2.$$

The second integral covers the t density, reducing the expression to

$$\frac{1}{1-F_T(a \mid \mu,\sigma^2,\nu)}\left[\int_a^\infty\frac{\nu\sigma^2\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sigma\sqrt{\nu\pi}}\left(1+\frac{(x-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu-1}{2}}dx\right]$$
$$-\nu\sigma^2+2\mu\mathrm{E}(X)-\mu^2.$$

If we replace ν in the denominator of the parentheses with $(\nu-2)\frac{\nu}{\nu-2}$ and use again some results from the proof of a), we can re-write the integral as

$$\frac{1}{1-F_T(a \mid \mu,\sigma^2,\nu)}\left[\int_a^\infty\frac{\nu\sigma^2(\nu-1)}{\nu-2}f_T\left(x\,\middle|\,\mu,\sigma^2\frac{\nu}{\nu-2},\nu-2\right)dx\right]$$
$$=\frac{\nu\sigma^2(\nu-1)\left(1-F_T\left(a\,\middle|\,\mu,\sigma^2\frac{\nu}{\nu-2},\nu-2\right)\right)}{(\nu-2)(1-F_T(a \mid \mu,\sigma^2,\nu))}.$$

Adding the remaining terms $(-\nu\sigma^2+2\mu E(X)-\mu^2)$ to this integral yields the desired result. □

In the next step, we introduce the skew-normal distribution, which will be used for the stochastic representation of the skew-t distribution. The density of the skew-normal distribution can be represented as a product of the normal cdf and the normal pdf as stated in the definition below.

Definition 2.7 (Univariate skew-normal distribution)**.** *Let $\mu,\lambda\in\mathbb{R}$ and $\sigma^2\in(0,\infty)$. A random variable X is said to follow a skew-normal distribution, denoted by $X\sim\mathcal{SN}(\mu,\sigma^2,\lambda)$ if its pdf is*

$$f(x \mid \mu,\sigma^2,\lambda)=2f_Z(x \mid \mu,\sigma^2+\lambda^2)F_Z\left(\frac{\lambda(x-\mu)}{\sigma^2+\lambda^2}\,\middle|\,0,\frac{\sigma^2}{\sigma^2+\lambda^2}\right),\ x\in\mathbb{R},$$

where $f_Z(\cdot)$ is the pdf and $F_Z(\cdot)$ is the cdf of the normal distribution (Sahu et al. 2003).

Note, when the skewness parameter $\lambda = 0$ the density collapses to a symmetric normal density:

$$f(x \mid \mu, \sigma^2, 0) = 2 f_Z(x \mid \mu, \sigma^2) \underbrace{F_Z(0 \mid 0, 1)}_{=0.5} = f_Z(x \mid \mu, \sigma^2).$$

The cdf of the skew-normal cannot be written in closed form but can be calculated with the help of Owen's T function (Owen 1980), for which fast algorithms exist. To draw random samples from the skew-normal distribution using Markov Chain Monte Carlo (MCMC) methods, we need the hierarchical representation of the skew-normal random variable. Thus, the next proposition shows the stochastic representation of a skew-normal random variable expressed as a half-normal (an at zero-truncated standard normal distribution) and a normal random variable; see Appendix A.6 for details and the definition of the half-normal distribution.

Proposition 2.8. *Let $X \sim \mathcal{SN}(\mu, \sigma^2, \lambda)$. Then $X = \lambda Z + Y$, where $Z \sim \mathcal{HN}(0, 1)$ is independent from $Y \sim \mathcal{N}(\mu, \sigma^2)$ (Arellano-Valle et al. 2007).*

The following definition extends the skew-normal to a skew-t distributed random variable, thus resulting in a random variable that we want to use as a component for the mixture in Definition 2.1.

Definition 2.9 (Hierachical representation of univariate skew-t distributions)**.** *A random variable X is said to follow a skew-t distribution if it can be represented as*

$$X = \mu + \frac{Y}{\sqrt{G}}, \; Y \sim \mathcal{SN}(0, \sigma^2, \lambda), \; G \sim \mathcal{G}\left(\frac{\nu}{2}, \frac{\nu}{2}\right),$$

with Y and G being independent. It is denoted by $X \sim \mathcal{ST}(\mu, \sigma^2, \lambda, \nu)$, where μ, $\lambda \in \mathbb{R}$, $\sigma^2 \in (0, \infty)$ and $\nu \in (0, \infty)$ (Sahu et al. 2003).

Now after defining the skew-t distribution in Definition 2.9, we need to derive the pdf for further calculations. The derivation exploits the hierarchical representation of a skew-t distributed random variable.

Theorem 2.10 (Density of univariate skew-t distributions)**.** *Let $X \sim \mathcal{ST}(\mu, \sigma^2, \lambda, \nu)$ with $\mu, \lambda \in \mathbb{R}$, $\sigma^2, \nu \in (0, \infty)$. Then*

$$f(x \mid \mu, \sigma^2, \lambda, \nu) = 2 f_T(x \mid \mu, \sigma^2 + \lambda^2, \nu) F_T\left(\frac{\lambda(x-\mu)}{\sigma^2 + \lambda^2} \sqrt{\frac{\nu + 1}{\nu + \frac{(x-\mu)^2}{\sigma^2+\lambda^2}}} \,\middle|\, 0, \frac{\sigma^2}{\sigma^2 + \lambda^2}, \nu + 1\right),$$

where $f_T(\cdot)$ is the density and $F_T(\cdot)$ is the cumulative distribution function of the t distribution.

Note that for $\lambda = 0$ the distribution of X reduces to a t distribution, and as the degrees of freedom ν goes to infinity, the distribution of X converges to a skew-normal distribution. The following lemma by Sahu et al. (2003) introduces the hierarchical representation for the skew-t distribution, which we later utilize for fitting.

Lemma 2.11. *Skew-t distributed random variables have the following hierarchical representation (Sahu et al. 2003, p. 201):*

$$\begin{aligned} X \mid (Z = z, G = g) &\sim \mathcal{N}\left(\mu + \lambda z, \frac{\sigma^2}{g}\right) \\ Z \mid (G = g) &\sim \mathcal{HN}\left(0, \frac{1}{g}\right) \\ G &\sim \mathcal{G}\left(\frac{\nu}{2}, \frac{\nu}{2}\right). \end{aligned}$$

Proof. Definition 2.9 states that $X \sim \mathcal{ST}(\mu, \sigma^2, \lambda, \nu)$ can be written as $X = \mu + \frac{Y}{\sqrt{G}}$, where $Y \sim \mathcal{SN}(0, \sigma^2, \lambda)$ is independent of $G \sim G\left(\frac{\nu}{2}, \frac{\nu}{2}\right)$. Proposition 2.8 expresses Y as $Y = \lambda Z_0 + Z_1$, where $Z_0 \sim \mathcal{HN}(0,1)$ is independent of $Z_1 = \mathcal{N}(0, \sigma^2)$. Therefore, $X = \mu + \lambda \frac{1}{\sqrt{G}} Z_0 + \frac{1}{\sqrt{G}} Z_1 = \mu + \lambda Z + W$, where $Z = \frac{1}{\sqrt{G}} Z_0 \sim \mathcal{HN}\left(0, \frac{1}{G}\right)$ and $W \sim \mathcal{N}\left(0, \frac{\sigma^2}{G}\right)$. This implies $X \mid (Z = z, G = g) \sim \mathcal{N}(\mu + \lambda z, \frac{\sigma^2}{g})$, $Z \mid (G = g) \sim \mathcal{HN}\left(0, \frac{1}{g}\right)$. □

The hierarchical representation indicates how to draw a sample of skew-t distributed random variables. First, we draw a sample of gamma distributed random variables, insert them and draw from a half-normal distribution, and then insert both samples in a normal distribution and draw from that distribution. Utilizing this representation, we can derive the expectation and variance of the skew-t distribution with the law of iterated expectations. The expectation and the variance will be used to get the initial values for the estimation. The idea behind is that we can calculate sample expectation, variance, and skewness and achieve some simple initial estimates for the parameters.

Proposition 2.12. *Let $X \sim \mathcal{ST}(\mu, \sigma^2, \lambda, \nu)$.*

a) The expectation of X is $\mathrm{E}(X) = \mu + \lambda \sqrt{\frac{\nu}{\pi}} \frac{\Gamma\left(\frac{\nu-1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}$.

b) The variance of X is $\mathrm{Var}(X) = \frac{\nu}{\nu-2}\left(\sigma^2 + \lambda^2\right) - \frac{\nu}{\pi}\left(\frac{\Gamma\left(\frac{\nu-1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\right)^2 \lambda^2$.

Proof. a) We apply the law of iterated expectation to the representation in Lemma 2.11:

$$\mathrm{E}(Z) = \mathrm{E}(\mathrm{E}(Z \mid G)) = \mathrm{E}\left(\frac{1}{\sqrt{G}}\sqrt{\frac{2}{\pi}}\right),$$

where we used that the mean of the half-normal (denoted with $\mathcal{HN}(0, \sigma^2)$) variable is $\frac{1}{\sqrt{G}}\sqrt{\frac{2}{\pi}}$ (given under Def. A.6). Manipulation yields

$$\sqrt{\frac{2}{\pi}} \mathrm{E}\left(\frac{1}{\sqrt{G}}\right) = \sqrt{\frac{2}{\pi}} \int_0^\infty \frac{1}{\sqrt{g}} \frac{\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)} g^{\frac{\nu}{2}-1} e^{-\frac{\nu}{2}g} dg$$

and simplified to

$$\begin{aligned}
&\sqrt{\frac{\nu}{\pi}}\frac{\Gamma\left(\frac{\nu-1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\int_0^\infty \frac{\left(\frac{\nu}{2}\right)^{\frac{\nu-1}{2}}}{\Gamma\left(\frac{\nu-1}{2}\right)}g^{\frac{\nu-1}{2}-1}e^{-\frac{\nu}{2}g}dg\\
&=\sqrt{\frac{\nu}{\pi}}\frac{\Gamma\left(\frac{\nu-1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}.
\end{aligned}$$

Note that $\mathrm{E}(X \mid Z = z, G = g) = \mu + \lambda z$ does not depend on g. Therefore, $\mathrm{E}(X \mid Z = z, G = g) = \mathrm{E}(X \mid Z = z)$ and

$$\mathrm{E}(X) = \mathrm{E}(\mathrm{E}(X \mid Z)) = \mathrm{E}(\mu + \lambda Z) = \mu + \lambda \mathrm{E}(Z) = \mu + \lambda\sqrt{\frac{\nu}{\pi}}\frac{\Gamma\left(\frac{\nu-1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}.$$

b) Using again the hierarchical representation it follows from the law of total variance that

$$\begin{aligned}
\mathrm{Var}(X) &= \mathrm{E}(\mathrm{Var}(X \mid Z, G)) + \mathrm{Var}(\mathrm{E}(X \mid Z, G))\\
&= \mathrm{E}\left(\frac{\sigma^2}{G}\right) + \mathrm{Var}(\mu + \lambda Z) = \sigma^2\mathrm{E}\left(\frac{1}{G}\right) + \mathrm{Var}(\lambda Z)\\
&= \sigma^2\mathrm{E}\left(\frac{1}{G}\right) + \lambda^2\mathrm{E}(\mathrm{Var}(Z \mid G)) + \lambda^2\mathrm{Var}(\mathrm{E}(Z \mid G))\\
&= \sigma^2\mathrm{E}\left(\frac{1}{G}\right) + \lambda^2\left(1 - \frac{2}{\pi}\right)\mathrm{E}\left(\frac{1}{G}\right) + \lambda^2\mathrm{Var}\left(\sqrt{\frac{1}{G}}\sqrt{\frac{2}{\pi}}\right).
\end{aligned}$$

Using that the variance can be expressed as the second moment and the squared expectation yields

$$\mathrm{E}\left(\frac{1}{G}\right)\left(\sigma^2 + \lambda^2\left(1 - \frac{2}{\pi}\right)\right) + \lambda^2\frac{2}{\pi}\mathrm{E}\left(\frac{1}{G}\right) - \lambda^2\mathrm{E}\left(\frac{1}{\sqrt{G}}\sqrt{\frac{2}{\pi}}\right)^2.$$

For the second expectation, we can use the proof of a), which yields

$$\mathrm{E}\left(\frac{1}{G}\right)\left(\sigma^2 + \lambda^2\left(1 - \frac{2}{\pi}\right)\right) + \lambda^2\frac{2}{\pi}\mathrm{E}\left(\frac{1}{G}\right) - \lambda^2\left(\sqrt{\frac{\nu}{\pi}}\frac{\Gamma\left(\frac{\nu-1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\right)^2.$$

For the variance of the skew-t distribution, we calculate the expectation

$$\begin{aligned}
\mathrm{E}\left(\frac{1}{G}\right) &= \int_0^\infty \frac{1}{g}\frac{\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)}g^{\frac{\nu}{2}-1}e^{\frac{\nu}{2}g}dg\\
&= \frac{\frac{\nu}{2}\Gamma\left(\frac{\nu}{2}-1\right)}{\Gamma\left(\frac{\nu}{2}\right)}\int_0^\infty \frac{\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}-1}}{\Gamma\left(\frac{\nu}{2}-1\right)}g^{\frac{\nu}{2}-2}e^{\frac{\nu}{2}g}dg.
\end{aligned}$$

The integral equals one since it is the integral of the gamma density over its support. The ratio of gamma functions can be simplified using that $\Gamma(x + 1) = x\Gamma(x)$ to obtain $\frac{\nu}{\nu-2}$. Inserting the calculated expectation, $\mathrm{E}\left(\frac{1}{G}\right) = \frac{\nu}{\nu-2}$, the variance becomes

$$\begin{aligned} &\frac{\nu}{\nu-2}\left(\sigma^2+\lambda^2\left(1-\frac{2}{\pi}\right)\right)+\lambda^2\frac{2}{\pi}\frac{\nu}{\nu-2}-\lambda^2\left(\sqrt{\frac{\nu}{\pi}}\frac{\Gamma\left(\frac{\nu-1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\right)^2 \\ &=\frac{\nu}{\nu-2}\left(\sigma^2+\lambda^2\left(1-\frac{2}{\pi}\right)\right)+\frac{2}{\pi}\left(\frac{\nu}{\nu-2}-\left(\frac{\Gamma\left(\frac{\nu-1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\right)^2\frac{\nu}{2}\right)\lambda^2 \\ &=\frac{\nu}{\nu-2}\left(\sigma^2+\lambda^2\right)-\frac{\nu}{\pi}\left(\frac{\Gamma\left(\frac{\nu-1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\right)^2\lambda^2. \end{aligned}$$

□

Figure 3 shows how the skew-t density is formed from Theorem 2.10. The density function of the skew-t distribution is not symmetric, and hence the mode, mean, and median are different. This property should be kept in mind for regression with skew-t distributed errors, as will be done later. In a regression setting, it would mean that the error of the distribution does not necessarily have mean zero nor a symmetric distribution, which would affect the interpretation of the regression coefficients.

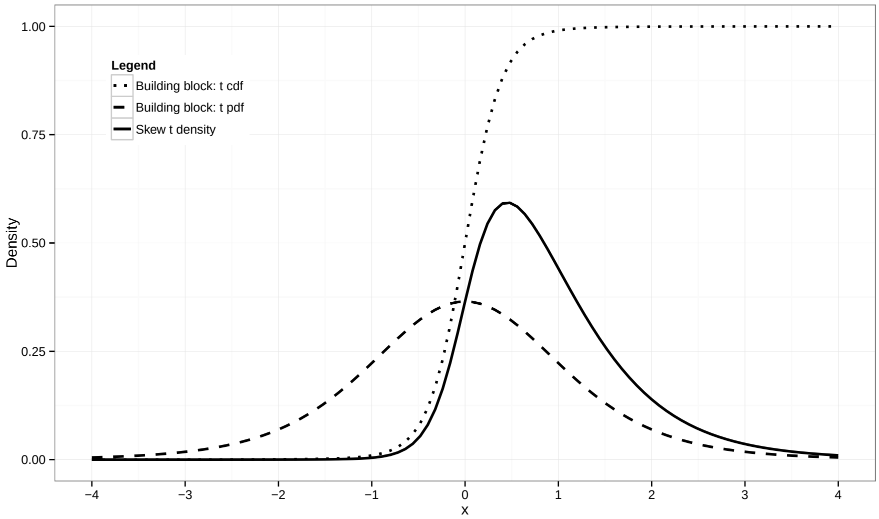


Figure 3 The skew-t density ($\mu = 0$, $\sigma^2 = 0.1$, $\lambda = 1$ and $\nu = 6$) along with its building blocks, the t density ($\mu = 0$, $\sigma^2 = 1.1$, $\nu = 6$) and t distribution function ($\mu = 0$, $\sigma^2 = \frac{0.1}{1.1}$, $\nu = 7$).

As the last lemma for the univariate case, the hierarchical representation of a skew-t mixture with c components is shown. The representation of the skew-t distribution is extended conditioning on a variable M_{ij}, which equals one if X_i belongs to the group j.

Lemma 2.13. *The hierarchical representation of a skew-t mixture with $i = 1, \ldots, n$ observation and $j = 1, \ldots, c$ components can be written as*

$$
\begin{aligned}
X_i \mid (Z_i = z_i, G_i = g_i, M_{ij} = 1) &\sim \mathcal{N}\left(\mu_j + \lambda_j z_i, \frac{\sigma_j^2}{g_i}\right) \\
Z_i \mid (G_i = g_i, M_{ij} = 1) &\sim \mathcal{HN}\left(0, \frac{1}{g_i}\right) \\
G_i \mid (M_{ij} = 1) &\sim \mathcal{G}_i\left(\frac{\nu_j}{2}, \frac{\nu_j}{2}\right) \\
M_i &\sim \mathcal{M}(1; \omega_1, \ldots, \omega_c),
\end{aligned}
$$

where $\mathcal{M}$ denotes the multinomial distribution and M_{ij} equals one if X_i belongs to group j.

Example 2.14. As motivation for the mixture models, we look at the blood marker PSA (prostate-specific antigen, measured in ng/ml) from $n = 546$ men participating in the San Antonio Biomarker Of Risk of prostate cancer study (SABOR). PSA is a commonly used marker for predicting whether a patient is likely to have prostate cancer and should undergo biopsy. This kind of biomarker is called prognostic biomarker. Later we will discuss this data set in more detail in the application part of the thesis. Suppose one has a group of PSA measurements from patients about to undergo biopsy. These patients cannot be assumed to be homogeneous because some will have cancer detected on biopsy and some not.

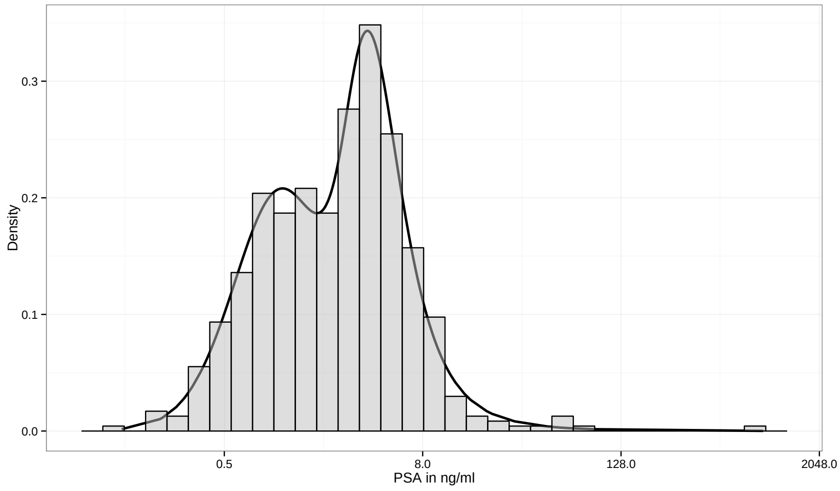


Figure 4 Distribution of the blood marker PSA. Best fitting two component skew-t distribution is overlaid on the histogram of the 546 observations.

Figure 4 shows a histogram of the PSA values with the best fitting two-component skew-t distribution. The PSA values were first $\log_2$-transformed, and then a mixture of skew-t distributions was fit. The histogram looks bimodal, and hence it can be suspected that the different subgroups induce this structure. We can see in this example that the two-component skew-t distribution has the excellent flexibility to model the distribution of PSA. This example illustrates how mixtures of skew-t distributions could be used as a framework for clustering. Patients could be classified as high-risk and recommended to biopsy if they were assigned with high probability to the upper cluster, and low risk otherwise.

2.2 Multivariate mixtures of skew-t distributions

Analogously to the univariate case, multivariate mixtures of skew-t distributions are defined as a sum of weighted density functions. This section presents fewer details and focuses mainly on definition and properties because the ideas and steps are very similar for univariate and multivariate mixtures of skew-t distributions. Note that to improve the readability vectors and matrices will be shown in bold font.

Definition 2.15 (Multivariate mixtures of skew-t distributions)**.** *A random variable $\boldsymbol{X}$ is said to follow a multivariate skew-t mixture distribution with c components if its density is defined as*

$$f(\boldsymbol{x}) = \sum_{i=1}^{c} \omega_i f_{\mathcal{MST}}(\boldsymbol{x} \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \boldsymbol{\lambda}_i, \nu_i),$$

where $f_{\mathcal{MST}}(\cdot)$ are multivariate skew-t densities with location vector $\boldsymbol{\mu}_i$, scale matrix $\boldsymbol{\Sigma}_i$, skewness vector $\boldsymbol{\lambda}_i$, degrees of freedom ν_i, number of components c (finite), and mixing proportions $\omega_i > 0$ with $\sum_{i=1}^{c} \omega_i = 1$.

Properties and construction of univariate mixtures of skew-t distributions have been shown in detail in the previous section. The derivation of the multivariate case is similar, but since matrix multiplication is not commutative, some steps differ. First, we define the p-dimensional normal distribution, which is well known. All marginal distributions of the multivariate normal distribution are also normally distributed.

Definition 2.16 (Multivariate normal distribution)**.** *Let $\boldsymbol{\mu} \in \mathbb{R}^p$ be a mean vector and $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$ a positive definite covariance matrix. A p-dimensional random vector $\boldsymbol{X} = (X_1, \dots, X_p)^\top$ is said to follow the p-variate normal distribution, denoted by $\boldsymbol{X} \sim \mathcal{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, if its joint probability density function is given by*

$$f(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^p |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right), \ \boldsymbol{x} \in \mathbb{R}^p,$$

where $|\boldsymbol{\Sigma}|$ is the determinant of $\boldsymbol{\Sigma}$.

It is important to know that there are several definitions of the multivariate t distribution, which differ. In this thesis, we follow the definition by Kotz and Nadarajah (2004).

Definition 2.17 (Multivariate t distribution)**.** *A p-dimensional random vector $\boldsymbol{X} = (X_1, \dots, X_p)^\top$ is said to follow a p-variate t distribution, denoted by $\boldsymbol{X} \sim \mathcal{MVT}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$, with mean vector $\boldsymbol{\mu} \in \mathbb{R}^p$, positive definite scale matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$, and degrees of freedom $\nu > 0$ if its joint probability density function is given by*

$$f(\boldsymbol{x}) = \frac{\Gamma\left(\frac{\nu+p}{2}\right)}{(\pi\nu)^{\frac{p}{2}} \Gamma\left(\frac{\nu}{2}\right) |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \left[1 + \frac{1}{\nu}(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right]^{-\frac{\nu+p}{2}},$$

where $|\boldsymbol{\Sigma}|$ is the determinant of $\boldsymbol{\Sigma}$ (Kotz and Nadarajah 2004).

For the t distribution, the scale matrix $\boldsymbol{\Sigma}$ is not the covariance matrix of the distribution. The mean vector of the distribution is $\boldsymbol{\mu}$, and the covariance matrix is $\frac{\nu}{\nu-2}\boldsymbol{\Sigma}$. The covariance matrix only exists for $\nu > 2$. As the degrees of freedom ν go to infinity, the multivariate t density approaches the multivariate normal density.

As the next step, we introduce the multivariate skew-normal distribution. A p-dimensional vector $\boldsymbol{\lambda}$ parameterizes the skewness of the distribution. Analogously to the univariate case, the multivariate skew-normal distribution coincides with the multivariate normal distribution for $\boldsymbol{\lambda} = (0, \ldots, 0)^\top$.

Definition 2.18 (Multivariate skew-normal distribution)**.** *Let $\boldsymbol{\mu}, \boldsymbol{\lambda} \in \mathbb{R}^p$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$ be positive definite. A random vector $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ is said to follow a p-dimensional skew-normal distribution, denoted by $\boldsymbol{X} \sim \mathcal{MSN}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$, with location $\boldsymbol{\mu}$, scale $\boldsymbol{\Sigma}$ and shape $\boldsymbol{\lambda}$, if its probability density function is*

$$f(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}) = 2^p f_{\boldsymbol{Z}}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Omega}) F_{\boldsymbol{Z}}\left(\boldsymbol{\Lambda}\boldsymbol{\Omega}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \;\middle|\; \boldsymbol{0}, \boldsymbol{\Delta}\right), \ \boldsymbol{x} \in \mathbb{R}^p,$$

where $\boldsymbol{\Lambda} = diag(\boldsymbol{\lambda}) \in \mathbb{R}^{p \times p}$, the diagonal matrix with diagonal elements $\boldsymbol{\lambda}$, $\boldsymbol{\Omega} = \boldsymbol{\Sigma} + \boldsymbol{\Lambda}^2$, $\boldsymbol{\Delta} = \boldsymbol{I}_p + \boldsymbol{\Lambda}\boldsymbol{\Omega}^{-1}\boldsymbol{\Lambda}$ with $\boldsymbol{I}_p$ the p-dimensional identity matrix, $f_{\boldsymbol{Z}}(\cdot)$ is the density and $F_{\boldsymbol{Z}}(\cdot)$ is the cumulative distribution function of the multivariate normal distribution (see Definition 2.16) (Sahu et al. 2003).

Proposition 2.19. *Let $\boldsymbol{X} \sim \mathcal{MSN}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$. Then*

$$\boldsymbol{X} = \boldsymbol{\Lambda}|\boldsymbol{Z}| + \boldsymbol{Y},$$

where $\boldsymbol{\Lambda}$ and $\boldsymbol{\Sigma}$ as in Definition 2.18, $\boldsymbol{Z} \sim \mathcal{MVN}(\boldsymbol{0}, \boldsymbol{I}_p)$ (see Definition 2.16) is independent from $\boldsymbol{Y} \sim \mathcal{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $|\boldsymbol{Z}|$ denotes the elementwise absolute value of the random vector $\boldsymbol{Z}$ (Arellano-Valle et al. 2007).

Note that $|\boldsymbol{Z}|$ follows a p-dimensional truncated at the hyperplane $\boldsymbol{x} \geq \boldsymbol{0}$ normal distribution, denoted by $\mathcal{MHN}(\boldsymbol{0}, \boldsymbol{I}_p)$. The definition of the multivariate half-normal distribution is given in Appendix A.9. The mean vector of $\boldsymbol{X} \sim \mathcal{MSN}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$ is $\boldsymbol{\mu} + \boldsymbol{\lambda}\sqrt{\frac{2}{\pi}}$. The covariance matrix is $\boldsymbol{\Sigma} + \left(1 - \frac{2}{\pi}\right)\boldsymbol{\Lambda}^2$. We next define a multivariate skew-t distribution in terms of multivariate skew-normal and gamma distributions.

Definition 2.20 (Hierachical representation of multivariate skew-t distributions)**.** *A random vector $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ is said to follow a p-dimensional skew-t distribution if it can be represented by*

$$\boldsymbol{X} = \boldsymbol{\mu} + \frac{\boldsymbol{Y}}{\sqrt{G}}, \ \boldsymbol{Y} \sim \mathcal{MSN}(\boldsymbol{0}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}), \ G \sim \mathcal{G}\left(\frac{\nu}{2}, \frac{\nu}{2}\right),$$

with $\boldsymbol{Y}$ and G independent. It is denoted by $\boldsymbol{X} \sim \mathcal{MST}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu)$, where $\boldsymbol{\mu} \in \mathbb{R}^p$, $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$ positive definite scale matrix, $\boldsymbol{\lambda} \in \mathbb{R}^p$ and $\nu \in (0, \infty)$ (Sahu et al. 2003).

Since the multivariate skew-t distribution is only stochastically defined, we need to derive the pdf. Since we are interested in estimating the parameters of the multivariate skew-t density function, it is, of course, necessary to derive the density function.

Theorem 2.21 (Density of multivariate skew-t distributions)**.** *Let $\boldsymbol{X} \sim \mathcal{MST}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu)$ with $\boldsymbol{\mu}, \boldsymbol{\lambda} \in \mathbb{R}^p$, positive definite $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}, \nu \in (0, \infty)$. Then the density function $f(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu)$ is defined as*

$$2^p f_{\boldsymbol{T}}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Omega}, \nu) F_{\boldsymbol{T}}\left(\boldsymbol{\Lambda}\boldsymbol{\Omega}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\sqrt{\frac{\nu+p}{\nu+(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Omega}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}} \,\middle|\, \boldsymbol{0}, \boldsymbol{\Delta}, \nu+p\right),$$

where $\boldsymbol{\Lambda} = diag(\boldsymbol{\lambda})$, $\boldsymbol{\Omega} = \boldsymbol{\Sigma} + \boldsymbol{\Lambda}^2$, $\boldsymbol{\Delta} = \boldsymbol{I}_p + \boldsymbol{\Lambda}\boldsymbol{\Omega}^{-1}\boldsymbol{\Lambda}$, $f_{\boldsymbol{T}}(\cdot)$ is the density and $F_{\boldsymbol{T}}(\cdot)$ is the cumulative distribution function of the multivariate t distribution (Sahu et al. 2003).

As ν goes to infinity, the multivariate skew-t distribution converges to a multivariate skew-normal distribution. If the skewness vector $\boldsymbol{\lambda}$ is set to a p-dimensional zero vector, the distribution reduces to a multivariate normal distribution.

The multivariate skew-t distribution provides great flexibility as it can be seen in Figure 5, where the contour lines represent the different bivariate skew-t densities. Depending on the parameters for scale, skewness, and degrees of freedom, the distribution can almost adapt to any shape.

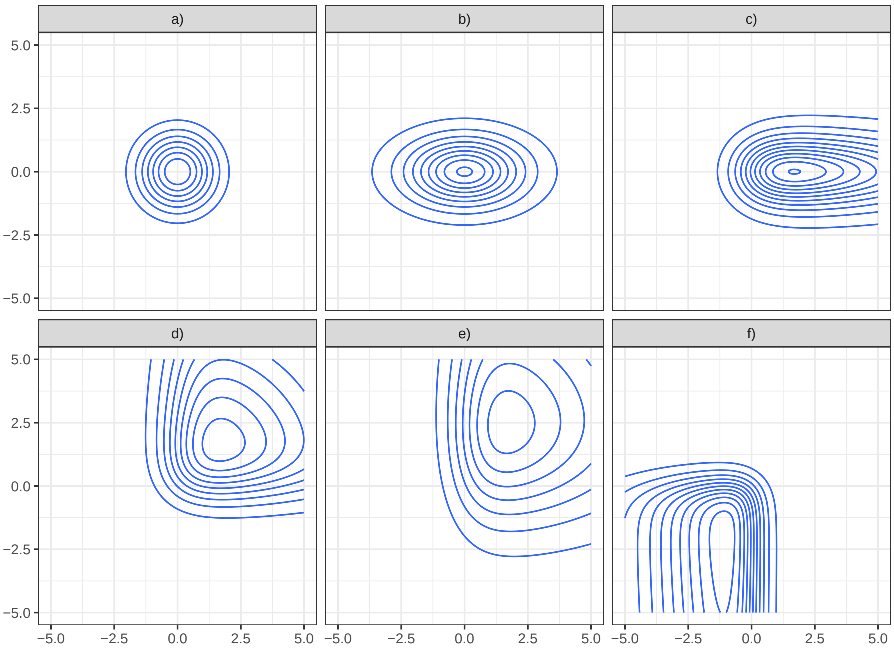


Figure 5 Countour plot of bivariate skew-t densitys with location parameter $\mu_{1,2} = 0$ and other parameters varying. In panel a) $\Sigma_{11,22} = 1$, $\Sigma_{12,21} = 0$, $\lambda_{1,2} = 0$, $\nu = 300$; in b) $\Sigma_{11} = 3$, $\Sigma_{22} = 1$, $\Sigma_{12,21} = 0$, $\lambda_{1,2} = 0$, $\nu = 6$; in c) $\Sigma_{11,22} = 1$, $\Sigma_{12,21} = 0$, $\lambda_1 = 5$, $\lambda_2 = 0$, $\nu = 6$; in d) $\Sigma_{11,22} = 1$,$\Sigma_{12,21} = 0$, $\lambda_{1,2} = 5$, $\nu = 6$; in e) $\Sigma_{11} = 1$, $\Sigma_{22} = 4$, $\Sigma_{12,21} = 0$, $\lambda_{1,2} = 5$, $\nu = 6$; in f) $\Sigma_{11,22} = 0.5$, $\Sigma_{12,21} = 0$, $\lambda_1 = -3$, $\lambda_2 = -15$, $\nu = 6$.

It is important to note that there are many different parameterizations and versions of multivariate skew-t distributions. For example, the R package `sn` (Azzalini 2019) uses a different parameterization for the skew-normal and skew-t distributions. In the next step, we introduce the hierarchical representation of a multivariate skew-t distribution. The hierarchical representation will be used to calculate needed expectations for the Expectation-Maximization algorithm (EM algorithm) and for implementing the MCMC model.

Lemma 2.22. *Multivariate skew-t distributed random variables have the following hierarchical representation (Sahu et al. 2003):*

$$\begin{aligned} \boldsymbol{X} \mid (Z = z, G = g) &\sim \mathcal{MVN}\left(\boldsymbol{\mu} + \boldsymbol{\Lambda z}, \boldsymbol{\Sigma}\frac{1}{g}\right) \\ \boldsymbol{Z} \mid (G = g) &\sim \mathcal{MHN}\left(\boldsymbol{0}, \frac{1}{g}\boldsymbol{I}_p\right) \\ G &\sim \mathcal{G}\left(\frac{\nu}{2}, \frac{\nu}{2}\right). \end{aligned}$$

Proof. Definition 2.20 states $\boldsymbol{X} \sim \mathcal{MST}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu)$ can be written as $\boldsymbol{X} = \boldsymbol{\mu} + \frac{1}{\sqrt{G}}\boldsymbol{Y}$, where $\boldsymbol{Y} \sim \mathcal{MSN}(\boldsymbol{0}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$ is independent of $G \sim G\left(\frac{\nu}{2}, \frac{\nu}{2}\right)$. Proposition 2.19 expresses $\boldsymbol{Y}$ as $\boldsymbol{Y} = \boldsymbol{\Lambda Z}_0 + \boldsymbol{Z}_1$, where $\boldsymbol{Z}_0 \sim \mathcal{MHN}(\boldsymbol{0}, \boldsymbol{I}_p)$ is independent of $\boldsymbol{Z}_1 = \mathcal{MVN}(\boldsymbol{0}, \boldsymbol{\Sigma})$. Therefore, $\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\frac{1}{\sqrt{G}}\boldsymbol{Z}_0 + \frac{1}{\sqrt{G}}\boldsymbol{Z}_1 = \boldsymbol{\mu} + \boldsymbol{\Lambda Z} + \boldsymbol{W}$, where $\boldsymbol{Z} = \frac{1}{\sqrt{G}}\boldsymbol{Z}_0 \sim \mathcal{MHN}\left(\boldsymbol{0}, \frac{1}{G}\boldsymbol{I}_p\right)$ and $\boldsymbol{W} \sim \mathcal{MVN}\left(\boldsymbol{0}, \frac{\boldsymbol{\Sigma}}{G}\right)$. This implies $\boldsymbol{X} \mid (\boldsymbol{Z} = \boldsymbol{z}, G = g) \sim \mathcal{MVN}(\boldsymbol{\mu} + \boldsymbol{\Lambda z}, \frac{\boldsymbol{\Sigma}}{g})$, $\boldsymbol{Z} \mid (G = g) \sim \mathcal{MHN}\left(\boldsymbol{0}, \frac{1}{g}\boldsymbol{I}_p\right)$. □

The expectation and variance of the multivariate skew-t distribution are derived in the proposition below. As already mentioned in the univariate case, expectation and variance will be used for the initial values of the algorithm to estimate skew-t mixtures. Lin (2010) pointed out an error in the original derivation of Sahu et al. (2003). The proof of this statement is explicitly given in Appendix B.7 since only the result is given in Lin (2010) and Sahu et al. (2003). For the goal to estimate the skew-t distribution with an EM algorithm starting values are necessary. Since the empirical mean and variance are easy to estimate on a given data set, we can use the representation of expectation and variance of skew-t distributions to get good starting values. This procedure will be explained in detail in the EM algorithm chapter.

Proposition 2.23. *Let* $\boldsymbol{X} \sim \mathcal{MST}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu)$.

a) *The expectation vector of* $\boldsymbol{X}$ *is* $\mathrm{E}(\boldsymbol{X}) = \boldsymbol{\mu} + \sqrt{\frac{\nu}{\pi}}\frac{\Gamma\left(\frac{\nu-2}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\boldsymbol{\lambda}$.

b) *The covariance matrix of* $\boldsymbol{X}$ *is* $\mathrm{Cov}(\boldsymbol{X}) = \frac{\nu}{\nu-2}\left(\boldsymbol{\Sigma} + \left(1 - \frac{2}{\pi}\right)\boldsymbol{\Lambda}^2\right)$
$+ \frac{2}{\pi}\left(\frac{\nu}{\nu-2} - \left(\frac{\Gamma\left(\frac{\nu-1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\right)^2 \frac{\nu}{2}\right)\boldsymbol{\lambda\lambda}^\top$, *where* $\boldsymbol{\Lambda} = diag(\boldsymbol{\lambda})$.

The mean vector of the distribution is determined by $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$, and the covariance matrix by $\boldsymbol{\Sigma}$, $\boldsymbol{\lambda}$, and ν.

For the hierarchical representation of multivariate skew-t mixtures, we introduce an allocation variable $\boldsymbol{M}$. This allocation variable specifies the mixture component for each observation. Hence, M_{ij} is one if the ith observation belongs to mixture component j, zero otherwise. This property implies that $\sum_{j=1}^{c} M_{ij} = 1$ for all observations i, where c is the number of mixture components. We can see from these properties that $\boldsymbol{M}_i$ follows a multinomial distribution with one trial and cell probabilities $\omega_1, \ldots, \omega_c$. This setting, where $\boldsymbol{M}$ is not directly observed, is called an incomplete data problem. The term incomplete data problem will be formally introduced in the EM algorithm chapter.

Lemma 2.24. *The hierarchical representation of a multivariate skew-t mixture with* $i = 1, \ldots, n$ *observation and* $j = 1, \ldots, c$ *components can be written as*

$$\begin{aligned}
\boldsymbol{X}_i \mid (\boldsymbol{Z}_i = \boldsymbol{z}_i, G_i = g_i, M_{ij} = 1) &\sim \mathcal{MVN}\left(\boldsymbol{\mu}_j + \boldsymbol{\Lambda}_j \boldsymbol{z}_i, \boldsymbol{\Sigma}_j \frac{1}{g_i}\right) \\
\boldsymbol{Z}_i \mid (G_i = g_i, M_{ij} = 1) &\sim \mathcal{MHN}\left(\boldsymbol{0}, \frac{1}{g_i} \boldsymbol{I}_p\right) \\
G_i \mid (M_{ij} = 1) &\sim \mathcal{G}\left(\frac{\nu_j}{2}, \frac{\nu_j}{2}\right) \\
\boldsymbol{M}_i &\sim \mathcal{M}(1; \omega_1, \ldots, \omega_c),
\end{aligned}$$

where $\mathcal{M}(1; \omega_1, \ldots, \omega_c)$ *denotes the multinomial distribution, and* M_{ij} *equals one if* $\boldsymbol{X}_i$ *belongs to mixture component* j*, zero otherwise.*

Example 2.25. We continue the example from the univariate case to motivate the advantages of the multivariate skew-t distribution. In the following example, we consider the bivariate case. As previously, we have PSA measurements from 546 patients about to undergo biopsy. Additionally, on the same patients, a second marker, percent free PSA, is measured before the biopsy. Both markers are known in the literature to be prognostic for the disease, and therefore we want to use the information of both markers in one model.

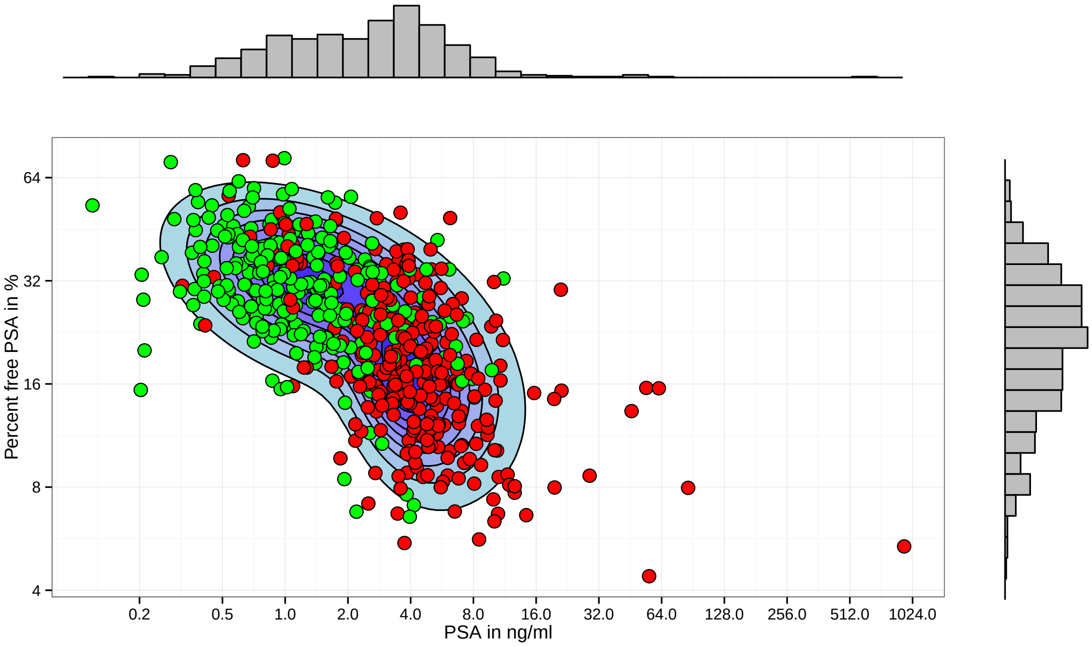


Figure 6 Scatterplot of the blood markers PSA and percent free PSA where red is indicating cancer and green no cancer. Best fitting two component bivariate skew-t distribution is shown with overlaid contour lines. Histograms visualize the marginal distribution.

Figure 6 shows a scatterplot of the observed values for PSA and percent free PSA on $\log_2$ scaled axes. The result of the biopsy is indicated by the color where, red means the biopsy was positive for cancer, green negative. The contour lines visualize the estimated density for the best suitable two-component bivariate skew-t mixture. Histograms show the marginal distributions on the respective axes. From the figure, we see that PSA and percent free PSA separate quite well between the patients with cancer and without cancer. The centers of the mixture components seem to coincide with cancer and no cancer subgroups. The overlaid contour lines of the multivariate skew-t mixture seem to depict the patient characteristics well. Already the marginal histogram of PSA gives the impression of a possibly bimodal distribution, which indicates that a mixture distribution is appropriate to model the data.

2.3 Updating prediction models

Prediction models that have been built on data collected over many years are an essential tool to predict, e.g., the risk with the help of specific patient characteristics. One example of a widely used risk calculator is the Breast Cancer Risk Assessment Tool (BCRAT) (Gail et al. 1989; Gail 2015), which predicts the breast cancer risk of women for specific characteristics. However, methods and standard of care are evolving. Furthermore, research is going forward, and more biomarkers associated with this type of cancer are identified. Large trials, which cannot be redone every couple of years to update this kind of data are the basis of existing prediction models. To resolve this problem, we discuss a method to update an existing risk prediction tool in the following.

As the first step, we show an example of how to update a prediction model using Bayes' rule. Furthermore, we develop a general method to classify multivariate data. For this method, the notation to update a multinomial risk prediction model is shown.

2.3.1 Bayes' theorem

The idea of updating prediction models will be explained using an example of prostate cancer. Updating prediction models in this context means adapting an existing model by incorporating additional information, e.g., new biomarkers that were not available when initially establishing the prediction model. The idea of the used updating mechanism is to apply the Bayes' theorem. Therefore, the recap of the theorem as it can be found in, e.g., Georgii (2013) is outlined below:

Theorem 2.26. *Let $(\Omega, \mathcal{F}, P)$ be a probability space and $\Omega = \cup_{i \in I} B_i$ an at most countable partition of Ω into pairwise disjoint events $B_i \in \mathcal{F}$. Then the following holds. For every $A \in \mathcal{F}$ with $P(A) > 0$ and every $k \in I$, the reverse conditional probability can be expressed as*

$$P(B_k \mid A) = \frac{P(B_k)P(A \mid B_k)}{\sum_{i \in I} P(B_i)P(A \mid B_i)}.$$

In words, the theorem shows that we can express the probability of B_k conditioned on A as the fraction of probabilities, which are not conditioned on A. Taking advantage of this for our case means that A represents a newly measured biomarker, where we do not know the probability of B_k conditioned on A. However, we can estimate the quantities needed in the Bayes' theorem. With this approach, we can calculate the probability of B_k conditioned on A. Below the derivation, for example, with three pairwise disjunct events is shown.

For updating the risk calculator in Section 5.1, the partitions B_i from the Bayes' theorem are the groups high-grade cancer (B_2), low-grade cancer (B_1) and no cancer (B_0). Let X denote the risk factors of the original risk calculator, Y the new marker that should be incorporated. The initial risk factors can include any variables such as continuous, binary, or categorical characteristics.

The PCPT risk calculator (Ankerst et al. 2008) predicts the risk for high-grade, low-grade, and no cancer based on the information X. This risk calculated from the original risk calculator $P(B_i \mid X)$ is denoted as prior risk, where $i \in \{0, 1, 2\}$ with 0 corresponding to no cancer, 1 to low-grade and 2 to high-grade. To update the risk calculator, we are interested in the posterior risk of high-grade, low-grade, and no cancer, which is denoted as $P(B \mid X, Y)$. This posterior risk includes the information of the new marker. By applying Bayes' theorem, the posterior risk for high-grade cancer can be written as

$$\begin{aligned}
&P(B_2 \mid X, Y)\\
&\quad= \frac{P(X, Y \mid B_2)P(B_2)}{P(X, Y \mid B_2)P(B_2) + P(X, Y \mid B_1)P(B_1) + P(X, Y \mid B_0)P(B_0)}\\
&\quad= \frac{\frac{P(X,Y,B_2)}{P(B_2)}P(B_2)}{\frac{P(X,Y,B_2)}{P(B_2)}P(B_2) + \frac{P(X,Y,B_1)}{P(B_1)}P(B_1) + \frac{P(X,Y,B_0)}{P(B_0)}P(B_0)}\\
&\quad= \frac{P(X, Y, B_2)}{P(X, Y, B_2) + P(X, Y, B_1) + P(X, Y, B_0)}\\
&\quad= \frac{P(Y \mid X, B_2)P(X, B_2)}{P(Y \mid X, B_2)P(X, B_2) + P(Y \mid X, B_1)P(X, B_1) + P(Y \mid X, B_0)P(X, B_0)}\\
&\quad= \frac{P(Y \mid X, B_2)P(B_2 \mid X)}{P(Y \mid X, B_2)P(B_2 \mid X) + P(Y \mid X, B_1)P(B_1 \mid X) + P(Y \mid X, B_0)P(B_0 \mid X)}.
\end{aligned}$$

Dividing the numerator and the denominator by $P(Y \mid X, B_1)P(B_1 \mid X) + P(Y \mid X, B_0)P(B_0 \mid X)$ the posterior risk for high-grade cancer in terms of odds can be written as

$$P(B_2 \mid X, Y) = \frac{O(B_2 \mid X, Y)}{O(B_2 \mid X, Y) + 1},$$

where

$$\begin{aligned}
O(B_2 \mid X, Y) &= \frac{P(Y \mid X, B_2)P(B_2 \mid X)}{P(Y \mid X, B_1)P(B_1 \mid X) + P(Y \mid X, B_0)P(B_0 \mid X)}\\
&= \frac{P(Y \mid X, B_2)P(B_2 \mid X)}{P(Y \mid X, B_2^c)P(B_2^c \mid X)}.
\end{aligned}$$

Using the law of total probability shows that

$$P(Y \mid X) = P(Y \mid X, B_2)P(B_2 \mid X) + P(Y \mid X, B_1)P(B_1 \mid X) + P(Y \mid X, B_0)P(B_0 \mid X).$$

The derived results above will be used for updating an existing risk prediction tool incorporating a new marker. In Section 5.1, an example of updating a current tool will be shown and extensively discussed.

2.3.2 Density ratios as a classifier

A proposal for classifying multivariate clinical data is suggested in this chapter. The idea of this method is to fit skew-t mixtures separately to cancer cases and controls. The density ratio is calculated as a function of the fitted densities. Each density models characteristics for cases or controls, respectively. The resulting ratio can assist as a tool to see differences between the two groups. It is important to note that the ratio is uncalibrated and thus, cannot be interpreted as a probability.

In the following, we discuss the form of the density ratio for different situations, such as for the constraint of having equal variance in the two groups. For the constraint of equal variance, the density ratio corresponds to classical linear discriminant analysis. Furthermore, the connection between the density ratio and the well-known logistic regression is established.

Definition 2.27. *Let X and Y be two random variables with their corresponding pdf. The density ratio is then defined as*

$$f(z) = \frac{f_X(z)}{f_Y(z)} \textit{if } z \in [\max(\min(\boldsymbol{x}), \min(\boldsymbol{y})), \min(\max(\boldsymbol{x}), \max(\boldsymbol{y}))],$$

*where **x** and **y** are the realisations of X and Y.*

Note that in the case-control setup, we use the pdf of the control group as the denominator in the ratio for a consistent interpretation of the ratio. The density ratio is set equal to one outside of the range where both densities are defined.

First, we look at a simple situation to motivate the density ratio. This setup is taken from a medical background. X is a prognostic biomarker, and therefore the distribution of the biomarker measurements in cases and controls differs. For simplicity, we assume that the biomarker measurements of cases and controls, in general, follow the same distribution but have different parameters. Formally, the biomarker can be represented by a random variable X that is normally distributed with different means and variances for the two groups. Hence, in the simple case where the biomarker follows a normal distribution, we can write

$$\begin{aligned} X \mid Y = 0 &\sim \mathcal{N}(\mu_0, \sigma_0^2), \\ X \mid Y = 1 &\sim \mathcal{N}(\mu_1, \sigma_1^2). \end{aligned}$$

To calculate log-odds for cases and controls conditional on the distribution of the biomarker results in:

$$\begin{aligned}\ln \frac{P(Y=1 \mid X)}{1-P(Y=1 \mid X)} &= \ln \frac{P(Y=1 \mid X)}{P(Y=0 \mid X)} = \ln \frac{P(Y=1, X)P(X)}{P(Y=0, X)P(X)} \\ &= \ln \frac{P(X \mid Y=1)}{P(X \mid Y=0)} + \ln \frac{P(Y=1)}{P(Y=0)}.\end{aligned}$$

With $\ln \frac{P(Y=1)}{1-P(Y=1)} = \beta_0$ we can write the term as

$$\begin{aligned}&= \ln \frac{P(X \mid Y=1)}{P(X \mid Y=0)} + \beta_0 \\ &= \ln P(X \mid Y=1) - \ln P(X \mid Y=0) + \beta_0.\end{aligned}$$

The equation reminds of the closely related equation for logistic regression. Since we know the distribution of $P(X \mid Y=0)$ and $P(X \mid Y=1)$ we can plug in the definition of the pdf for both probabilities. That yields

$$\begin{aligned}&= \beta_0 + \ln\left(\frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right)\right) - \ln\left(\frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(x-\mu_0)^2}{2\sigma_0^2}\right)\right) \\ &= \beta_0 + \ln\frac{\sigma_1}{\sigma_0} - \frac{(x-\mu_1)^2}{2\sigma_1^2} + \frac{(x-\mu_0)^2}{2\sigma_0^2} \\ &= \beta_0 + \ln\frac{\sigma_1}{\sigma_0} + \frac{\sigma_1^2(x^2-2x\mu_0+\mu_0^2) - \sigma_0^2(x^2-2x\mu_1+\mu_1^2)}{2\sigma_0^2\sigma_1^2} \\ &= \beta_0 + \ln\frac{\sigma_1}{\sigma_0} + \frac{\sigma_1^2\mu_0^2 - \sigma_0^2\mu_1^2}{2\sigma_0^2\sigma_1^2} - \frac{\sigma_1^2\mu_0 - \sigma_0^2\mu_1}{\sigma_0^2\sigma_1^2}x + \frac{\sigma_1^2-\sigma_0^2}{2\sigma_0^2\sigma_1^2}x^2.\end{aligned}$$

We can see that the log odds are a quadratic function of x, the result of the biomarker measurement. Note that for equal variance $\sigma_0^2 = \sigma_1^2$ the expression simplifies to a linear form. We define $\tilde{\beta}_0 = \beta_0 + \ln\frac{\sigma_1}{\sigma_0} + \frac{\sigma_1^2\mu_0^2-\sigma_0^2\mu_1^2}{2\sigma_0^2\sigma_1^2}$, $\tilde{\beta}_1 = -\frac{(\sigma_1^2\mu_0-\sigma_0^2\mu_1)}{\sigma_0^2\sigma_1^2}$ and $\tilde{\beta}_2 = \frac{\sigma_1^2-\sigma_0^2}{2\sigma_0^2\sigma_1^2}$ and can write the above equation as

$$= \tilde{\beta}_0 + \tilde{\beta}_1 x + \tilde{\beta}_2 x^2.$$

This form is similar to logistic regression with covariates x and x^2. Note that for equal variance, the quadratic term vanishes. The equation without the quadratic term for equal variances means that the updated risk curve in this simple case will have a monotonic form. Please not that this only holds for the simple case with normal distribtions.

2.4 Skew-t regression

Besides the theory about skew-t distributions, we want to be able to incorporate covariates in the model resulting in a regression model with the skew-t distributed error term. The usage of a skew-t distributed error term gives more flexibility to fit the regression model to skewed data.

Often a transformation is used for skewed data in regression problems. With the following approach, a transformation is not necessary, and the estimates can be interpreted without the need for back-transformation or interpretation on the transformed scale.

For the skew-t regression model, we assume to have an $n \times p$ covariate matrix $\boldsymbol{X}$ with full column rank and a p-variate vector of regression coefficients $\boldsymbol{\beta}$.

We have n observations y_i, where $y_i \sim \mathcal{ST}(\mu_i, \sigma^2, \lambda, \nu)$ independently. For the regression model, we set $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$ for the location parameter of the skew-t distribution. Hence, the likelihood function of the skew-t regression with parameters $\boldsymbol{\beta}$, σ^2, λ, and ν is given by the product of the skew-t densities:

$$\mathcal{L}(\boldsymbol{\beta}, \sigma^2, \lambda, \nu; y_1, \dots, y_n) = \prod_{i=1}^{n} f(y_i \mid \boldsymbol{x}_i\boldsymbol{\beta}, \sigma^2, \lambda, \nu).$$

The likelihood function will be used to incorporate the skew-t regression into the EM algorithm.

Interpreting this as regression we can write the model as follows

$$\boldsymbol{Y} = \boldsymbol{\beta}^\top \boldsymbol{X} + \boldsymbol{\varepsilon},$$

where $\varepsilon_i \sim \mathcal{ST}(0, \sigma^2, \lambda, \nu)$, $\mathrm{E}(\varepsilon_i) = \lambda\sqrt{\frac{\nu}{\pi}}\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}$. That means that $\mathrm{E}(\boldsymbol{Y} \mid \boldsymbol{X}) = \boldsymbol{\beta}^\top \boldsymbol{X} + \lambda\sqrt{\frac{\nu}{\pi}}\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}$. It is crucial to understand that the conditional expectation is not only dependent on $\boldsymbol{\beta}$ but also on skewness and degrees of freedom of the skew-t distribution. Hence, the interpretation of the model results is not straightforward and must be done with caution.

3 Algorithms

In this chapter, we first give an introduction to EM algorithms and their properties (Dempster et al. 1977; Tanner and Wong 1987). Secondly, we review EM algorithms for fitting multivariate skew-t mixtures that were introduced by Lin (2010) and Lee and McLachlan (2011). We motivate collapsed clusters, introduce the notion, and present how to extend the EM algorithm to accommodate collapsed clusters. We discuss how an MCMC algorithm can be programmed in R using JAGS (Hornik et al. 2003; Plummer 2012) or WinBugs (Lunn et al. 2000) with the help of the hierarchical representation of the mixtures. A brief insight into the advantages and disadvantages of MCMC and EM method is discussed.

3.1 EM algorithm

The EM algorithm is a general iterative optimization method. Although the method was used before for specific problems, it was first popularized by the name EM in the publication by Dempster et al. (1977). Since then, the method has been extended to solve many types of statistical problems. The EM algorithm is often used to find maximum likelihood estimates in incomplete data problems or in situations where it is needed to calculate the maximum likelihood estimates iteratively. The term incomplete data means in this sense that the observed data vector, which is regarded as incomplete, is an observable function of the complete data (McLachlan and Krishnan 2007). Incompleteness can also arise from censored or truncated distributions or latent variables. Latent variables are unobservable variables that are introduced to simplify the analysis (Rizzo 2007).

First, we introduce and review the main ideas and results from EM algorithms. The EM algorithm belongs to the class of data augmentation algorithms. The data are augmented with latent variables, such as missing data or parameters, to simplify the calculations (McLachlan and Krishnan 2007) and hence to avoid computational maximization or simulation. Thus, the EM algorithm computes the maximum likelihood given the complete data (Zhai 2007). The main idea of the EM algorithm is to find the global maximum by alternating between an E (expectation) and M (maximization) step. In the E-step, the expectation of the log-likelihood function given the current parameters, and the observed data is calculated. In the M-step, the conditional expectation is maximized with respect to the target parameters. The estimates are updated, repeating the E-and M-steps iteratively until the algorithm converges according to a chosen criterion. In some cases, computing the conditional expectation in the E-step is complicated.

Following Tanner and Wong (1987) the EM algorithm is provided in terms of the observed data $\boldsymbol{y}$, augmented data $\boldsymbol{x}$, and latent data $\boldsymbol{z}$. Hence, we can write $\boldsymbol{x} = (\boldsymbol{y}, \boldsymbol{z})$. The goal is to maximize the marginal likelihood of $\boldsymbol{y}$ with respect to the complete likelihood $p(\boldsymbol{x} \mid \boldsymbol{\theta})$ for parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

One of the desirable qualities of the EM algorithm is its monotonic convergence to a maximum.

Algorithm 3.1 EM algorithm

(i) Set $\boldsymbol{\theta}^{(0)}$ to be a starting value for the maximum likelihood estimates. Iterate between the 2-steps until a stopping criterion is fulfilled, for example, $\|\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^{(k)}\|$ or $|Q(\boldsymbol{\theta}^{(k+1)} \mid \boldsymbol{\theta}^{(k)}) - Q(\boldsymbol{\theta}^{(k)} \mid \boldsymbol{\theta}^{(k)})|$ is sufficiently small.

(ii) E-step: Compute

$$\begin{aligned} Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)}) &= \int_{\boldsymbol{Z}} \ln(p(\boldsymbol{\theta} \mid \boldsymbol{Z}, \boldsymbol{Y}))p(\boldsymbol{Z} \mid \boldsymbol{\theta}^{(k)}, \boldsymbol{Y})d\boldsymbol{Z} \\ &= \mathrm{E}(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{\theta}^{(k)}). \end{aligned}$$

(iii) M-step: Set $\hat{\boldsymbol{\theta}}^{(k+1)} := arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})$.

Let $\boldsymbol{Y}$ be a random vector corresponding to the observed data, $\boldsymbol{Z}$ the latent data, $\boldsymbol{X} = (\boldsymbol{Y}, \boldsymbol{Z})$ the augmented data, $p(\boldsymbol{x} \mid \boldsymbol{\theta})$ the density of $\boldsymbol{X}$, $\boldsymbol{\theta} \in \mathbb{R}^p$ the parameters of interest, $p(\boldsymbol{\theta})$ a prior for $\boldsymbol{\theta}$, $l(\boldsymbol{\theta} \mid \boldsymbol{Y})$ likelihood, $p(\boldsymbol{\theta} \mid \boldsymbol{X}) \propto l(\boldsymbol{\theta})p(\boldsymbol{\theta})$ posterior, $p(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{Z})$ augmented posterior.

The following theorem shows this property.

Theorem 3.1 (Monotonicity of the EM algorithm)**.** *Every EM algorithm increases the posterior $p(\boldsymbol{\theta} \mid \boldsymbol{Y})$ at each iteration, i.e.,*

$$p(\boldsymbol{\theta}^{(k+1)} \mid \boldsymbol{Y}) \geq p(\boldsymbol{\theta}^{(k)} \mid \boldsymbol{Y}).$$

Theorem 3.1 shows that if $\boldsymbol{\theta}$ converges it converges to a stationary point $p(\boldsymbol{\theta} \mid \boldsymbol{Y})$. However, when there are multiple stationary points, such as local minima, maxima or saddle points, the EM algorithm may not converge to the corresponding global maximum or minimum, respectively. Dempster et al. (1977) showed that the EM algorithm converges at a linear rate, with the rate depending on the proportion of information about $\boldsymbol{\theta}$ in $p(\boldsymbol{\theta} \mid \boldsymbol{Y})$, that is, the amount of observed information. This implies that convergence might be slow, particularly when a large portion of the data is missing. In the following, we define some extensions of the EM algorithm, which will be useful for fitting skew-t distributions. The first is the Expectation Conditional Maximization (ECM) algorithm.

Meng and Rubin (1993) introduced the ECM as an extension of the EM algorithm. The ECM algorithm replaces the M-step by several conditional maximization steps, the so-called CM-steps. The advantage of the CM-steps is that they are more straightforward to calculate. Furthermore, in the M-step, we need to optimize over the full parameter space of $\boldsymbol{\theta}$, in contrast to the CM-step where we need only optimize over consecutive subsets of $\boldsymbol{\theta}$. In general, $S > 1$ steps replace the M-step. Let $s = 1, \ldots, S$ and $\boldsymbol{\theta}^{(k+s/S)}$ denote the value of $\boldsymbol{\theta}$ on the sth CM-step of the iteration k+1. If we partition $\boldsymbol{\theta}$ into S subvectors denoted by $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_S)$ then this corresponds to the CM-steps described above. Hence the sth CM-step is the maximization of the Q-function with respect to the subvector $\boldsymbol{\theta}_s$ conditional on the other $S-1$ subvectors. E.g., in case of the EM algorithm for skew-t mixtures, the maximization is done for each parameter separately.

Formally we are maximizing $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})$, constraint to $\boldsymbol{g}_s(\boldsymbol{\theta}) = \boldsymbol{g}_s(\boldsymbol{\theta}^{(k+(s-1)/S)})$, where

$\{\boldsymbol{g}_s(\boldsymbol{\theta}), s = 1, \ldots, S\}$ is a set of preselected vector functions. Hence, with this we get $Q(\boldsymbol{\theta}^{(k+s/S)} \mid \boldsymbol{\theta}) \geq Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})$ for all $\boldsymbol{\theta}$ satisfying the constraint $\boldsymbol{g}_s(\boldsymbol{\theta}) = \boldsymbol{g}_s(\boldsymbol{\theta}^{(k+(s-1)/S)})$. This means that the ECM algorithm has the desired convergence properties and therefore $L(\boldsymbol{\theta}^{(k+1)}) \geq L(\boldsymbol{\theta}^{(k)})$ holds. The convergence properties and the rate of convergence of the ECM algorithm have been discussed in Meng and Rubin (1993) and Sexton and Swensen (2000), where was shown that $Q(\boldsymbol{\theta}^{(k+1)} \mid \boldsymbol{\theta}^{(k+1)}) \geq Q(\boldsymbol{\theta}^{(k+s/S)} \mid \boldsymbol{\theta}^{(k)}) \geq Q(\boldsymbol{\theta}^{(k)} \mid \boldsymbol{\theta}^{(k)})$. Computation of the E-step may be faster than the CM-step. Hence, Meng and Rubin proposed the multi-cycle ECM algorithm. In the multi-cycle ECM algorithm, the expectation in the E-step is calculated after each CM-step. In general, the convergence of the ECM algorithm may be slower than the EM algorithm. However, it can be faster in total computation time. It might be that more iterations are needed, but in total the computational costs for each iteration are lower.

Another extension to the EM and ECM algorithm is the Expectation Conditional Maximization Either (ECME) algorithm. The algorithm operates like the ECM algorithm, but the "either" means that some or all CM-steps are replaced by steps that conditionally maximize the incomplete data log-likelihood function. Hence, the ECME algorithm either maximizes the conditional expectation of the complete data log-likelihood $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})$ or the actual incomplete data log-likelihood function. The convergence results of the EM and ECM algorithms hold for the ECME algorithm as well. The ECME algorithm is preferable since it is nearly always faster than EM and ECM algorithm regarding the number of iterations (Liu and Rubin 1994). For the implementation of our package, we will use the ECM and the ECME algorithm depending if different degrees of freedom are chosen in the mixture components. If only one parameter is selected for all components, the faster ECME algorithm can be used.

3.1.1 EM algorithm for univariate skew-t distributions

In this section, we define the univariate algorithm in detail, which makes it easier to understand the algorithm for the mixture of skew-t distributions as well as for the multivariate skew-t distribution. For the algorithm, we first need to calculate $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})$, which represents the expectation conditional on the data and the current parameters $\boldsymbol{\theta}$ of the likelihood with respect to the joint density. Since we use the ECM algorithm, we maximize the Q-function consecutively for each parameter. Let $\boldsymbol{x} = (x_1, \ldots, x_n)$, $\boldsymbol{z} = (z_1, \ldots, z_n)$ and $\boldsymbol{g} = (g_1, \ldots, g_n)$. Hence, the likelihood function of $\boldsymbol{\theta} = (\mu, \sigma^2, \lambda, \nu)$ given $(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{g})$ is defined as

$$\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{g}) = f(x_1, \ldots x_n, z_1, \ldots, z_n, g_1, \ldots, g_n \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i, z_i, g_i \mid \boldsymbol{\theta}),$$

and the log-likelihood function

$$l(\boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{g}) = \sum_{i=1}^{n} \ln f(x_i, z_i, g_i \mid \boldsymbol{\theta}).$$

The joint density function of X, Z and G denoted as $f(x, z, g)$ equals $\frac{\left(g\frac{\nu}{2}\right)^{\frac{\nu}{2}}}{\sigma\pi\Gamma\left(\frac{\nu}{2}\right)} \exp\left(-\frac{g}{2}\left(\nu + \frac{(x-\mu)^2}{\sigma^2+\lambda^2} + \frac{\sigma^2+\lambda^2}{\sigma^2}\left(z - \frac{\lambda(x-\mu)}{\sigma^2+\lambda^2}\right)^2\right)\right) 1_{\{z\geq 0,\, g\geq 0\}}$. Expressions of density functions, which will be needed either for the log-likelihood function or the EM algorithms, are shown in the following lemmata. These expressions are straightforward to derive using properties of distribution functions (Lin 2010; Lee and McLachlan 2011).

The complete data log-likelihood function $l(\boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{g})$ equals

$$\begin{aligned}
&\sum_{i=1}^{n} \ln \left[\frac{(g_i \frac{\nu}{2})^{\frac{\nu}{2}}}{\sigma \pi \Gamma\left(\frac{\nu}{2}\right)} \exp\left(-\frac{g_i}{2} \left(\nu + \frac{(x_i - \mu)^2}{\sigma^2 + \lambda^2} + \frac{\sigma^2 + \lambda^2}{\sigma^2} \left(z_i - \frac{\lambda(x_i - \mu)}{\sigma^2 + \lambda^2} \right)^2 \right) \right) \right] \\
&= \frac{\nu}{2} \sum_{i=1}^{n} \ln g_i - \sum_{i=1}^{n} \left[\frac{g_i}{2} \left(\nu + \frac{(x_i - \mu)^2}{\sigma^2 + \lambda^2} + \frac{\sigma^2 + \lambda^2}{\sigma^2} \left(z_i - \frac{\lambda(x_i - \mu)}{\sigma^2 + \lambda^2} \right)^2 \right) \right] \\
&+ n \left[\frac{\nu}{2} \ln\left(\frac{\nu}{2}\right) - \ln\left(\sigma \pi \Gamma\left(\frac{\nu}{2}\right) \right) \right].
\end{aligned}$$

The EM algorithm does not maximize the likelihood function directly, but rather the Q-function, $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})$, which is the expectation of the log-likelihood function conditional on the data. $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})$ equals the expectation of the complete data log-likelihood evaluated at $\boldsymbol{\theta}^{(k)}$ conditional on the observed data $\boldsymbol{x}$ $\mathrm{E}(l(\boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{g}) \mid \boldsymbol{x}, \boldsymbol{\theta}^{(k)})$ and is needed in the E-step, where $\boldsymbol{\theta} = (\mu, \sigma^2, \lambda, \nu)$ for this the conditional derivatives of $\boldsymbol{x}, \boldsymbol{z}, g$ are used. Calculating this expectation yields

$$\begin{aligned}
Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)}) = &\frac{\nu^{(k)}}{2} \sum_{i=1}^{n} \mathrm{E}(\ln(g_i) \mid x_i, \boldsymbol{\theta}^{(k)}) - \frac{\nu^{(k)}}{2} \sum_{i=1}^{n} \mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)}) \\
&- \frac{\sum_{i=1}^{n} (x_i - \mu^{(k)})^2 \mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)})}{2((\sigma^{(k)})^2 + (\lambda^{(k)})^2)} \\
&- \frac{(\sigma^{(k)})^2 + (\lambda^{(k)})^2}{2(\sigma^{(k)})^2} \sum_{i=1}^{n} \mathrm{E}\left(g_i \left(z_i - \frac{\lambda^{(k)}(x_i - \mu^{(k)})}{(\sigma^{(k)})^2 + (\lambda^{(k)})^2} \right)^2 \middle| \, x_i, \boldsymbol{\theta}^{(k)} \right) \\
&+ n \left[\frac{\nu^{(k)}}{2} \ln\left(\frac{\nu^{(k)}}{2} \right) - \ln\left(\sigma^{(k)} \pi \Gamma\left(\frac{\nu^{(k)}}{2} \right) \right) \right].
\end{aligned}$$

For each E-step k of the EM algorithm $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})$ is analytically calculated, and afterward the M-step is performed. In the algorithm, the E- and CM-steps will be combined. Hence, $arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})$ is calculated by setting first derivatives to zero. Since we want to use the ECM algorithm, the expectations are maximized sequentially conditional on the other parameters in the CM-steps. Therefore, we maximize the conditional expectations explicitly. To have a fast implementation, we express the necessary expectations as closed-form and make use of t densities and distribution functions, which can be evaluated by efficient algorithms. Calculations of the conditional expectations, which are quite detailed, are shown in the appendix.

For the parameter ν, the optimization step cannot be re-written as a closed-form expression; hence, we use a one-dimensional root search algorithm. In the next step, we have calculated all parts needed for fitting a skew-t distribution with an ECM algorithm. In Definition 3.5, the explicit ECM algorithm is shown. Lemma 3.2 provides several results to calculate the expectations in the E-step that will be needed later for implementation of the ECM algorithm. To derive the conditional expectation $\mathrm{E}(\ln(G) \mid X = x)$ we need to calculate the derivative of the t distribution function with respect to the parameter ν. In Lemma 3.3, the result of this derivative is presented. The optimization steps for the ECM algorithms are as follows:

Lemma 3.2. *Let $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})$ be the expectation of the log-likelihood function conditional on the data and the parameter estimates at step k.*

a) $\mu^{(k+1)} = \frac{\sum_{i=1}^{n} x_i \mathrm{E}(g_i|x_i,\boldsymbol{\theta}^{(k)}) - \lambda^{(k)} \sum_{i=1}^{n} \mathrm{E}(g_i z_i|x_i,\boldsymbol{\theta}^{(k)})}{\sum_{i=1}^{n} \mathrm{E}(g_i|x_i,\boldsymbol{\theta}^{(k)})}.$

b) $\lambda^{(k+1)} = \frac{\sum_{i=1}^{n} (x_i - \mu^{(k+1)}) \mathrm{E}(g_i z_i|x_i,\boldsymbol{\theta}^{(k)})}{\sum_{i=1}^{n} \mathrm{E}(g_i z_i^2|x_i,\boldsymbol{\theta}^{(k)})}.$

c) $(\sigma^{(k+1)})^2 = \frac{\sum_{i=1}^{n} (x_i - \mu^{(k+1)})^2 \mathrm{E}(g_i|x_i,\boldsymbol{\theta}^{(k)}) + (\lambda^{(k+1)})^2 \sum_{i=1}^{n} \mathrm{E}(g_i z_i^2|x_i,\boldsymbol{\theta}^{(k)})}{n}$
$- \frac{2\lambda^{(k+1)} \sum_{i=1}^{n} (x_i - \mu^{(k+1)}) \mathrm{E}(g_i z_i|x_i,\boldsymbol{\theta}^{(k)})}{n}.$

d) $\nu^{(k+1)} = \underset{\nu}{solve}\left[\ln\left(\frac{\nu}{2}\right) - \psi\left(\frac{\nu}{2}\right) + 1 + \frac{\sum_{i=1}^{n} \mathrm{E}(\ln(g_i)|x_i,\boldsymbol{\theta}^{(k)}) - \sum_{i=1}^{n} \mathrm{E}(g_i|x_i,\boldsymbol{\theta}^{(k)})}{n} = 0 \right]$*, where $\psi(x)$ denotes the digamma function which is the derivative of the log gamma function $\frac{d}{dx} \ln \Gamma(x)$.*

To explicitly make use of the algorithms we calculate the expectations, which are required in the optimization steps. The following lemma provides results that are needed for the derivation of the expectations.

Lemma 3.3. *Let $T \sim \mathcal{T}\left(\frac{\lambda(x-\mu)}{\sigma^2+\lambda^2} \sqrt{\frac{\nu+1}{\nu + \frac{(x-\mu)^2}{\sigma^2+\lambda^2}}} \,\middle|\, 0, \frac{\sigma^2}{\sigma^2+\lambda^2}, \nu+1 \right)$ then the derivative with respect to the parameter ν of the t distribution function is*

$$\frac{1}{2} \int_{-\infty}^{\frac{\lambda(x-\mu)}{\sigma^2+\lambda^2}} f_T\left(t, 0, \frac{\sigma^2 \left(\nu + \frac{(x-\mu)^2}{\sigma^2+\lambda^2} \right)}{(\sigma^2+\lambda^2)(\nu+1)}, \nu+1 \right) c(t)\, dt,$$

where

$$\begin{aligned} c(t) =& \psi\left(\frac{\nu}{2}+1\right) - \psi\left(\frac{\nu+1}{2}\right) - \left(\nu + \frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)^{-1} \\ &+ \frac{t^2 \left(\frac{\sigma^2+\lambda^2}{\sigma^2}\right)(\nu+2)}{\left(\nu + \frac{(x-\mu)^2}{\lambda^2+\sigma^2}\right)\left(\nu + \frac{(x-\mu)^2}{\sigma^2+\lambda^2} + \frac{t^2(\sigma^2+\lambda^2)}{\sigma^2}\right)} \\ &- \ln\left(1 + \frac{t^2(\sigma^2+\lambda^2)}{\sigma^2\left(\nu + \frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)}\right). \end{aligned}$$

In the following lemma, it is explicitly shown how to calculate the conditional expectations, which are used in each step of the ECM algorithm, based on the actual value of the parameter at the kth iteration and the data.

Lemma 3.4. *Let $\mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)})$, $\mathrm{E}(\ln(g_i) \mid x_i, \boldsymbol{\theta}^{(k)})$, $\mathrm{E}(g_i z_i \mid x_i, \boldsymbol{\theta}^{(k)})$ and $\mathrm{E}(g_i z_i^2 \mid x_i, \boldsymbol{\theta}^{(k)})$ be the expectations, which are needed for the ECM algorithm shown in 3.2. These expectations can be calculated as follows:*

a) $e_{1i}^{(k)} := \mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)}) =$

$$\frac{(\nu^{(k)}+1) f_T\left(\frac{\lambda^{(k)}(x_i-\mu^{(k)})}{(\sigma^{(k)})^2+(\lambda^{(k)})^2}\sqrt{\frac{\nu^{(k)}+3}{\nu^{(k)}+\frac{(x_i-\mu^{(k)})^2}{(\sigma^{(k)})^2+(\lambda^{(k)})^2}}}\,\middle|\, 0, \frac{(\sigma^{(k)})^2}{(\sigma^{(k)})^2+(\lambda^{(k)})^2}, \nu^{(k)}+3\right)}{\left(\nu^{(k)}+\frac{(x_i-\mu^{(k)})^2}{(\sigma^{(k)})^2+(\lambda^{(k)})^2}\right) f_T\left(\frac{\lambda^{(k)}(x_i-\mu^{(k)})}{(\sigma^{(k)})^2+(\lambda^{(k)})^2}\sqrt{\frac{\nu^{(k)}+1}{\nu^{(k)}+\frac{(x_i-\mu^{(k)})^2}{(\sigma^{(k)})^2+(\lambda^{(k)})^2}}}\,\middle|\, 0, \frac{(\sigma^{(k)})^2}{(\sigma^{(k)})^2+(\lambda^{(k)})^2}, \nu^{(k)}+1\right)}.$$

b) $e_{2i}^{(k)} := \mathrm{E}(\ln(g_i) \mid x_i, \boldsymbol{\theta}^{(k)}) =$

$$e_{1i}^{(k)} - \frac{\nu^{(k)}+1}{\left(\nu^{(k)}+\frac{(x_i-\mu^{(k)})^2}{(\sigma^{(k)})^2+(\lambda^{(k)})^2}\right)} - \ln\left(\frac{\nu^{(k)}}{2}+\frac{(x_i-\mu^{(k)})^2}{2((\sigma^{(k)})^2+(\lambda^{(k)}))}\right) + \psi\left(\frac{\nu^{(k)}+1}{2}\right)$$
$$+\int_{-\infty}^{\frac{\lambda^{(k)}(x_i-\mu^{(k)})}{(\sigma^{(k)})^2+(\lambda^{(k)})^2}} \frac{f_T\left(t, 0, \frac{(\sigma^{(k)})^2\left(\nu^{(k)}+\frac{(x_i-\mu^{(k)})^2}{(\sigma^{(k)})^2+(\lambda^{(k)})^2}\right)}{((\sigma^{(k)})^2+(\lambda^{(k)})^2)(\nu^{(k)}+1)}, \nu^{(k)}+1\right)}{F_T\left(\frac{\lambda^{(k)}(x_i-\mu^{(k)})}{(\sigma^{(k)})^2+(\lambda^{(k)})^2}\sqrt{\frac{\nu^{(k)}+1}{\nu^{(k)}+\frac{(x_i-\mu^{(k)})^2}{(\sigma^{(k)})^2+(\lambda^{(k)})^2}}}\,\middle|\, \frac{(\sigma^{(k)})^2}{(\sigma^{(k)})^2+(\lambda^{(k)})^2}, \nu^{(k)}+1\right)} c(t)dt,$$

with $c(t)$ defined as in Lemma 3.3.

c) $e_{3i}^{(k)} := \mathrm{E}(g_i z_i \mid x_i, \boldsymbol{\theta}^{(k)}) = \mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)})\mathrm{E}(z_i \mid x_i, \boldsymbol{\theta}^{(k)}) =$

$$e_{1i}^{(k)}\left(\mu^{(k)} + \frac{\nu^{(k)}(\sigma^{(k)})^2 f_T\left(0 \,\middle|\, \mu^{(k)}, (\sigma^{(k)})^2\frac{\nu^{(k)}}{\nu^{(k)}-2}, \nu^{(k)}-2\right)}{(\nu^{(k)}-2)(1-F_T(0 \mid \mu^{(k)}, (\sigma^{(k)})^2, \nu^{(k)}))}\right).$$

d) $e_{4i}^{(k)} := \mathrm{E}(g_i z_i^2 \mid x_i, \boldsymbol{\theta}^{(k)}) =$

$$e_{1i}^{(k)}\left(\frac{\nu^{(k)}(\sigma^{(k)})^2(\nu^{(k)}-1)\left(1-F_T\left(0 \,\middle|\, \mu^{(k)}, (\sigma^{(k)})^2\frac{\nu^{(k)}}{\nu^{(k)}-2}, \nu^{(k)}-2\right)\right)}{(\nu^{(k)}-2)(1-F_T(0 \mid \mu^{(k)}, (\sigma^{(k)})^2, \nu^{(k)}))} - \nu^{(k)}(\sigma^{(k)})^2 \right.$$
$$\left. -(\mu^{(k)})^2 + 2\mu^{(k)}e_{3i}^{(k)}\right).$$

Note that besides $e_{2i}^{(k)}$, all conditional expectations can be calculated in a closed-form only with the help of the t pdf and cdf. Having a closed-form is an advantage for the implementation of fast algorithms for evaluation of the pdf and cdf of the t distribution exist. An alternative is to use Monte Carlo integration to calculate the conditional expectations. Monte Carlo integration can help to speed up the ECM algorithm in higher dimensions. However, the estimates for the expectations are less precise. These representations are used for the implementation in the R package.

The following definition is showing the ECM algorithm for the skew-t distribution. In the first step, it is necessary to choose starting values, which is a crucial point, since the algorithm is sensitive to starting values and can lead to different results. Afterward, E-step and CM-steps are performed until the ECM algorithm converges.

Definition 3.5 (ECM algorithm for fitting skew-t distributions)**.**

(i) Choose starting estimates $\boldsymbol{\theta}^{(0)}$ for all parameters.

(ii) E-step: Given $\boldsymbol{\theta} = \boldsymbol{\theta}^{(k)}$, we compute the needed expectations $e_{1i}^{(k)}, e_{2i}^{(k)}, e_{3i}^{(k)}, e_{4i}^{(k)}$ (see Lemma 3.4).

(iii) CM-steps: Maximizing yields the parameters for the $(k+1)$ step.

a) $\mu^{(k+1)} = \frac{\sum_{i=1}^n x_i e_{1i}^{(k)} - \lambda^{(k)} \sum_{i=1}^n e_{3i}^{(k)}}{\sum_{i=1}^n e_{1i}^{(k)}}$.

b) $\lambda^{(k+1)} = \frac{\sum_{i=1}^n (x_i - \mu^{(k+1)}) e_{3i}^{(k)}}{\sum_{i=1}^n e_{4i}^{(k)}}$.

c) $(\sigma^{(k+1)})^2 = \frac{\sum_{i=1}^n (x_i - \mu^{(k+1)})^2 e_{1i}^{(k)} + (\lambda^{(k+1)})^2 \sum_{i=1}^n e_{4i}^{(k)} - 2\lambda^{(k+1)} \sum_{i=1}^n (x_i - \mu^{(k+1)}) e_{3i}^{(k)}}{n}$.

d) $\nu^{(k+1)} = \underset{\nu}{solve}\left[\ln\left(\frac{\nu}{2}\right) - \psi\left(\frac{\nu}{2}\right) + 1 + \frac{\sum_{i=1}^n e_{2i}^{(k)} - \sum_{i=1}^n e_{1i}^{(k)}}{n} = 0\right]$.

In the definition above, we did not state a stopping criterion. The stopping criterion will be discussed later, as well as how to choose the starting values in practice.

3.1.2 EM algorithm for univariate mixtures of skew-t distributions

In this section, the EM algorithm for the skew-t distribution is extended by an additional parameter vector $\boldsymbol{m}$, which is introduced to specify for each observation to which group it belongs. Recall that the probability density function of the skew-t mixture was defined as

$$f(x) = \sum_{j=1}^{c} \omega_j f_{\mathcal{ST}}(x \mid \mu_j, \sigma_j^2, \lambda_j, \nu_j), \text{ with } \sum_{j=1}^{c} \omega_j = 1.$$

Let $\boldsymbol{x} = (x_1, \ldots, x_n)$, $\boldsymbol{z} = (z_1, \ldots, z_n)$, $\boldsymbol{g} = (g_1, \ldots, g_n)$ and $\boldsymbol{m} = (\boldsymbol{m}_1, \ldots, \boldsymbol{m}_n)$, where $\boldsymbol{m}_i$ follows a multinomial distribution. Hence, the likelihood function of $\boldsymbol{\theta} = (\mu_1, \sigma_1^2, \lambda_1, \nu_1, \ldots, \mu_c, \sigma_c^2, \lambda_c, \nu_c)$ for all parameter given $(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{g}, \boldsymbol{m})$ is defined as

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{g}, \boldsymbol{m}) &= f(x_1, \ldots x_n, z_1, \ldots, z_n, g_1, \ldots, g_n, m_1, \ldots, m_n \mid \boldsymbol{\theta}) \\ &= \prod_{j=1}^{c} \prod_{i=1}^{n} f(x_i, z_i, g_i, m_{ij} \mid \boldsymbol{\theta}) \end{aligned}$$

and the log-likelihood function

$$l(\boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{g}, \boldsymbol{m}) = \sum_{j=1}^{c} \sum_{i=1}^{n} \ln f(x_i, z_i, g_i, m_{ij} \mid \boldsymbol{\theta}).$$

The log-likelihood function $l(\boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{g}, \boldsymbol{m})$ equals

$$
\begin{aligned}
&\sum_{j=1}^{c}\sum_{i=1}^{n} m_{ij} \ln\left[\frac{\omega_j \left(g_i \frac{\nu_j}{2}\right)^{\frac{\nu_j}{2}}}{\sigma_j \pi \Gamma\left(\frac{\nu_j}{2}\right)} \exp\left(-\frac{g_i}{2}\left(\nu_j + \frac{(x_i-\mu_j)^2}{\sigma_j^2+\lambda_j^2} + \frac{\sigma_j^2+\lambda_j^2}{\sigma_j^2}\left(z_i - \frac{\lambda_j(x_i-\mu_j)}{\sigma_j^2+\lambda_j^2}\right)^2\right)\right)\right] \\
&= \sum_{j=1}^{c}\sum_{i=1}^{n} \frac{\nu_j m_{ij} \ln g_i}{2} - \sum_{j=1}^{c}\sum_{i=1}^{n} m_{ij}\left[\frac{g_i}{2}\left(\nu_j + \frac{(x_i-\mu_j)^2}{\sigma_j^2+\lambda_j^2} + \frac{\sigma_j^2+\lambda_j^2}{\sigma_j^2}\left(z_i - \frac{\lambda_j(x_i-\mu_j)}{\sigma_j^2+\lambda_j^2}\right)^2\right)\right] \\
&+ \sum_{j=1}^{c}\sum_{i=1}^{n} m_{ij}\left[\frac{\nu_j}{2}\ln\left(\frac{\nu_j}{2}\right) - \ln\left(\sigma_j \pi \Gamma\left(\frac{\nu_j}{2}\right)\right)\right] + \sum_{j=1}^{c}\sum_{i=1}^{n} m_{ij} \ln \omega_j,
\end{aligned}
$$

which is the previous log-likelihood summed over the c components. The parameter $\boldsymbol{m}_i$ follows a multinomial distribution. Hence, we calculate the expectation of the log-likelihood function $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)}) = \mathrm{E}(l(\boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{g}, \boldsymbol{m}) \mid \boldsymbol{x}, \boldsymbol{\theta}^{(k)})$, as

$$
\begin{aligned}
Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)}) &= \sum_{j=1}^{c}\sum_{i=1}^{n} \frac{\nu_j^{(k)}}{2}\mathrm{E}(m_{ij}\ln(g_i) \mid x_i, \boldsymbol{\theta}^{(k)}) - \sum_{j=1}^{c}\sum_{i=1}^{n} \frac{\nu_j^{(k)}}{2}\mathrm{E}(m_{ij} g_i \mid x_i, \boldsymbol{\theta}^{(k)}) \\
&- \sum_{j=1}^{c}\sum_{i=1}^{n} \frac{(x_i - \mu_j^{(k)})^2 \mathrm{E}(m_{ij} z_i g_i \mid x_i, \boldsymbol{\theta}^{(k)})}{2((\sigma_j^{(k)})^2 + (\lambda_j^{(k)})^2)} \\
&- \sum_{j=1}^{c}\sum_{i=1}^{n} \frac{(\sigma_j^{(k)})^2 + (\lambda_j^{(k)})^2}{2(\sigma_j^{(k)})^2}\mathrm{E}\left(m_{ij} g_i \left(z_i - \frac{\lambda_j^{(k)}(x_i-\mu_j^{(k)})}{(\sigma_j^{(k)})^2 + (\lambda_j^{(k)})^2}\right)^2 \middle| \, x_i, \boldsymbol{\theta}^{(k)}\right) \\
&- \sum_{j=1}^{c}\sum_{i=1}^{n} \mathrm{E}(m_{ij} \mid x_i, \boldsymbol{\theta}^{(k)})\left[\frac{\nu_j^{(k)}}{2}\ln\left(\frac{\nu_j^{(k)}}{2}\right) - \ln\left(\sigma_j^{(k)} \pi \Gamma\left(\frac{\nu_j^{(k)}}{2}\right)\right)\right] \\
&+ \sum_{j=1}^{c}\sum_{i=1}^{n} \mathrm{E}(m_{ij} \ln \omega_j^{(k)} \mid x_i, \boldsymbol{\theta}^{(k)}).
\end{aligned}
$$

The maximization $\underset{\boldsymbol{\theta}}{argmax}\ Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})$ subject to $\sum_{j=1}^{c} \omega_j^{(k)} = 1$, $\omega_j^{(k)} \geq 0, j = 1, \ldots, n$ is calculated by setting first derivatives to zero. The Lagrangian is used to optimize with respect to the constraints on $w_j^{(k)}$.

Lemma 3.6. *Let $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})$ be the expectation of the log-likelihood function conditional on the data and the parameter estimates at step k of component j of the skew-t mixture. The optimization steps for the ECM algorithms are as follows:*

a) $\omega_j^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}(m_{ij} \mid x_i, \boldsymbol{\theta}^{(k)})$.

b) $\mu_j^{(k+1)} = \frac{\sum_{i=1}^{n} x_i \mathrm{E}(m_{ij} g_i | x_i, \boldsymbol{\theta}^{(k)}) - \lambda_j^{(k)} \sum_{i=1}^{n} \mathrm{E}(m_{ij} g_i z_i | x_i, \boldsymbol{\theta}^{(k)})}{\sum_{i=1}^{n} \mathrm{E}(m_{ij} g_i | x_i, \boldsymbol{\theta}^{(k)})}$.

c) $\lambda_j^{(k+1)} = \frac{\sum_{i=1}^{n}(x_i - \mu_j^{(k+1)}) \mathrm{E}(m_{ij} g_i z_i | x_i, \boldsymbol{\theta}^{(k)})}{\sum_{i=1}^{n} \mathrm{E}(m_{ij} g_i z_i^2 | x_i, \boldsymbol{\theta}^{(k)})}$.

d) $(\sigma_j^{(k+1)})^2 = \frac{1}{\sum_{i=1}^{n} \mathrm{E}(m_{ij}|\boldsymbol{x}_i, \boldsymbol{\theta}^{(k)})} \sum_{i=1}^{n}(x_i - \mu_j^{(k+1)})^2 \mathrm{E}(m_{ij} g_i \mid x_i, \boldsymbol{\theta}^{(k)})$
$+ \frac{1}{\sum_{i=1}^{n} \mathrm{E}(m_{ij}|\boldsymbol{x}_i, \boldsymbol{\theta}^{(k)})} (\lambda_j^{(k+1)})^2 \sum_{i=1}^{n} \mathrm{E}(m_{ij} g_i z_i^2 \mid x_i, \boldsymbol{\theta}^{(k)})$
$- \frac{1}{\sum_{i=1}^{n} \mathrm{E}(m_{ij}|\boldsymbol{x}_i, \boldsymbol{\theta}^{(k)})} 2\lambda_j^{(k+1)} \sum_{i=1}^{n}(x_i - \mu_j^{(k+1)}) \mathrm{E}(m_{ij} g_i z_i \mid x_i, \boldsymbol{\theta}^{(k)})$.

e) $\nu^{(k+1)} = \underset{\nu_j}{sol\,ve}\Bigg[\ln\left(\frac{\nu_j}{2}\right) - \psi\left(\frac{\nu_j}{2}\right) + 1 + \frac{1}{\sum_{i=1}^{n}\mathrm{E}(m_{ij}|\boldsymbol{x}_i,\boldsymbol{\theta}^{(k)})}\sum_{i=1}^{n}\mathrm{E}(m_{ij}\ln(g_i) \mid x_i, \boldsymbol{\theta}^{(k)})$

$- \frac{1}{\sum_{i=1}^{n}\mathrm{E}(m_{ij}|\boldsymbol{x}_i,\boldsymbol{\theta}^{(k)})}\sum_{i=1}^{n}\mathrm{E}(m_{ij}g_i \mid x_i, \boldsymbol{\theta}^{(k)}) = 0\Bigg]$*, where* $\psi(x)$ *denotes the digamma function equal to* $\frac{d}{dx}\ln\Gamma(x)$*.*

The maxima conditional on the allocation of the data to the mixture components can be calculated independently for each component. Hence, the parameter of component j can be calculated as shown above. To implement the steps from Lemma 3.6 $a) - e)$ the conditional expectations need to be calculated. The lemma below shows the conditional expectations expressed as t density and t distribution functions.

Lemma 3.7. *Let* $\mathrm{E}(m_{ij} \mid x_i, \boldsymbol{\theta}^{(k)}), \mathrm{E}(m_{ij}g_i \mid x_i, \boldsymbol{\theta}^{(k)})$*,* $\mathrm{E}(m_{ij}\ln(g_i) \mid x_i, \boldsymbol{\theta}^{(k)})$*,* $\mathrm{E}(m_{ij}g_iz_i \mid x_i, \boldsymbol{\theta}^{(k)})$ *and* $\mathrm{E}(g_iz_i^2 \mid x_i, \boldsymbol{\theta}^{(k)})$ *be the expectations, which are needed for the EM algorithm. These expectations can be calculated as follows:*

a) $\pi_{ij}^{(k)} := \mathrm{E}(m_{ij} \mid x_i, \boldsymbol{\theta}^{(k)}) = P(m_{ij} = 1 \mid x_i, \boldsymbol{\theta}^{(k)}) = \frac{\omega_j^{(k)} f_{\mathcal{ST}}(x_i|\mu_j^{(k)},(\sigma_j^{(k)})^2,\lambda_j^{(k)},\nu_j^{(k)})}{f(x_i|\boldsymbol{\theta}^{(k)})}$*,*
where $f(x_i \mid \boldsymbol{\theta}^{(k)})$ *denotes probability density function of the skew-t mixture.*

b) $e_{1ij}^{(k)} := \mathrm{E}(m_{ij}g_i \mid x_i, \boldsymbol{\theta}^{(k)}) \overset{\textit{law of total probability}}{=} \pi_{ij}^{(k)}\mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)}, m_{ij} = 1)$*.*

c) $e_{2ij}^{(k)} := \mathrm{E}(m_{ij}\ln(g_i) \mid x_i, \boldsymbol{\theta}^{(k)}) = \pi_{ij}^{(k)}\mathrm{E}(\ln(g_i) \mid x_i, \boldsymbol{\theta}^{(k)}, m_{ij} = 1)$*.*

d) $e_{3ij}^{(k)} := \mathrm{E}(m_{ij}g_iz_i \mid x_i, \boldsymbol{\theta}^{(k)}) = \pi_{ij}^{(k)}\mathrm{E}(g_iz_i \mid x_i, \boldsymbol{\theta}^{(k)}, m_{ij} = 1)$ *.*

e) $e_{4ij}^{(k)} := \mathrm{E}(m_{ij}g_iz_i^2 \mid x_i, \boldsymbol{\theta}^{(k)}) = \pi_{ij}^{(k)}\mathrm{E}(g_iz_i^2 \mid x_i, \boldsymbol{\theta}^{(k)}, m_{ij} = 1)$*.*

The solution for $\pi_{ij}^{(k)}$ can be interpreted as posterior probabilities of observations belonging to component j. With the representation from this lemma, we can use the expectation derived in Lemma 3.4 for each mixture component. The expectation is weighted with the probability of each observation belonging to component j. Based on these results, we write down the ECM algorithm for skew-t mixtures. These algorithms were previously provided by Lin (2010) and Lee and McLachlan (2011). However, there are some missing pieces in the derivations and a mistake in the second expectation of Algorithm 3.2 that is corrected in this work. There are many possibilities for stopping criteria since no stopping criterion was mentioned in the original paper by Dempster et al. (1977). For example, the absolute or the relative change in the likelihood function or the parameters would be possibilities to achieve a certain precision for the estimates. As pointed out by Lin (2010), the stopping criterion can also be based on Aitken's acceleration (McLachlan and Krishnan 2007). Aitken's acceleration is defined as

$$a^{(k)} = \frac{l(\boldsymbol{\theta}^{(k+1)}) - l(\boldsymbol{\theta}^{(k)})}{l(\boldsymbol{\theta}^{(k)}) - l(\boldsymbol{\theta}^{(k-1)})} = \frac{l^{(k+1)} - l^{(k)}}{l^{(k)} - l^{(k-1)}},$$

where $l(\cdot)$ denotes the log-likelihood function and $\boldsymbol{\theta}^{(k)}$ the parameter estimates at the kth iteration. The asymptotic estimate of the log-likelihood at iteration $k+1$ is given by

$$l_\infty^{(k+1)} = l^{(k)} + \frac{1}{1-a^{(k)}}(l^{(k+1)} - l^{(k)}).$$

Lindsay (1995) proposed that the algorithm can be considered to have converged when $|l_\infty^{(k+1)} - l^{(k+1)}| < \epsilon$ and the derivation of the so-called Aitken δ^2 method was described. One interpretation of this method is that it is a linear extrapolation for going through the points $(l^{(k-1)}, l^{(k-2)} - l^{(k-1)})$ and $(l^{(k-1)}, l^{(k)} - l^{(k-1)})$. Aitken's acceleration involves not only the difference between the last two likelihoods and therefore seems to be a robust convergence criterion.

The following algorithm introduces the ECM algorithm, which we used for the implementation.

Algorithm 3.2 ECM algorithm for mixtures of univariate skew-t

```
(i) Let $\boldsymbol{\theta}^{(0)}$ be a starting value for posterior mode.
(ii) E-step: Given $\boldsymbol{\theta} = \boldsymbol{\theta}^{(k)}$, we compute the needed expectations $e_{1ij}^{(k)}, e_{2ij}^{(k)}, e_{3ij}^{(k)}, e_{4ij}^{(k)}$.
(iii) CM-steps: Maximizing yields the parameters for the $(k+1)$ step.
    a) $\omega_j^{(k+1)} = \frac{1}{n}\sum_{i=1}^n \pi_{ij}^{(k)}$.
    b) $\mu_j^{(k+1)} = \frac{\sum_{i=1}^n x_i e_{1ij}^{(k)} - \lambda_j^{(k)} \sum_{i=1}^n e_{3ij}^{(k)}}{\sum_{i=1}^n e_{1ij}^{(k)}}$.
    c) $\lambda_j^{(k+1)} = \frac{\sum_{i=1}^n (x_i - \mu_j^{(k+1)}) e_{3ij}^{(k)}}{\sum_{i=1}^n e_{4ij}^{(k)}}$.
    d) $(\sigma_j^{(k+1)})^2 = \frac{1}{\sum_{i=1}^n \pi_{ij}^{(k)}} \sum_{i=1}^n (x_i - \mu_j^{(k+1)})^2 e_{1ij}^{(k)} + \frac{1}{\sum_{i=1}^n \pi_{ij}^{(k)}} (\lambda_j^{(k+1)})^2 \sum_{i=1}^n e_{4ij}^{(k)}$
       $- \frac{1}{\sum_{i=1}^n \pi_{ij}^{(k)}} 2\lambda_j^{(k+1)} \sum_{i=1}^n (x_i - \mu_j^{(k+1)}) e_{3ij}^{(k)}$.
    e) $\nu_j^{(k+1)} = \underset{\nu_j}{solve}\left[\ln\left(\frac{\nu_j}{2}\right) - \psi\left(\frac{\nu_j}{2}\right) + 1 + \frac{\sum_{i=1}^n e_{2ij}^{(k)} - \sum_{i=1}^n e_{1ij}^{(k)}}{\sum_{i=1}^n \pi_{ij}^{(k)}} = 0\right]$.
```

The conditional expectations e_{1ij}, e_{2ij}, e_{3ij}, and e_{4ij}, which are needed for the algorithm are derived in Lemma 3.7.

3.1.3 EM algorithm for mixtures of multivariate skew-t

In this section, we directly begin with mixtures instead of first showing the EM algorithm for the multivariate skew-t distributions. The multivariate case is very similar to the univariate case. However, some simplifications are not possible in the multivariate case since matrix multiplication is not commutative. The first step is to derive the log-likelihood function. Let $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, $\boldsymbol{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$ and $\boldsymbol{g} = (g_1, \ldots, g_n)$ and $\boldsymbol{m} = (\boldsymbol{m}_1, \ldots, \boldsymbol{m}_n)$, where m_i follows a multinomial distribution. Hence the likelihood function of $\boldsymbol{\theta} = (\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \boldsymbol{\lambda}_1, \nu_1, \ldots, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c, \boldsymbol{\lambda}_c, \nu_c)$ given $(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{g})$ is defined as

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{g}) &= f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_n, g_1, \ldots, g_n, \ldots, \boldsymbol{m}_1, \ldots, \boldsymbol{m}_n \mid \boldsymbol{\theta}) \\ &= \prod_{j=1}^{c} \prod_{i=1}^{n} f(\boldsymbol{x}_i, \boldsymbol{z}_i, g_i, m_{ij} \mid \boldsymbol{\theta}),\end{aligned}$$

and the log-likelihood function as

$$l(\boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{g}) = \sum_{j=1}^{c} \sum_{i=1}^{n} \ln f(\boldsymbol{x}_i, \boldsymbol{z}_i, g_i, m_{ij} \mid \boldsymbol{\theta}).$$

In the next step, we calculate densities, which will be useful for deriving expectations in the E-step.

Lemma 3.8. *Let $\boldsymbol{X}, \boldsymbol{Z}$ and G be as in Lemma 2.22. Then:*

a) The joint density function of $\boldsymbol{X}, \boldsymbol{Z}$ and G is given by:

$$f(\boldsymbol{x}, \boldsymbol{z}, g) = \frac{\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}} g^{\frac{\nu}{2}+p-1}}{\pi^p |\boldsymbol{\Sigma}|^{\frac{1}{2}} \Gamma\left(\frac{\nu}{2}\right)} \exp\Bigg(-\frac{g}{2}\Big(\nu + (\boldsymbol{x}-\boldsymbol{\mu})^\top (\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \\ +(\boldsymbol{z}-\boldsymbol{q})^\top \boldsymbol{\Delta}^{-1}(\boldsymbol{z}-\boldsymbol{q})\Big)\Bigg) 1_{\{\boldsymbol{z} \geq \boldsymbol{0},\, g \geq 0\}},$$

where $\boldsymbol{q} = \boldsymbol{\Lambda}(\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$ and $\boldsymbol{\Delta} = (\boldsymbol{I}_p + \boldsymbol{\Lambda}^2\boldsymbol{\Sigma}^{-1})^{-1} = \boldsymbol{I}_p - \boldsymbol{\Lambda}(\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)^{-1}\boldsymbol{\Lambda}$.

b) The joint density function of $\boldsymbol{X}$ and G is given by:

$$f(\boldsymbol{x}, g) = \left(\frac{2}{\pi}\right)^{\frac{p}{2}} \frac{\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}} g^{\frac{\nu+p}{2}-1} |\boldsymbol{I}_p - \boldsymbol{\Lambda}(\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)^{-1}\boldsymbol{\Lambda}|^{\frac{1}{2}}}{|\boldsymbol{\Sigma}|^{\frac{1}{2}} \Gamma\left(\frac{\nu}{2}\right)} \\ \cdot \exp\left(-\frac{g}{2}\left(\nu + (\boldsymbol{x}-\boldsymbol{\mu})^\top (\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)\right) \\ \cdot F_{\boldsymbol{Z}}\left(\sqrt{g}\boldsymbol{\Lambda}(\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \,\Big|\, \boldsymbol{0}, \boldsymbol{I}_p - \boldsymbol{\Lambda}(\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)^{-1}\boldsymbol{\Lambda}\right) 1_{\{g \geq 0\}},$$

where $F_{\boldsymbol{Z}}(\cdot)$ denotes the distribution function of a multivariate normal distribution.

c) The joint density of $\boldsymbol{X}$ and $\boldsymbol{Z}$ is:

$$f(\boldsymbol{x}, \boldsymbol{z}) = \frac{\Gamma(\frac{\nu+2p}{2})\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}}{|\boldsymbol{\Sigma}|^{\frac{1}{2}} \pi^p \Gamma\left(\frac{\nu}{2}\right)} \Bigg(\frac{1}{2}(\nu + (\boldsymbol{x}-\boldsymbol{\mu})^\top (\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \\ +(\boldsymbol{z}-\boldsymbol{q})^\top \boldsymbol{\Delta}^{-1}(\boldsymbol{z}-\boldsymbol{q}))\Bigg)^{-\frac{\nu+2p}{2}} 1_{\{\boldsymbol{z} \geq \boldsymbol{0}\}},$$

where $\boldsymbol{q} = \boldsymbol{\Lambda}(\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$ and $\boldsymbol{\Delta} = \boldsymbol{I}_p - \boldsymbol{\Lambda}(\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)^{-1}\boldsymbol{\Lambda}$.

d) The density of $\boldsymbol{Z}$ conditional on $\boldsymbol{X}$ and G is:

$$f_{\boldsymbol{Z}|\boldsymbol{X}=\boldsymbol{x},G=g}(\boldsymbol{z}) = \frac{f_{\boldsymbol{Z}}\left(\boldsymbol{z} \,\Big|\, \boldsymbol{\Lambda}(\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)^{-1}(\boldsymbol{x}-\boldsymbol{\mu}), \frac{1}{g}\left(\boldsymbol{I}_p - \boldsymbol{\Lambda}(\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)^{-1}\boldsymbol{\Lambda}\right)\right)}{1 - F_{\boldsymbol{Z}}\left(\boldsymbol{0} \,\Big|\, \boldsymbol{\Lambda}(\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)^{-1}(\boldsymbol{x}-\boldsymbol{\mu}), \frac{1}{g}\left(\boldsymbol{I}_p - \boldsymbol{\Lambda}(\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)^{-1}\boldsymbol{\Lambda}\right)\right)} 1_{\{\boldsymbol{z} \geq \boldsymbol{0}\}},$$

showing that $\boldsymbol{Z} \mid \boldsymbol{X}, G \sim \mathcal{MHN}\left(\boldsymbol{\Lambda}(\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)^{-1}(\boldsymbol{x}-\boldsymbol{\mu}), \frac{1}{g}\left(\boldsymbol{I}_p - \boldsymbol{\Lambda}(\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)^{-1}\boldsymbol{\Lambda}\right)\right)$.

e) The density of G conditional on $\boldsymbol{X}$ and $\boldsymbol{Z}$ is:

$$\begin{aligned} f_{G|\boldsymbol{X}=\boldsymbol{x},\boldsymbol{Z}=\boldsymbol{z}}(g) = {} & \frac{g^{\frac{\nu+2p}{2}-1}}{\Gamma\left(\frac{\nu+2p}{2}\right)} \\ & \cdot \frac{\exp\left(-\frac{g}{2}(\nu+(\boldsymbol{x}-\boldsymbol{\mu})^\top(\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)^{-1}(\boldsymbol{x}-\boldsymbol{\mu})+(\boldsymbol{z}-\boldsymbol{q})^\top\boldsymbol{\Delta}^{-1}(\boldsymbol{z}-\boldsymbol{q}))\right)}{\left(\frac{1}{2}\left(\nu+(\boldsymbol{x}-\boldsymbol{\mu})^\top(\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)^{-1}(\boldsymbol{x}-\boldsymbol{\mu})+(\boldsymbol{z}-\boldsymbol{q})^\top\boldsymbol{\Delta}^{-1}(\boldsymbol{z}-\boldsymbol{q})\right)\right)^{-\frac{\nu+2p}{2}}} 1_{\{g\geq 0\}}, \end{aligned}$$

where $\boldsymbol{q} = \boldsymbol{\Lambda}(\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$ and $\boldsymbol{\Delta} = \boldsymbol{I}_p - \boldsymbol{\Lambda}(\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)^{-1}\boldsymbol{\Lambda}$.

f) The density of G conditional on $\boldsymbol{X}$, $f_{G|\boldsymbol{X}=\boldsymbol{x}}(g)$ equals:

$$\begin{aligned} & \frac{g^{\frac{\nu+p}{2}-1}\exp\left(-\frac{g}{2}\left(\nu+(\boldsymbol{x}-\boldsymbol{\mu})^\top(\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)\right)}{\Gamma\left(\frac{\nu+p}{2}\right)\left(\frac{1}{2}\left((\boldsymbol{x}-\boldsymbol{\mu})^\top(\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)^{-1}(\boldsymbol{x}-\boldsymbol{\mu})+\nu\right)\right)^{-\frac{\nu+p}{2}}} \\ & \cdot \frac{F_Z\left(\sqrt{g}\boldsymbol{\Lambda}(\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \;\middle|\; \boldsymbol{0}, \boldsymbol{I}_p - \boldsymbol{\Lambda}(\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)^{-1}\boldsymbol{\Lambda}\right)}{F_{\boldsymbol{T}}\left(\frac{\boldsymbol{\Lambda}(\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\sqrt{\nu+p}}{\sqrt{\nu+(\boldsymbol{x}-\boldsymbol{\mu})^\top(\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}} \;\middle|\; \boldsymbol{0}, \boldsymbol{I}_p - \boldsymbol{\Lambda}(\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)^{-1}\boldsymbol{\Lambda}, \nu+p\right)} 1_{\{g\geq 0\}}. \end{aligned}$$

g) The density of $\boldsymbol{Z}$ conditional on $\boldsymbol{X}$, $f_{\boldsymbol{Z}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{z})$ is:

$$\frac{f_{\boldsymbol{T}}\left(\boldsymbol{z} \;\middle|\; \boldsymbol{\Lambda}(\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)^{-1}(\boldsymbol{x}-\boldsymbol{\mu}), \frac{\left(\nu+(\boldsymbol{x}-\boldsymbol{\mu})^\top(\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)\left(\boldsymbol{I}_p-\boldsymbol{\Lambda}(\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)^{-1}\boldsymbol{\Lambda}\right)}{\nu+p}, \nu+p\right) 1_{\{\boldsymbol{z}\geq\boldsymbol{0}\}}}{1-F_{\boldsymbol{T}}\left(\boldsymbol{0} \;\middle|\; \boldsymbol{\Lambda}(\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)^{-1}(\boldsymbol{x}-\boldsymbol{\mu}), \frac{\left(\nu+(\boldsymbol{x}-\boldsymbol{\mu})^\top(\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)\left(\boldsymbol{I}_p-\boldsymbol{\Lambda}(\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)^{-1}\boldsymbol{\Lambda}\right)}{\nu+p}, \nu+p\right)},$$

showing that $\boldsymbol{Z} \mid \boldsymbol{X}$ is multivariated truncated t distributed.

Using the joint mixture density $f(\boldsymbol{x},\boldsymbol{z},\boldsymbol{g})$, the log-likelihood function $l(\boldsymbol{\theta} \mid \boldsymbol{x},\boldsymbol{z},\boldsymbol{g})$ can be written as

$$\begin{aligned} & \sum_{j=1}^{c}\sum_{i=1}^{n} m_{ij}\ln\left[\omega_j\frac{\left(\frac{\nu_j}{2}\right)^{\frac{\nu_j}{2}} g_i^{\frac{\nu_j}{2}+p-1}}{|\boldsymbol{\Sigma}_j|^{\frac{1}{2}}\pi^p\Gamma\left(\frac{\nu_j}{2}\right)}\exp\left(-\frac{g_i}{2}(\nu_j+(\boldsymbol{x}_i-\boldsymbol{\mu}_j)^\top(\boldsymbol{\Sigma}_j+\boldsymbol{\Lambda}_j^2)^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_j)\right.\right. \\ & \left.\left.+(\boldsymbol{z}_i-\boldsymbol{q}_{ij})^\top\boldsymbol{\Delta}_j^{-1}(\boldsymbol{z}_i-\boldsymbol{q}_{ij}))\right)\right] \\ & = \sum_{j=1}^{c}\sum_{i=1}^{n}\left(\frac{\nu_j}{2}+p-1\right)m_{ij}\ln(g_i) - \sum_{j=1}^{c}\sum_{i=1}^{n} m_{ij}\left[\frac{g_i}{2}(\nu_j+(\boldsymbol{x}_i-\boldsymbol{\mu}_j)^\top(\boldsymbol{\Sigma}_j+\boldsymbol{\Lambda}_j^2)^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_j)\right. \\ & \left.+(\boldsymbol{z}_i-\boldsymbol{q}_{ij})^\top\boldsymbol{\Delta}_j^{-1}(\boldsymbol{z}_i-\boldsymbol{q}_{ij}))\right] \\ & + \sum_{j=1}^{c}\sum_{i=1}^{n} m_{ij}\left[\frac{\nu_j}{2}\ln\left(\frac{\nu_j}{2}\right) - p\ln(\pi) - \ln\left(\Gamma\left(\frac{\nu_j}{2}\right)\right) - \frac{1}{2}\ln(|\boldsymbol{\Sigma}_j|)\right] \\ & + \sum_{j=1}^{c}\sum_{i=1}^{n} m_{ij}\ln\omega_j, \end{aligned}$$

where $\boldsymbol{q}_{ij} = \boldsymbol{\Lambda}_j(\boldsymbol{\Sigma}_j+\boldsymbol{\Lambda}_j^2)^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_j)$ and $\boldsymbol{\Delta}_j = \boldsymbol{I}_p - \boldsymbol{\Lambda}_j(\boldsymbol{\Sigma}_j+\boldsymbol{\Lambda}_j^2)^{-1}\boldsymbol{\Lambda}_j$.

The Q-function can be calculated by taking the expectation of the log-likelihood function yielding

$$
\begin{aligned}
&Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)}) = \mathrm{E}(l(\boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{g}) \mid \boldsymbol{x}, \boldsymbol{\theta}^{(k)}) \\
&= \sum_{j=1}^{c} \sum_{i=1}^{n} \left(\frac{\nu_j^{(k)}}{2} + p - 1 \right) \mathrm{E}(m_{ij} \ln(g_i) \mid \boldsymbol{x}_i, \boldsymbol{\theta}^{(k)}) \\
&- \sum_{j=1}^{c} \sum_{i=1}^{n} \frac{\nu_j^{(k)}}{2} \mathrm{E}(m_{ij} g_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}^{(k)}) \\
&- \sum_{j=1}^{c} \sum_{i=1}^{n} \mathrm{E}(m_{ij} g_i (\boldsymbol{x}_i - \boldsymbol{\mu}_j^{(k)})^\top (\boldsymbol{\Sigma}_j^{(k)} + (\boldsymbol{\Lambda}_j^{(k)})^2)^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_j^{(k)}) \mid \boldsymbol{x}_i, \boldsymbol{\theta}^{(k)}) \\
&- \sum_{j=1}^{c} \sum_{i=1}^{n} \mathrm{E}\left(m_{ij} g_i (\boldsymbol{z}_i - \boldsymbol{q}_{ij}^{(k)})^\top (\boldsymbol{\Delta}_j^{(k)})^{-1} (\boldsymbol{z}_i - \boldsymbol{q}_{ij}^{(k)}) \;\middle|\; \boldsymbol{x}_i, \boldsymbol{\theta}^{(k)} \right) \\
&+ \sum_{j=1}^{c} \sum_{i=1}^{n} \mathrm{E}(m_{ij} \mid \boldsymbol{x}_i, \boldsymbol{\theta}^{(k)}) \left[\frac{\nu_j^{(k)}}{2} \ln\left(\frac{\nu_j^{(k)}}{2} \right) - p \ln(\pi) - \ln\left(\Gamma\left(\frac{\nu_j^{(k)}}{2} \right) \right) - \frac{1}{2} \ln(|\boldsymbol{\Sigma}_j^{(k)}|) \right] \\
&+ \sum_{j=1}^{c} \sum_{i=1}^{n} \mathrm{E}(m_{ij} \ln \omega_j^{(k)} \mid \boldsymbol{x}_i, \boldsymbol{\theta}^{(k)}),
\end{aligned}
$$

where $\boldsymbol{q}_{ij}^{(k)} = \boldsymbol{\Lambda}_j^{(k)} (\boldsymbol{\Sigma}_j^{(k)} + (\boldsymbol{\Lambda}_j^{(k)})^2)^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_j^{(k)})$ and $\boldsymbol{\Delta}_j^{(k)} = \boldsymbol{I}_p - \boldsymbol{\Lambda}_j^{(k)} (\boldsymbol{\Sigma}_j^{(k)} + (\boldsymbol{\Lambda}_j^{(k)})^2)^{-1} \boldsymbol{\Lambda}_j^{(k)}$. Again the Q-function is a sum of several conditional expectations. These conditional expectations are later explicitly stated for the multivariate skew-t mixtures. In the following, the results of the CM-steps are shown. Hence, for each parameter, the optimal value of the parameter conditional on the previous value and conditional on the data can be calculated with those conditional expectations.

Lemma 3.9. *Let $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})$ be the expectation of the log-likelihood function conditional on the data and the parameter estimates at step k. The optimization steps for the ECM algorithms are as follows:*

a) $\omega_j^{(k+1)} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}(m_{ij} \mid \boldsymbol{x}_i, \boldsymbol{\theta}^{(k)})$.

b) $\boldsymbol{\mu}_j^{(k+1)} = \left[\sum_{i=1}^{n} \boldsymbol{x}_i \mathrm{E}(m_{ij} g_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}^{(k)}) - \boldsymbol{\Lambda}_j^{(k)} \sum_{i=1}^{n} \mathrm{E}(m_{ij} g_i \boldsymbol{z}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}^{(k)}) \right]$
$\cdot \left(\sum_{i=1}^{n} \mathrm{E}(m_{ij} g_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}^{(k)}) \right)^{-1}$.

c) $\boldsymbol{\lambda}_j^{(k+1)} = \left[(\boldsymbol{\Sigma}_j^{(k)})^{-1} \odot \boldsymbol{B}_{1j}^{(k)} \right]^{-1} \left[(\boldsymbol{\Sigma}_j^{(k)})^{-1} \odot \boldsymbol{B}_{2j}^{(k)} \right]$, *where $\odot$ denotes component-wise matrix multiplication (Hadamard product).*

d) $\boldsymbol{\Sigma}_j^{(k+1)} = \frac{1}{\sum_{i=1}^{n} \mathrm{E}(m_{ij} \mid \boldsymbol{x}_i, \boldsymbol{\theta}^{(k)})} \Big[\sum_{i=1}^{n} \mathrm{E}(m_{ij} g_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}^{(k)}) (\boldsymbol{x}_i - \boldsymbol{\mu}_j^{(k+1)}) (\boldsymbol{x}_i - \boldsymbol{\mu}_j^{(k+1)})^\top +$
$\boldsymbol{\Lambda}_j^{(k+1)} \boldsymbol{B}_{1j}^{(k)} \boldsymbol{\Lambda}_j^{(k+1)} - \boldsymbol{\Lambda}_j^{(k+1)} \boldsymbol{B}_{2j}^{(k)} - (\boldsymbol{B}_{2j}^{(k)})^\top \boldsymbol{\Lambda}_j^{(k+1)} \Big]$.

e) $\nu_j^{(k+1)} = \underset{\nu_j}{solve} \Bigg[\ln\left(\frac{\nu_j}{2} \right) - \psi\left(\frac{\nu_j}{2} \right) + 1 + \frac{1}{\sum_{i=1}^{n} \mathrm{E}(m_{ij} \mid \boldsymbol{x}_i, \boldsymbol{\theta}^{(k)})} \Big[\sum_{i=1}^{n} \mathrm{E}(m_{ij} \ln(g_i) \mid \boldsymbol{x}_i, \boldsymbol{\theta}^{(k)})$
$- \sum_{i=1}^{n} \mathrm{E}(m_{ij} g_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}^{(k)}) \Big] = 0 \Bigg]$, *where $\psi(x)$ denotes the digamma function which*

is equal to $\frac{d}{dx}\ln\Gamma(x)$*, where* $\boldsymbol{B}_{1j}^{(k)} = \sum_{i=1}^{n}\mathrm{E}(m_{ij}g_i\boldsymbol{z}_i\boldsymbol{z}_i^\top \mid \boldsymbol{x}_i,\boldsymbol{\theta}^{(k)})$ *and* $\boldsymbol{B}_{2j}^{(k)} = \sum_{i=1}^{n}\mathrm{E}(m_{ij}g_i\boldsymbol{z}_i \mid \boldsymbol{x}_i,\boldsymbol{\theta}^{(k)})(\boldsymbol{x}_i-\boldsymbol{\mu}_j^{(k+1)})^\top$.

Like for the univariate case, the derivative of a multivariate t distribution is calculated. The derivation is done to explicitly to calculate the conditional expectations, which are summands of the Q-function.

Lemma 3.10. *Let* $\boldsymbol{T} \sim \mathcal{MVT}\left(\frac{\boldsymbol{\Lambda}(\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\sqrt{\nu+p}}{\sqrt{\nu+(\boldsymbol{x}-\boldsymbol{\mu})^\top(\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}} \,\middle|\, \boldsymbol{0}, \boldsymbol{I}_p - \boldsymbol{\Lambda}(\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)^{-1}\boldsymbol{\Lambda}, \nu+p\right)$*. Then the derivative with respect to the parameter* ν *of the t distribution function is*

$$\frac{1}{2}\int_{\boldsymbol{A}} f_{\boldsymbol{T}}\left(\boldsymbol{t}, \boldsymbol{0}, \frac{\left(\nu+(\boldsymbol{x}-\boldsymbol{\mu})^\top(\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)(\boldsymbol{I}_p-\boldsymbol{\Lambda}(\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)^{-1}\boldsymbol{\Lambda})}{\nu+p}, \nu+p\right) c(\boldsymbol{t})\, d\boldsymbol{t},$$

where the set $\boldsymbol{A}$ *is defined as* $\{\boldsymbol{t} \mid \boldsymbol{t} \leq \boldsymbol{\Lambda}(\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\}$ *and* $c(\boldsymbol{t})$ *equals*

$$\begin{aligned}
&\psi\left(\frac{\nu}{2}+p\right) - \psi\left(\frac{\nu+p}{2}\right) - \frac{p}{\nu+(\boldsymbol{x}-\boldsymbol{\mu})^\top(\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)^{-1}(\boldsymbol{x}-\boldsymbol{\mu})} \\
&+ \frac{\left(\nu+(\boldsymbol{x}-\boldsymbol{\mu})^\top(\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)^{-1}(\boldsymbol{t}^\top(\boldsymbol{I}_p-\boldsymbol{\Lambda}(\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)^{-1}\boldsymbol{\Lambda})^{-1}\boldsymbol{t})(\nu+2p)}{\left(\nu+(\boldsymbol{x}-\boldsymbol{\mu})^\top(\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)^{-1}(\boldsymbol{x}-\boldsymbol{\mu})+\boldsymbol{t}^\top(\boldsymbol{I}_p-\boldsymbol{\Lambda}(\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)^{-1}\boldsymbol{\Lambda})^{-1}\boldsymbol{t}\right)} \\
&- \ln\left(1+\frac{\boldsymbol{t}^\top(\boldsymbol{I}_p-\boldsymbol{\Lambda}(\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)^{-1}\boldsymbol{\Lambda})^{-1}\boldsymbol{t}}{\left(\nu+(\boldsymbol{x}-\boldsymbol{\mu})^\top(\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)}\right).
\end{aligned}$$

In the next lemma, we explicitly calculate the quantities needed for the E-step.

Lemma 3.11. *Let* $\mathrm{E}(m_{ij} \mid \boldsymbol{x}_i,\boldsymbol{\theta}^{(k)})$*,* $\mathrm{E}(g_i \mid \boldsymbol{x}_i,\boldsymbol{\theta}^{(k)})$*,* $\mathrm{E}(\ln(g_i) \mid \boldsymbol{x}_i,\boldsymbol{\theta}^{(k)})$*,* $\mathrm{E}(g_i\boldsymbol{z}_i \mid \boldsymbol{x}_i,\boldsymbol{\theta}^{(k)})$ *and* $\mathrm{E}(g_i\boldsymbol{z}_i^\top\boldsymbol{z}_i \mid \boldsymbol{x}_i,\boldsymbol{\theta}^{(k)})$ *be the expectations, which are needed for the EM algorithm. These expectations can be calculated as follows:*

a) $\pi_{ij}^{(k)} := \mathrm{E}(m_{ij} \mid \boldsymbol{x}_i,\boldsymbol{\theta}^{(k)}) = P(m_{ij}=1 \mid \boldsymbol{x}_i,\boldsymbol{\theta}^{(k)}) = \frac{\omega_j^{(k)} f_{\mathcal{MST}}(\boldsymbol{x}_i|\boldsymbol{\mu}_j^{(k)},\boldsymbol{\Sigma}_j^{(k)},\boldsymbol{\lambda}_j^{(k)},\nu_j^{(k)})}{f(\boldsymbol{x}_i|\boldsymbol{\theta}^{(k)})}$*, where* $f(\boldsymbol{x}_i \mid \boldsymbol{\theta}^{(k)})$ *denotes probability density function of the multivariate skew-t mixture.*

b) $e_{1ij}^{(k)} := \mathrm{E}(m_{ij}g_i \mid \boldsymbol{x}_i,\boldsymbol{\theta}^{(k)}) = \pi_{ij}^{(k)}\mathrm{E}(g_i \mid \boldsymbol{x}_i,\boldsymbol{\theta}^{(k)}, m_{ij}=1)$

$$\begin{aligned}
&= \frac{\pi_{ij}^{(k)}(\nu_j^{(k)}+p)}{\left(\nu_j^{(k)}+(\boldsymbol{x}_i-\boldsymbol{\mu}_j^{(k)})^\top(\boldsymbol{\Sigma}_j^{(k)}+(\boldsymbol{\Lambda}_j^{(k)})^2)^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_j^{(k)})\right)} \\
&\cdot \frac{F_{\boldsymbol{T}}\left(\frac{\boldsymbol{\Lambda}_j^{(k)}(\boldsymbol{\Sigma}_j^{(k)}+(\boldsymbol{\Lambda}_j^{(k)})^2)^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_j^{(k)})\sqrt{\nu_j^{(k)}+p}}{\sqrt{\nu_j^{(k)}+(\boldsymbol{x}_i-\boldsymbol{\mu}_j^{(k)})^\top(\boldsymbol{\Sigma}_j^{(k)}+(\boldsymbol{\Lambda}_j^{(k)})^2)^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_j^{(k)})}} \,\middle|\, \boldsymbol{0}, \boldsymbol{\Delta}_j^{(k)}, \nu_j^{(k)}+p+2\right)}{F_{\boldsymbol{T}}\left(\frac{\boldsymbol{\Lambda}_j^{(k)}(\boldsymbol{\Sigma}_j^{(k)}+(\boldsymbol{\Lambda}_j^{(k)})^2)^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_j^{(k)})\sqrt{\nu_j^{(k)}+p}}{\sqrt{\nu_j^{(k)}+(\boldsymbol{x}_i-\boldsymbol{\mu}_j^{(k)})^\top(\boldsymbol{\Sigma}_j^{(k)}+(\boldsymbol{\Lambda}_j^{(k)})^2)^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_j^{(k)})}} \,\middle|\, \boldsymbol{0}, \boldsymbol{\Delta}_j^{(k)}, \nu_j^{(k)}+p\right)},
\end{aligned}$$

where $\boldsymbol{\Delta}_j^{(k)} = \boldsymbol{I}_p - \boldsymbol{\Lambda}_j^{(k)}(\boldsymbol{\Sigma}_j^{(k)}+(\boldsymbol{\Lambda}_j^{(k)})^2)^{-1}\boldsymbol{\Lambda}_j^{(k)}$.

c) $e_{2ij}^{(k)} := \mathrm{E}(m_{ij}\ln(g_i) \mid \boldsymbol{x}_i, \boldsymbol{\theta}^{(k)}) = \pi_{ij}^{(k)}\mathrm{E}(\ln(g_i) \mid \boldsymbol{x}_i, \boldsymbol{\theta}^{(k)}, m_{ij} = 1)$

$$
\begin{aligned}
&= e_{1ij}^{(k)} - \pi_{ij}^{(k)} \frac{\nu_j^{(k)} + p}{\nu_j^{(k)} + (\boldsymbol{x}_i - \boldsymbol{\mu}_j^{(k)})^\top (\boldsymbol{\Sigma}_j^{(k)} + (\boldsymbol{\Lambda}_j^{(k)})^2)^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_j^{(k)})} \\
&- \pi_{ij}^{(k)} \ln\left(\frac{1}{2}\left[\nu_j^{(k)} + (\boldsymbol{x}_i - \boldsymbol{\mu}_j^{(k)})^\top (\boldsymbol{\Sigma}_j^{(k)} + (\boldsymbol{\Lambda}_j^{(k)})^2)^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_j^{(k)})\right]\right) \\
&+ \pi_{ij}^{(k)} \psi\left(\frac{\nu_j^{(k)} + p}{2}\right) \\
&+ \pi_{ij}^{(k)} \int_{\boldsymbol{A}_{ij}^{(k)}} \frac{f_{\boldsymbol{T}}\left(\boldsymbol{t}, \boldsymbol{0}, \frac{\left(\nu_j^{(k)} + (\boldsymbol{x}_i - \boldsymbol{\mu}_j^{(k)})^\top (\boldsymbol{\Sigma}_j^{(k)} + (\boldsymbol{\Lambda}_j^{(k)})^2)^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_j^{(k)})\right)\boldsymbol{\Delta}_j^{(k)}}{\nu_j^{(k)} + p}, \nu_j^{(k)} + p\right)}{F_{\boldsymbol{T}}\left(\boldsymbol{x}_j^{(k)*} \,\middle|\, \boldsymbol{0}, \boldsymbol{I}_p - \boldsymbol{\Lambda}_j^{(k)}(\boldsymbol{\Sigma}_j^{(k)} + (\boldsymbol{\Lambda}_j^{(k)})^2)^{-1}\boldsymbol{\Lambda}_j^{(k)}, \nu_j^{(k)} + p\right)} c(\boldsymbol{t})d\boldsymbol{t},
\end{aligned}
$$

where $\boldsymbol{x}_j^{(k)*} = \frac{\boldsymbol{\Lambda}_j^{(k)}(\boldsymbol{\Sigma}_j^{(k)} + (\boldsymbol{\Lambda}_j^{(k)})^2)^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_j^{(k)})\sqrt{\nu_j^{(k)} + p}}{\sqrt{\nu_j^{(k)} + (\boldsymbol{x}_i - \boldsymbol{\mu}_j^{(k)})^\top (\boldsymbol{\Sigma}_j^{(k)} + (\boldsymbol{\Lambda}_j^{(k)})^2)^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_j^{(k)})}}$, *the set* $\boldsymbol{A}_{ij}^{(k)}$ *is defined as* $\{\boldsymbol{t} \mid \boldsymbol{t} \leq \boldsymbol{\Lambda}_j^{(k)}(\boldsymbol{\Sigma}_j^{(k)} + (\boldsymbol{\Lambda}_j^{(k)})^2)^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_j^{(k)})\}$, *and* $\boldsymbol{\Delta}_j^{(k)} = \boldsymbol{I}_p - \boldsymbol{\Lambda}_j^{(k)}(\boldsymbol{\Sigma}_j^{(k)} + (\boldsymbol{\Lambda}_j^{(k)})^2)^{-1}\boldsymbol{\Lambda}_j^{(k)}$.

d) $\boldsymbol{e}_{3ij}^{(k)} := \mathrm{E}(m_{ij}g_i\boldsymbol{z}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}^{(k)}) = e_{1ij}^{(k)}\mathrm{E}(\boldsymbol{z}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}^{(k)}, m_{ij} = 1)$.

e) $\boldsymbol{e}_{4ij}^{(k)} := \mathrm{E}(m_{ij}g_i\boldsymbol{z}_i\boldsymbol{z}_i^\top \mid \boldsymbol{x}_i, \boldsymbol{\theta}^{(k)}) = e_{1ij}^{(k)}\mathrm{E}(\boldsymbol{z}_i\boldsymbol{z}_i^\top \mid \boldsymbol{x}_i, \boldsymbol{\theta}^{(k)}, m_{ij} = 1)$.

The first and second moment of multivariate trunacted t distributions required for c) and d) can be calculated fast using the algorithm provided by O'hagan (1976). This algorithm was implemented in the R package `fitmixst4`.

In the following, the ECM algorithm is defined for multivariate skew-t mixtures.

Algorithm 3.3 ECM algorithm for multivariate skew-t mixtures

(i) Let $\boldsymbol{\theta}^{(0)}$ be a starting value for the maximum likelihood estimates.

(ii) E-step: Given $\boldsymbol{\theta} = \boldsymbol{\theta}^{(k)}$, we compute the needed expectations $e_{1ij}^{(k)}, e_{2ij}^{(k)}, \boldsymbol{e}_{3ij}^{(k)}, \boldsymbol{e}_{4ij}^{(k)}$. Note that all those quantities can be directly calculated using high efficient methods for evaluating t densities.

(iii) CM-steps: Maximizing yields the parameters for the $(k+1)$ step.

- a) $\omega_j^{(k+1)} = \frac{1}{n}\sum_{i=1}^n \pi_{ij}^{(k)}$.
- b) $\boldsymbol{\mu}_j^{(k+1)} = \left[\sum_{i=1}^n \boldsymbol{x}_i e_{1ij}^{(k)} - \boldsymbol{\Lambda}_j^{(k)}\sum_{i=1}^n \boldsymbol{e}_{3ij}^{(k)}\right]\left(\sum_{i=1}^n e_{1ij}^{(k)}\right)^{-1}$.
- c) $\boldsymbol{\lambda}_j^{(k+1)} = \left[(\boldsymbol{\Sigma}_j^{(k)})^{-1} \odot \boldsymbol{B}_{1j}^{(k)}\right]^{-1}\left[(\boldsymbol{\Sigma}_j^{(k)})^{-1} \odot \boldsymbol{B}_{2j}^{(k)}\right]$.
- d) $\boldsymbol{\Sigma}_j^{(k+1)} = \frac{1}{\sum_{i=1}^n \pi_{ij}^{(k)}}\Big[\sum_{i=1}^n e_{1ij}^{(k)}(\boldsymbol{x}_i - \boldsymbol{\mu}_j^{(k+1)})(\boldsymbol{x}_i - \boldsymbol{\mu}_j^{(k+1)})^\top + \boldsymbol{\Lambda}_j^{(k+1)}\boldsymbol{B}_{1j}^{(k)}\boldsymbol{\Lambda}_j^{(k+1)} - \boldsymbol{\Lambda}_j^{(k+1)}\boldsymbol{B}_{2j}^{(k)} - (\boldsymbol{B}_{2j}^{(k)})^\top\boldsymbol{\Lambda}_j^{(k+1)}\Big]$.
- e) $\nu_j^{(k+1)} = \underset{\nu_j}{solve}\left[\ln\left(\frac{\nu_j}{2}\right) - \psi\left(\frac{\nu_j}{2}\right) + 1 + \frac{1}{\sum_{i=1}^n \pi_{ij}^{(k)}}\left[\sum_{i=1}^n e_{2ij}^{(k)} - \sum_{i=1}^n e_{1ij}^{(k)}\right] = 0\right]$,

  where $\psi(x)$ denotes the digamma function which is equal to $\frac{d}{dx}\ln\Gamma(x)$, where $\boldsymbol{B}_{1j}^{(k)} = \sum_{i=1}^n \boldsymbol{e}_{4ij}^{(k)}$ and $\boldsymbol{B}_{2j}^{(k)} = \sum_{i=1}^n \boldsymbol{e}_{3ij}^{(k)}(\boldsymbol{x}_i - \boldsymbol{\mu}_j^{(k+1)})^\top$.

3.2 Constrained EM algorithm

In the literature, there is lots of discussion about parameter identifiability for mixture distributions. The problem is that if we simulate data from, e.g., a two-component skew-t mixture, there can be outliers. Since the model is fit by maximizing the likelihood function, we need to be aware that few outliers can produce a spike in the empirical density function. These spikes can contribute a large part to the likelihood. Hence, the EM algorithm might choose some of the peaks as an individual component of the mixture, even if this can be mainly outlier. Especially in these situations, the choice of starting values is crucial. There are many approaches in the literature to circumvent this problem.

However, in this part, we will focus on a different problem that is related to identifiability. It is important to note that if we work with skew-t distributions, the likelihood can be singular. Singularities appear in the simple univariate case if the scale parameter σ^2 approaches zero. In this case, the pdf goes to infinity, which leads to convergence problems. The EM algorithm can crash if it is not well programmed and the denominator is zero. Now, this situation may sound not so common, but with real-world data, this can happen quite quickly. One example of this effect can be that data is recorded with a measurement instrument which has a particular precision. So, if a value is below the limit of quantification, just the limit of the measurement instrument is recorded. Then the density of the measured parameter could look like a normal distribution which is truncated at the threshold value and instead of continuing has a peak at the threshold. Another possibility is that a variable is derived from a continuous measurement by a particular formula and not well defined. This truncation can happen in applications where practical measures are required for real-world decisions. In both situations, the variance of one component can approach zero, and hence, the EM algorithm fails.

Now one approach would be to use semi-continuous models such as a hurdle model or other hierarchical models, which will first determine if an observation belongs to the continuous component or not and afterward fit the density to each component. However, we assume the values are not naturally fixed. As described above the data would be continuous but is semi-continuous based on, e.g., the measurement method. However, when estimating a mixture model, the peak can lead to biased estimates or not estimable parameters. Hence, we want to restrict only the likelihood function of the collapsed component. Constraints for the likelihood function helps us to get unbiased estimates of the other components and still estimate the peak in the framework of mixture models.

To motivate the idea of the constrained ECM algorithm, we show a simple univariate problem.

Example 3.12. In this example we take a sample of $n_1 = 100$ skew-t distributed random variables with parameter $\mu = 1, \sigma^2 = 3, \lambda = 2$ and $\nu = 10$. Now we add $n_2 = 20$ observations which have the value -1. The total data set now has 120 observations and a collapsed cluster at -1. Figure 7 shows the histogram of the data set described above. The overlaid densities are estimated using the skew-t mixture with constrained scale (green), the skew-t distribution (red) and a non-parametric kernel density approach. We can see that the collapse cluster

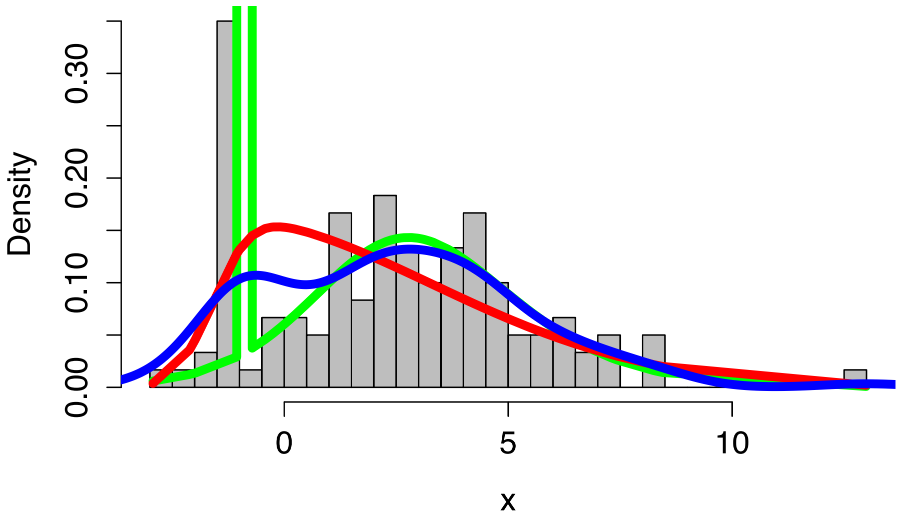


Figure 7 Histogram of univariate data with a collapsed cluster. Overlaid the best fitting one-component skew-t mixture (red), two-component mixture with constrained scale parameter (green) and kernel density estimator (blue).

heavily influences the estimates for the kernel density as well as for the one component skew-t distribution. In the case of the one component skew-t distribution, the estimated parameters do not reflect the true parameters from which the data were simulated. Now integrating an additional component to account for the collapsed cluster yields better estimates for the distribution.

Algorithm 3.4 Modified ECM algorithm for multivariate skew-t mixtures

```
  (i) Let $\boldsymbol{\theta}^{(0)}$ be a starting value for the maximum likelihood estimates.
 (ii) E-step: Given $\boldsymbol{\theta} = \boldsymbol{\theta}^{(k)}$, we compute the needed expectations $e_{1ij}^{(k)}, e_{2ij}^{(k)}, \boldsymbol{e}_{3ij}^{(k)}, \boldsymbol{e}_{4ij}^{(k)}$ Note that all those quantities can be directly calculated using high efficient methods for evaluating t densities.
(iii) CM-steps: Maximizing yields the parameters for the $(k+1)$ step.
      a) $\omega_j^{(k+1)} = \frac{1}{n} \sum_{i=1}^n \pi_{ij}^{(k)}$.
      b) $\boldsymbol{\mu}_j^{(k+1)} = \left[\sum_{i=1}^n \boldsymbol{x}_i e_{1ij}^{(k)} - \boldsymbol{\Lambda}_j^{(k)} \sum_{i=1}^n \boldsymbol{e}_{3ij}^{(k)}\right] \left(\sum_{i=1}^n e_{1ij}^{(k)}\right)^{-1}$.
      c) $\boldsymbol{\lambda}_j^{(k+1)} = \left[(\boldsymbol{\Sigma}_j^{(k)})^{-1} \odot \boldsymbol{B}_{1j}^{(k)}\right]^{-1} \left[(\boldsymbol{\Sigma}_j^{(k)})^{-1} \odot \boldsymbol{B}_{2j}^{(k)}\right]$.
      d) $\boldsymbol{\Sigma}_j^{(k+1)} = \frac{1}{\sum_{i=1}^n \pi_{ij}^{(k)}} \Big[\sum_{i=1}^n e_{1ij}^{(k)} (\boldsymbol{x}_i - \boldsymbol{\mu}_j^{(k+1)})(\boldsymbol{x}_i - \boldsymbol{\mu}_j^{(k+1)})^\top + \boldsymbol{\Lambda}_j^{(k+1)} \boldsymbol{B}_{1j}^{(k)} \boldsymbol{\Lambda}_j^{(k+1)} - \boldsymbol{\Lambda}_j^{(k+1)} \boldsymbol{B}_{2j}^{(k)} - (\boldsymbol{B}_{2j}^{(k)})^\top \boldsymbol{\Lambda}_j^{(k+1)}\Big]$.
      e) If the determinant $|\boldsymbol{\Sigma}_j^{(k+1)}| < \varepsilon$ then the diagonal elements smaller than $\varepsilon$ are set to $\varepsilon$ and all other elements in the respective column or row to zero.
      f) $\nu_j^{(k+1)} = \underset{\nu_j}{solve} \left[ \ln\left(\frac{\nu_j}{2}\right) - \psi\left(\frac{\nu_j}{2}\right) + 1 + \frac{1}{\sum_{i=1}^n \pi_{ij}^{(k)}} \left[\sum_{i=1}^n e_{2ij}^{(k)} - \sum_{i=1}^n e_{1ij}^{(k)}\right] = 0 \right]$,
         where $\psi(x)$ denotes the digamma function which is equal to $\frac{d}{dx} \ln \Gamma(x)$, where $\boldsymbol{B}_{1j}^{(k)} = \sum_{i=1}^n \boldsymbol{e}_{4ij}^{(k)}$ and $\boldsymbol{B}_{2j}^{(k)} = \sum_{i=1}^n \boldsymbol{e}_{3ij}^{(k)} (\boldsymbol{x}_i - \boldsymbol{\mu}_j^{(k+1)})^\top$.
 (iv) Repeat step (ii) and (ii) until $|l_\infty^{(k+1)} - l^{(k)}| < \epsilon$.
```

The adaption in the algorithm ensures that the matrix is invertible. If the determinant of the variance is smaller than a fixed value, we change the variance in this step. The result is that if we fit data, which has two components but the variance of one component is by far smaller than the variance of the other component/s, then the algorithm still can find an optimum considering the implemented constrained. This effect can be observed in the example shown in Figure 7.

As an alternative to changing matrix in step $d)$ we can also use the pseudoinverse of $\boldsymbol{\Sigma}_j$. The pseudoinverse is a generalization of the inverse matrix. The pseudoinverse can be intuitively understood as the closest solution identifying the inverse matrix in an equation with minimizing the L^2 norm. This approach has been implemented additionally.

3.2.1 Skew-t regression

The skew-t regression incorporates additional coefficients as covariates. The error distribution is a skew-t distribution. The EM algorithm is built in that way that $\boldsymbol{X}\boldsymbol{\beta}_j$ replaces the location parameter.

Algorithm 3.5 Modified ECM algorithm for multivariate skew-t mixtures

```
(i) Let $\boldsymbol{\theta}^{(0)}$ be a starting value for the maximum likelihood estimates.
(ii) E-step: Given $\boldsymbol{\theta} = \boldsymbol{\theta}^{(k)}$, we compute the needed expectations $e^{(k)}_{1ij}, e^{(k)}_{2ij}, \boldsymbol{e}^{(k)}_{3ij}, \boldsymbol{e}^{(k)}_{4ij}$ Note that all those quantities can be directly calculated using high efficient methods for evaluating t densities.
(iii) CM-steps: Maximizing yields the parameters for the $(k+1)$ step.
    a) $\omega_j^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n} \pi_{ij}^{(k)}$.
    b) $\boldsymbol{\mu}_j^{(k+1)} = \left[\sum_{i=1}^{n} \boldsymbol{x}_i e_{1ij}^{(k)} - \boldsymbol{\Lambda}_j^{(k)} \sum_{i=1}^{n} \boldsymbol{e}_{3ij}^{(k)}\right] \left(\sum_{i=1}^{n} e_{1ij}^{(k)}\right)^{-1}$.
    c) $\boldsymbol{\lambda}_j^{(k+1)} = \left[(\boldsymbol{\Sigma}_j^{(k)})^{-1} \odot \boldsymbol{B}_{1j}^{(k)}\right]^{-1} \left[(\boldsymbol{\Sigma}_j^{(k)})^{-1} \odot \boldsymbol{B}_{2j}^{(k)}\right]$.
    d) $\boldsymbol{\Sigma}_j^{(k+1)} = \frac{1}{\sum_{i=1}^{n} \pi_{ij}^{(k)}} \Big[ \sum_{i=1}^{n} e_{1ij}^{(k)} (\boldsymbol{x}_i - \boldsymbol{\mu}_j^{(k+1)})(\boldsymbol{x}_i - \boldsymbol{\mu}_j^{(k+1)})^\top + \boldsymbol{\Lambda}_j^{(k+1)} \boldsymbol{B}_{1j}^{(k)} \boldsymbol{\Lambda}_j^{(k+1)} - \boldsymbol{\Lambda}_j^{(k+1)} \boldsymbol{B}_{2j}^{(k)} - (\boldsymbol{B}_{2j}^{(k)})^\top \boldsymbol{\Lambda}_j^{(k+1)} \Big]$.
    e) $\boldsymbol{\beta}_j^{(k+1)} = \min_{\beta_j} arg \|\boldsymbol{Z}\boldsymbol{\beta}_j - \boldsymbol{\mu}_j^{(k+1)}\|$, where $\boldsymbol{Z}$ is the matrix of covariates.
    f) $\nu_j^{(k+1)} = \underset{\nu_j}{solve}\left[ \ln\left(\frac{\nu_j}{2}\right) - \psi\left(\frac{\nu_j}{2}\right) + 1 + \frac{1}{\sum_{i=1}^{n} \pi_{ij}^{(k)}} \left[\sum_{i=1}^{n} e_{2ij}^{(k)} - \sum_{i=1}^{n} e_{1ij}^{(k)}\right] = 0 \right]$,
       where $\psi(x)$ denotes the digamma function which is equal to $\frac{d}{dx}\ln\Gamma(x)$, where $\boldsymbol{B}_{1j}^{(k)} = \sum_{i=1}^{n} \boldsymbol{e}_{4ij}^{(k)}$ and $\boldsymbol{B}_{2j}^{(k)} = \sum_{i=1}^{n} \boldsymbol{e}_{3ij}^{(k)} (\boldsymbol{x}_i - \boldsymbol{\mu}_j^{(k+1)})^\top$.
(iv) Repeat step (ii) and (ii) until $|l_\infty^{(k+1)} - l^{(k)}| < \epsilon$.
```

The EM algorithm maximizes the likelihood function of the skew-t distribution with covariates. Note, since the $\boldsymbol{Z}\boldsymbol{\beta}_j$ replaces the location parameter $\boldsymbol{\mu}_j$ the error term of the regression does not have zero mean. The mean of the skew-t distribution consists of both the location and the skewness parameter.

The skew-t regression gives more flexibility and estimates the coefficients while maximizing the regarding likelihood function.

3.2.2 Profile likelihood

One aspect of assessing properties of a distribution is the profile likelihood. The profile likelihood shows the change of the likelihood function that is induced by variations of the input parameters. For univariate skew-t distributions, this means that the parameter location, scale, shape, and degrees of freedom can be varied. In the following example, 500 observations of skew-t data are simulated having the location of 1.0, the scale of 5.0, the shape of 5.0 and degrees of freedom of 10. The profile likelihood is generated during the process of fitting the appropriate model on this data.

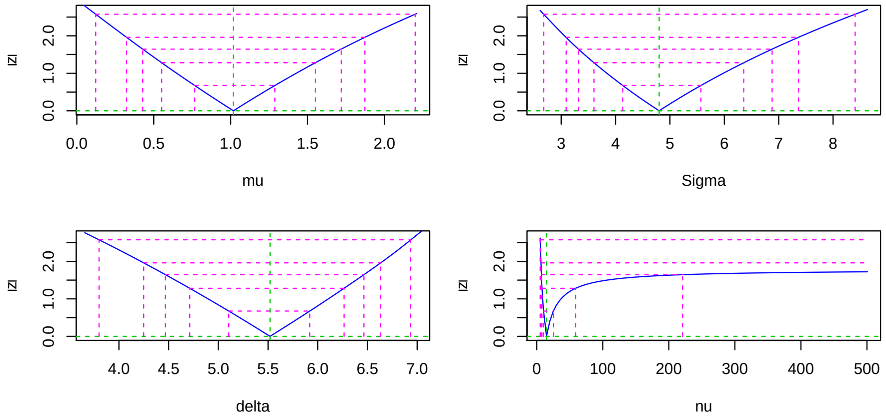


Figure 8 Profile of the log-likelihood function for the different parameters of a univariate skew t distribution. The y-axis of the figure shows $|Z|$, the absolute value of the log-likelihood and the x-axis the value of the corresponding parameters mu, Sigma, delta, and nu, respectively.

In Figure 8, the profile likelihood is presented. The minimum of the graph shows the parameter estimates. For the location parameter, the estimate exactly matches the true value from of the skew-t distribution, which was used to simulate data. For the scale and shape parameter, a slight difference to the true values can be observed. We see that the change of the likelihood is influenced similarly by these three parameters. Changing the location parameter by one unit has a similar impact as changing the scale or shape parameter. The behavior of the parameter degrees of freedom is different from this. As we already can see based on the scale, the impact of changing the degrees of freedom by one unit is rather low. Here the change of about 50 units has a similar effect on the likelihood as one unit for the other parameters. So, we expect that the true value of degrees of freedom is hard to estimate because the likelihood function is flat in the direction of this parameter. The estimated value of 14.3 is close to the true parameter, but the absolute difference is more prominent than for the other parameters.

The profile likelihood shows that for the estimates for the parameters location, scale, and shape can be quite precise. Already a small change in the parameter results in a change of the likelihood function. For the degrees of freedom, the likelihood is quite flat. For degrees of freedom higher than about 200, there is no change at all in the likelihood function.

3.2.3 Simulation study

The EM algorithm is implemented in the R package `fitmixst4`. All intensive computing functions are coded in C++ to achieve better speed. For each iteration of the EM algorithm, every data point is used. Thus, the order of the algorithm is $O(iter \cdot n)$ with $iter$ being the number of iterations and n being the number of observed data points for the univariate skew-t distribution. For a mixture or higher dimensions, the order is only growing by multiplying it with the dimension p, or the number of groups g. For the C++ coding, the linear algebra package Armadillo as well as the gnu scientific library GSL has been used to get high-performance matrix multiplications and optimizer. The C++ Code is embedded in the R package using the R package `Rcpp`. The package allows interacting with C++ code efficiently. The guessing of the initial values, input checks, and so forth are implemented in R. For faster calculation OpenMP is used to parallelize the loops in the C++ code. As a result, the calculations are almost four times faster using an Intel i7 with four cores.

The simulation study is done to verify how well the EM algorithm is finding the original parameters. So, in the first step, we draw random numbers of a skew-t distribution with a known parameter set (μ, σ^2, δ, ν) for different sample sizes. Afterward, we estimate the parameters ($\hat{\mu}$, $\hat{\sigma}^2$, $\hat{\delta}$, $\hat{\nu}$) using the EM algorithm. This procedure was done 100 times. Hence, we have 100 estimates which we are comparing to the real parameters of the distribution. The absolute bias and the mean squared error (MSE) is calculated for each parameter of the distribution. The formula for the absolute bias is

$$\frac{1}{B}\sum_{i=1}^{B}|\hat{\theta}_i - \theta|,$$

where B is the number of replications (in our case 100), θ being the true parameter, $\hat{\theta}_i$ for $i = 1, \ldots, B$ the estimated parameter of the ith replicate. Similar to the bias the MSE is defined as

$$\frac{1}{B}\sum_{i=1}^{B}(\hat{\theta}_i - \theta)^2,$$

where B is the number of replications (in our case 100), θ being the true parameter, $\hat{\theta}_i$ for $i = 1, \ldots, B$ the estimated parameter of the ith replicate.

Table 1 shows the bias and MSE for different parameters of the skew-t distribution. Also, the average number of observations and the average time in seconds is reported. The calculations were done on a Laptop with Intel Core i7 2.66ghz. The iteration maximum is set to $1,000$ iterations. The relative error, which is calculated based on the Aitken acceleration, is fixed to 10^{-3}.

In Table 1, we can see that the identification of the parameter ν is hard for small sample sizes. Looking at the profile likelihood of the skew-t distributions shows that changing ν by 10 has a minimal impact on the likelihood.

In the simulation study, we kept μ unchanged since it is only a shift parameter for the distribution and therefore, the shape of the distribution is not changing for different μ.

The scale parameter σ^2 is connected to the variance of the distribution. This connection can be seen from the analytical formula of the variance for skew-t mixtures, where the variance can be calculated from the parameters λ, σ^2, and ν. We expect that for higher values of σ^2 and therefore, a larger variance in the simulated data, it might be harder for the algorithm to detect the correct parameters. In the simulation we were using $\sigma^2 = \{1, 5\}$.

The skewness parameter λ was set in the simulation to $\lambda = \{0, 3, 10\}$ to assess the impact of this parameter on the fitting algorithm. Only positive λ were used because of symmetry reasons. For this distribution, it is crucial to estimate the skewness correctly since analytical mean, as well as variance, depend on it.

The likelihood of fitting the skew-t distribution is flat for the parameter ν. A small change in the parameter ν almost does not have any influence on the likelihood of the distribution. The parameter ν is also influencing the variance and was set to $\nu = \{6, 20, 50\}$ in this simulation study.

The simulation results show that with m observations reduce more observations, the absolute bias, and the MSE for each of the parameters.

The time for the algorithm to converge is increasing for a larger sample size. For example, in simulation with the setting $\mu = 0$, $\lambda = 0$, $\sigma^2 = 5$ and $\nu = 6$ the time is increasing from an average of 0.114 for 50 observations to 2.275 for $1,000$ observation. In this setup, the number of iterations until convergence was similar for the different number of observations ranging from 345 to 455 iterations.

For both scenarios with $\sigma^2 = 1$ and $\sigma^2 = 5$, the results are quite similar. It seems that the parameter σ^2 is always the parameter, which is closest to the real parameter. For higher values of λ, and ν, the accuracy of the estimates for σ^2 is getting lower. This behavior seems to be reasonable because the variance in the data is getting larger for higher values of λ and ν.

3.3 MCMC algorithm

This chapter introduces and discusses the first widely used sampling scheme for constructing a Markov chain with pre-specified limiting distribution π. It was first used for approximately sampling from the Gibbs distribution used in image analysis. Geman and Geman (1984) discuss this problem for several sampling schemes, and Gelfand and Smith (1990) were the first to point out to the statistical community at large that this sampling scheme could be used for other distributions as the Gibbs distribution. Gelfand and Smith (1990) also compared the Gibbs sampling scheme with the data augmentation algorithm and sampling-importance resampling.

Another method to calculate estimates of univariate or multivariate skew-t mixtures is the MCMC algorithm. Historically the MCMC method has been used in many fields and is becoming more popular. With programs such as WinBugs or JAGS at hand, it is easy to implement all kind of models using MCMC. Furthermore, the priors can be used to incorporate expert knowledge or other information in the model estimation. This greatest strength efficiently incorporating external information in the model estimation is, at the same time, the greatest problem. As shown in many publications, different prior distributions can lead to different results.

In the following section, we briefly show the implementation of the MCMC algorithm for skew-t mixtures in an existing framework such as WinBugs, OpenBugs, Stan or JAGS.

The main idea of the implementation of skew-t mixtures is the hierarchical representation in section 3.4. The hierarchical representation yields already all information that is necessary to implement the Bayesian approach for skew-t distributions. The hierarchical representation shows already the distribution of several random variables conditional on data and mixture components. Thus, this representation enables us to write down the MCMC algorithm to get posterior distributions of the mixture parameters.

First, we will present an example of a two-component mixture of normal distributions to explain the idea. The step to the more complicated skew-t distributions is rather small, but the idea behind is the same.

The two-component mixture is defined as

$$f^*(X) = pf_1(X) + (1-p)f_2(X),$$

where $0 \leq p \leq 1$ is the mixing parameter, and f_1 and f_2 are normal densities.

We assume that the densities f_1 and f_2 are entirely specified, and the problem is to estimate p. A given sample of data $z_1, z_2, \dots, z_n$ has been collected and is assumed to follow the two-component mixture distribution. A Bayesian analysis will be performed that specifies a prior distribution $\pi(p) = U(0,1)$ for the mixing parameter p. The goal is to draw samples of p from the posterior distribution.

The hierarchical Bayesian specification contains two crucial stages. The first stage is the specification of the likelihood, which in our case $z_1, z_2, \dots, z_n \overset{iid}{\sim} f^*$. The second stage is the definition of the prior distribution for the mixing parameter $p \sim \pi(p) = U(0,1)$. With both parts, we can now write down the posterior distribution as

$$\begin{aligned} f(p\,|z_1, z_2, \dots, z_n) &= \frac{f(z_1, z_2, \dots, z_n \mid p)\pi(p)}{\int f(z_1, z_2, \dots, z_n \mid p)\pi(p)dp} \propto f(z_1, z_2, \dots, z_n \mid p)\pi(p) \\ &= \prod_{i=1}^{n} (pf_1(z_i) + (1-p)f_2(z_i)) \end{aligned}$$

The most obvious choice is the $Beta(a,b)$ distribution $g(p) \propto p^{a-1}(1-p)^{b-1}$. Note, it is important that the $g(\cdot)$ is supported on the set of valid probabilities $p \in (0,1)$. Without prior information one possibility would be $Beta(1,1) = U(0,1)$, with $g(p) = 1$. Using the proposal density $g(p) \propto p^{a-1}(1-p)^{b-1}$ the acceptance probability for a new candidate p given the current sample point p_t is

$$\alpha(p_t, p) = \min\left(1, \frac{f(p)g(p_t)}{f(p_t)g(p)}\right) = \min\left(1, \frac{\prod_{i=1}^{n}(pf_1(z_i) + (1-p)f_2(z_i))p^{a-1}(1-p)^{b-1}}{\prod_{i=1}^{n}(p_tf_1(z_i) + (1-p_t)f_2(z_i))p^{a-1}(1-p_t)^{b-1}}\right)$$

Now we can sample from the posterior distribution to get an estimate for the mixing parameter. In this example, the mean, as well as the variance of both normal components, is known and only the mixing parameter is estimated from the posterior distribution. The sampling scheme above is called independence sampler and is a special case of the Metropolis-Hastings sampler.

One specific MCMC algorithm, which has proven to be very useful in practice is the Gibbs sampler. The Gibbs sampler is an algorithm, which draws samples from the joint probability distribution in cases this probability is not known, but all the conditional distributions of each parameter are known. Suppose the vector of interest $\theta = (\theta_1, \dots, \theta_d)$ consists of d subvectors θ_d, for $k = 1, \dots, d$. In each iteration, the Gibbs sampler steps through subvectors of θ_d and drawing samples of each subset conditional on all the other subvectors. So, at time t, d steps are performed to get a sample from θ. In each iteration, each θ_d^t is sampled conditional on all other subvectors of θ:

$$p(\theta_d^t \mid \theta_{-k}^{t-1}, y),$$

where θ_{-k}^{t-1} represents all the other subvectors of θ at the current time. This means that for all θ_j for $j < k$ draws for iteration t already exists, for $j > k$ the draws from iteration $t-1$ are used:

$$\theta_{-k}^{t-1} = (\theta_1^t, \dots, \theta_{k-1}^t, \theta_{k+1}^{t-1}, \dots, \theta_d^{t-1}).$$

Following this approach leads to the well-known Gibbs sampler.

Definition 3.13 (Gibbs sampler algorithm)**.**

(i) Set the initial counter to $t = 1$ and determine initial values $\theta^{(0)} = (\theta_1^{(0)}, \dots, \theta_d^{(0)})$

(ii) Obtain a new value for $\theta^{(t)} = (\theta_1^{(t)}, \dots, \theta_d^{(t)})$ through successive generation of values

$$\begin{aligned}
&\theta_1^{(t)} \sim \pi(\theta_1 \mid \theta_2^{(t-1)}, \dots, \theta_d^{(t-1)}) \\
&\theta_2^{(t)} \sim \pi(\theta_2 \mid \theta_1^{(t)}, \theta_3^{(t-1)}, \dots, \theta_d^{(t-1)}) \\
&\vdots \\
&\theta_d^{(t)} \sim \pi(\theta_d \mid \theta_1^{(t)}, , \dots, \theta_d^{(t)})
\end{aligned}$$

(iii) Change counter t to $t + 1$ and return to step 2 until convergence is reached.

Analytically deriving the posterior distribution is easy for simple examples. Fortunately, there are various software implementations available that simplify the specification of Bayesian models. Therefore, we will not derive the posterior distribution analytically for mixtures of skew-t distributions. Instead, we will show how to implement this hierarchical model to one of the commonly used software packages.

Let $x = (x_1, \ldots, x_n)$ be the given observations that have been collected. The data are assumed to follow a skew-t distribution with hierarchical representation defined in Definition 2.11 with unknown skewness parameter λ, location parameter μ, variance parameter σ^2 and degrees of freedom ν. The software was used to automatically calculate the conditionals, which are necessary for the Gibbs sampler. With the assumptions of the above-mentioned skew-t distribution, these are the necessary conditional distributions

$$\begin{aligned}
&p(\lambda \mid \mu, \sigma^2, \nu, y), \\
&p(\mu \mid \lambda, \sigma^2, \nu, y), \\
&p(\sigma^2 \mid \lambda, \mu, \nu, y), \\
&p(\nu \mid \lambda, \mu, \sigma^2, y).
\end{aligned}$$

To use JAGS, the model has to be written in the so-called JAGS language. Writing the hierarchical definition of the skew-t distribution in JAGS yields to following algorithm:

Algorithm 3.6 Implementation of skew-t distribution in JAGS

```
model {
for (i in 1:length(x)) {
x[i] ~ dnorm(alpha[i], beta[i])

alpha[i] <- mu + lambda * z[i]
beta[i] <- g[i] * tau
z[i] ~ dnorm(0, g[i]) T(0,)
g[i] ~ dgamma(nu / 2, nu / 2)
}
mu ~ dnorm(0.0, 0.01)
lambda ~ dnorm(0.0, 0.01)
nu ~ dexp(0.1) T(2,)
tau ~ dgamma(0.01, 0.01)
sigma2 <- 1/tau
}
```

Note that in JAGS the normal distribution is defined by the precision and not by the variance. The priors in the last part of the code are non-informative and can be modified accordingly. For the algorithm, we use non-informative prior distributions for the parameters of the skew-t distribution. For location and skewness parameter, a normal distribution with zero mean and standard deviation 100 is used. For the parameter ν, a truncated exponential distribution is used to ensure that $\nu \geq 2$. For the inverse variance, a wide gamma distribution with both parameters being 0.01 is used.

		Absolute bias				MSE					
Parameter	n	$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{\lambda}$	$\hat{\nu}$	$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{\lambda}$	$\hat{\nu}$	time	iterations
$\mu = 0,\ \sigma^2 = 1,$	50	0.709	0.350	0.882	29.809	0.629	0.178	1.003	2066.529	0.101	292.140
	100	0.576	0.233	0.671	16.403	0.433	0.084	0.621	972.487	0.243	370.050
$\lambda = 0,\ \nu = 6$	500	0.275	0.096	0.323	2.422	0.119	0.014	0.169	32.851	1.374	455.960
	1000	0.185	0.072	0.216	1.084	0.052	0.008	0.070	2.169	2.467	429.360
$\mu = 0,\ \sigma^2 = 1,$	50	0.748	0.401	0.908	43.148	0.685	0.210	0.983	2704.030	0.108	383.970
	100	0.693	0.338	0.857	32.581	0.582	0.154	0.877	1663.673	0.212	364.480
$\lambda = 0,\ \nu = 20$	500	0.441	0.139	0.536	20.492	0.232	0.028	0.350	858.792	1.739	669.490
	1000	0.320	0.092	0.393	16.246	0.141	0.012	0.211	624.945	4.517	897.500
$\mu = 0,\ \sigma^2 = 1,$	50	0.825	0.467	1.005	32.098	0.797	0.285	1.175	1387.658	0.093	302.390
	100	0.661	0.315	0.814	33.577	0.534	0.148	0.779	1457.628	0.202	381.090
$\lambda = 0,\ \nu = 50$	500	0.449	0.156	0.555	23.572	0.236	0.035	0.359	851.596	1.526	624.550
	1000	0.425	0.131	0.528	21.134	0.203	0.023	0.312	711.859	3.524	738.590
$\mu = 0,\ \sigma^2 = 1,$	50	0.399	0.632	0.625	28.466	0.275	0.708	0.600	1983.638	0.165	492.150
	100	0.313	0.490	0.543	10.961	0.179	0.440	0.466	426.404	0.299	378.530
$\lambda = 3,\ \nu = 6$	500	0.147	0.194	0.269	2.426	0.032	0.052	0.109	24.876	1.633	347.820
	1000	0.099	0.121	0.164	1.240	0.015	0.023	0.042	3.908	3.555	376.290
$\mu = 0,\ \sigma^2 = 1,$	50	0.514	0.566	0.710	44.320	0.513	0.554	0.957	2944.854	0.197	701.160
	100	0.407	0.451	0.516	36.525	0.237	0.348	0.433	2196.771	0.281	395.180
$\lambda = 3,\ \nu = 20$	500	0.143	0.227	0.200	18.246	0.032	0.081	0.063	810.033	1.674	426.080
	1000	0.088	0.117	0.137	14.665	0.012	0.022	0.030	463.923	3.667	462.790
$\mu = 0,\ \sigma^2 = 1,$	50	0.476	0.637	0.620	31.661	0.444	0.665	0.748	1302.812	0.164	548.710
	100	0.329	0.424	0.429	37.326	0.260	0.363	0.426	1659.210	0.231	334.530
$\lambda = 3,\ \nu = 50$	500	0.116	0.149	0.164	24.314	0.022	0.035	0.045	819.820	1.551	458.520
	1000	0.093	0.143	0.125	21.515	0.014	0.032	0.025	712.014	3.408	515.440

Table 1 Results from a simulation study fitting a skew-t distribution are presented. There are n observations sampled from a skew-t distribution with known parameters. The absolute bias as well as the MSE, the time in minutes and the number of iterations are shown.

		Absolute bias				MSE					
Parameter	n	$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{\lambda}$	$\hat{\nu}$	$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{\lambda}$	$\hat{\nu}$	time	iterations
$\mu = 0,\ \sigma^2 = 1,$	50	0.513	1.042	1.419	24.537	0.496	2.055	3.146	1778.869	0.955	3354.380
	100	0.427	0.796	1.214	14.585	0.291	0.969	1.989	922.295	1.404	2358.910
$\lambda = 10,\ \nu = 6$	500	0.189	0.310	0.549	1.779	0.054	0.157	0.435	8.407	3.027	790.330
	1000	0.120	0.206	0.335	1.068	0.022	0.062	0.166	2.537	6.112	801.460
$\mu = 0,\ \sigma^2 = 1,$	50	0.565	1.203	1.206	33.871	0.703	5.866	2.235	2160.291	0.670	2451.580
	100	0.415	0.738	0.863	32.603	0.271	0.944	1.317	2063.017	0.596	962.150
$\lambda = 10,\ \nu = 20$	500	0.182	0.320	0.413	16.739	0.051	0.170	0.276	728.272	2.366	659.770
	1000	0.119	0.178	0.289	10.536	0.023	0.055	0.128	324.274	4.184	538.030
$\mu = 0,\ \sigma^2 = 1,$	50	0.845	1.639	1.362	40.387	1.328	7.659	3.129	1829.533	0.379	1451.900
	100	0.397	0.747	0.819	35.780	0.248	0.867	1.085	1525.179	0.657	1092.850
$\lambda = 10,\ \nu = 50$	500	0.140	0.264	0.306	29.082	0.031	0.110	0.155	1097.947	2.188	606.050
	1000	0.123	0.204	0.258	25.592	0.025	0.069	0.104	833.198	4.443	620.300
$\mu = 0,\ \sigma^2 = 5,$	50	1.543	1.792	1.964	34.418	3.000	4.792	4.934	2547.063	0.114	345.580
	100	1.247	1.122	1.506	18.874	2.149	2.045	3.253	1123.145	0.221	364.330
$\lambda = 0,\ \nu = 6$	500	0.585	0.497	0.687	2.236	0.538	0.372	0.746	22.796	1.310	455.180
	1000	0.377	0.287	0.430	0.960	0.223	0.127	0.298	1.750	2.275	410.480
$\mu = 0,\ \sigma^2 = 5,$	50	1.673	2.045	2.066	39.054	3.264	5.626	5.059	2300.744	0.085	284.270
	100	1.342	1.508	1.601	32.813	2.314	3.123	3.299	1792.758	0.226	402.480
$\lambda = 0,\ \nu = 20$	500	0.928	0.691	1.140	21.489	1.103	0.698	1.683	851.375	1.787	706.430
	1000	0.768	0.447	0.930	14.768	0.751	0.306	1.115	426.356	4.360	889.260
$\mu = 0,\ \sigma^2 = 5,$	50	1.701	2.105	2.077	32.077	3.520	6.225	5.200	1401.977	0.108	408.120
	100	1.376	1.623	1.710	28.932	2.326	3.690	3.554	1198.315	0.214	402.000
$\lambda = 0,\ \nu = 50$	500	1.027	0.784	1.249	23.889	1.189	0.835	1.746	886.933	1.589	653.750
	1000	0.957	0.655	1.183	20.589	1.063	0.582	1.614	681.723	3.944	838.030

		Absolute bias				MSE					
Parameter	n	$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{\lambda}$	$\hat{\nu}$	$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{\lambda}$	$\hat{\nu}$	time	iterations
$\mu = 0,\ \sigma^2 = 5,$ $\lambda = 3,\ \nu = 6$	50	1.382	1.967	1.822	33.014	3.862	6.490	6.255	2260.921	0.110	297.320
	100	0.832	1.610	1.116	14.028	1.330	4.223	2.198	713.659	0.230	300.810
	500	0.326	0.699	0.455	2.038	0.165	0.802	0.318	27.204	1.438	335.920
	1000	0.234	0.524	0.320	1.032	0.090	0.466	0.157	2.302	3.374	392.120
$\mu = 0,\ \sigma^2 = 5,$ $\lambda = 3,\ \nu = 20$	50	1.557	1.980	1.980	48.463	4.854	5.480	7.592	3203.703	0.112	388.680
	100	1.035	1.556	1.311	33.229	2.293	3.610	3.503	1865.082	0.229	370.490
	500	0.384	0.802	0.479	18.599	0.354	0.983	0.519	633.163	1.765	570.270
	1000	0.251	0.560	0.319	19.384	0.097	0.478	0.156	733.501	3.253	549.790
$\mu = 0,\ \sigma^2 = 5,$ $\lambda = 3,\ \nu = 50$	50	1.845	1.929	2.241	31.483	6.562	5.466	9.459	1317.032	0.084	275.050
	100	1.168	1.565	1.422	35.373	3.030	3.418	4.432	1533.599	0.199	369.400
	500	0.351	0.727	0.405	18.942	0.285	0.809	0.425	605.493	1.371	508.560
	1000	0.259	0.616	0.327	18.683	0.115	0.569	0.184	607.081	3.304	622.740
$\mu = 0,\ \sigma^2 = 5,$ $\lambda = 10,\ \nu = 6$	50	1.073	3.631	1.840	30.426	1.895	22.426	4.651	2302.011	0.304	1019.430
	100	0.594	2.729	1.203	13.534	0.558	14.892	2.177	743.320	0.287	347.750
	500	0.339	1.041	0.606	1.676	0.188	1.846	0.587	5.203	1.458	336.220
	1000	0.217	0.810	0.464	1.159	0.078	0.956	0.331	2.602	3.395	393.020
$\mu = 0,\ \sigma^2 = 5,$ $\lambda = 10,\ \nu = 20$	50	1.065	3.509	1.521	38.453	1.651	23.208	3.458	2457.647	0.143	427.760
	100	0.648	2.472	1.104	34.289	0.655	10.370	1.815	2101.314	0.311	435.750
	500	0.343	1.066	0.464	20.624	0.171	1.666	0.308	1061.194	1.856	469.370
	1000	0.194	0.867	0.360	11.094	0.061	1.085	0.200	309.389	3.182	386.390
$\mu = 0,\ \sigma^2 = 5,$ $\lambda = 10,\ \nu = 50$	50	1.233	3.973	1.663	38.210	5.038	40.793	8.381	1673.247	0.250	908.500
	100	0.650	2.430	0.941	36.208	0.748	8.873	1.477	1550.161	0.275	373.540
	500	0.301	1.027	0.441	26.442	0.141	1.586	0.312	917.908	1.833	502.420
	1000	0.208	0.780	0.379	24.350	0.080	0.987	0.234	788.751	4.042	551.680

4 Software

4.1 R package fitmixst4

The R package `fitmixst4` (Höfler 2019), publicly available at http://R-Forge.R-project.org, was developed to fit skew-t mixtures and implements the algorithm described in chapter 3.1. The package provides highly efficient functions implemented in C++ embedded in a simple to use R package. In order to utilize the power of modern multi-core processors, the functions are parallelized using Open Multi-Processing (OpenMP) (Dagum and Menon 1998). OpenMP is a C++ library that enables users to perform multi-processing with shared memory. The R package `roxygen2` (Wickham et al. 2018) was used for the development of the R package to generate the documentation of function. The functions in this package use S4 classes. In R, there are different object orientated systems implemented. While S3 is the older and more common way of implementing objects in R, S4 is closer to other object-orientated programming languages. For S3 classes, generic functions are called, which decides the used method, e.g., `lm.summary` for the result of a linear model. S4 formalizes the same idea. That means S4 classes have a more rigid class definition specifying all elements of an object, the so-called slots. The slots of an object can be accessed using the symbol, e.g., `out@logLik`. One advantage is that the validity checks can be carried out when calling a new instance of the class. Thus, with S4 classes, we can restrict entries to be in a pre-specified range, which is necessary for variables such as the degrees of freedom. The usage of S4 classes simplifies the checks for valid inputs at the beginning of the EM algorithm. The core function of the package is the `fitmixst`, which calls C++ routines and executes the EM algorithm on the data. In the most straightforward case, the user needs to enter a vector of data and the number of components of the mixture distribution. Additionally, other starting options, as well as starting values, can be used as input parameters of the function. Besides the efficiently implemented call of C++ using the explicit form of the moments, it is possible to use legacy R code with either calculating the moments with the explicit expressions or with MCMC integration. Furthermore, a verbose option is available to trace the steps of the EM algorithm, e.g., for debugging. The `rmixst` function can be used to generate multivariate mixtures of skew-t distributed observations.

Furthermore, generic functions `plot`, `points`, `print`, `predict` and `logLik` are implemented, which allow the user to apply a wide range of R tools. The `stderror` function is used to derive standard errors of the estimates. The Fisher information matrix is the basis for the derivation and can be calculated in each step of the EM algorithm. The internal function `truncatedt` proved useful, allowing first and second moments of truncated multivariate t distributions to be derived quickly.

4.1.1 Installation

The R package `fitmixst4` can be installed from R-Forge using the following command.

```r
install.packages("fitmixst4",
                 repos="http://R-Forge.R-project.org")
```

For compiling the package from source, the R packages `Rcpp`, `RcppGSL` and `RcppArmadillo` are required. To make use of the `RcppGSL` package when compiling the C++ libraries, it is required to install the GNU Scientific Library (GSL) (Galassi et al. 2009).

4.1.2 Key data structure

Different input parameters used in the package require different structures. The following list summarizes essential structures.

- Data: Y matrix $n \times p$, where n is the number of observations and p the number of dimensions.
- Parameters: a list containing the parameters of the skew-t mixture or an S4 object of the class mixpara.
- Result: an S4 object of the class fitmixout containing the class mixpara with the fitted parameters, log-likelihood, observations, empirical covariance and the posterior probability of each data point belonging to the kth mixture component.

Below are the elements, the so-called slots from introduced S4 classes.

```r
getSlots("fitmixout")
```

```
##              y              X           beta          resid
##       "matrix"       "matrix"       "matrix"       "matrix"
##           para         logLik           iter           call
##      "mixpara"      "numeric"      "numeric"         "call"
##         empcov     posteriori likConvergence
##       "matrix"       "matrix"      "numeric"
```

The statement above shows all slots of the S4 class fitmixout. For each slot the type is predefined, and the user can only enter this data type. This class is the output object that is returned by the `fitmixst` function. In each iteration of the fitting process, the algorithm adds the current numerical result to the respective slots y, X, beta, and resid of the fitmixout object. The matrices y and X contain the data matrix and the covariate matrix, respectively. The matrix beta holds the values of the regression coefficient and resid the residuals of the skew-t regression. The slot mixpara is itself an S4 class (see description below) and holds the information of all estimated values generated during the fitting process of the skew-t mixture such as an estimate for location parameter or logLik that stores the numeric value of the log-likelihood at the current iteration given in the iter slot. The call slot stores the information about the function call, which enables the user to see the setting of the call. The parameter empcov is the empirical covariance matrix of the skew-t mixture parameters. The parameter posteriori is a matrix showing the posterior probability of each data point belonging to the kth component. The last slot is a vector with the log-likelihood value in each iteration to see the progress of the algorithm and check the convergence.

```
getSlots("mixpara")
```

```
##       pro        mu     Sigma     delta        nu         p         g
## "numeric"    "list"    "list"    "list"    "list" "integer" "integer"
```

The mixpara S4 class consists of seven slots with the information to define skew-t mixture distributions. The pro vector is numeric and represents the mixture proportions summing up to one. The parameters mu, Sigma, delta, and nu are list objects, where each list element represents one mixture component. The parameter mu corresponds to the location vector, Sigma to the scale matrix, delta to the skewness parameter, and nu to the degrees of freedom. The parameters p and g are integers, which define the dimension and the number of components, respectively.

4.1.3 Example

The `fitmixst4` package fits skew-t mixture models to data. The package requires a data matrix and the number of mixture components as minimal input. In the R package, the default iteration maximum is set to $1,000$ and the model components are initialized using k-means (Lloyd 1957; MacQueen et al. 1967). The algorithm stops iterating either after reaching the maximum iterations or if the relative error of the likelihood function is smaller than a certain predefined threshold, with the default being 10^{-3}.

The code below illustrates a simple example of simulating skew-t distributed data with the `rmixst` function, followed by use of the `fitmixst` function to determine the parameters. In the first step, the mixture parameter pro is set to one since we only have one component. The parameter mu corresponds to the location parameter, which we set to zero for this example. The scale parameter is set to three, the skewness parameter delta to zero, and the degrees of freedom nu is 30. This parametrization implies that the generated data should look quite like a normal distribution. In the next line, the list object para defines all parameters of the skew-t mixture distribution. The function `rmixst` generates a data vector y of length 1000 with the help of the para list object defining the parameterization of the skew-t mixture. The `fitmixst` function now fits the skew-t mixture to the data y. The number of components was set to $g = 1$ and the method to initialize the components is set to $method = "kmeans"$, which does not have an influence in this setting with only one component. Requesting out in the R console calls the generic show function of this object and presents the output S4 class fitmixout as specified in the R package.

```
        pro = 1; mu=10; Sigma=0.5; delta=-2; nu=16;
        para = list(pro=pro,mu=mu,Sigma=Sigma,delta=delta,nu=nu)
        y <- rmixst(1000,para)
        out <- fitmixst(y,g=1,method="kmeans")
        out
```

```
## Fitmixst
## fitmixst(y, g = 1, method = "kmeans")
## Number of iteration: 388
##  Log Likelhood: -1836.6486
##
## Mixture parameter:
## pro
## [1] 1
## mu
## [[1]]
## [1] 10.15715
##
## Sigma
## [[1]]
## [1] 0.4595311
##
## delta
## [[1]]
## [1] -2.188008
##
## nu
## [[1]]
## [1] 15.35657
```

In the output, we can see the number of iterations, the log-likelihood as well as the mixture parameters. It shows in the first few lines the call of the function followed by the number of iterations done until the stopping criterion was fulfilled. In this case, the algorithm stopped after 388 iterations with a log-likelihood of -1836.6486. The mixture parameter pro is equal to one as by definition with only one component. The estimated value of mu is at 10.15715 close to the true parameter of 10. Similarly, the scale parameter estimated 0.4595311 is close to the true value of 0.5. In the data generation step, we set the skewness to -2. Hence, the estimated parameter for delta is at -2.188008 quite close to -2. The degrees of freedom nu has a value of 15.35657, which is almost 16 the value for generating the data. Overall, we see that the estimated parameters correspond quite well to the true parameters that were used to simulate the data. If we want to assess the convergence of the parameters, we can graph the log-likelihood values per iteration as follows. We see that after about 50 iterations the convergence levels. To reach the stopping criterion, which is defined relative to three estimates of the log-likelihood, 388 iterations were required.

```
plot(out@likConvergence)
```

Figure 9 shows the log-likelihood values for each iteration step. It can be observed that the first few iterations of the log-likelihood increases faster than during later iteration steps. Finally, the log-likelihood values are converging when the optimal parameters were estimated.

Generic functions, such as `logLik`, `plot`, `points`, `print`, and `BIC` extracting the log-likelihood,

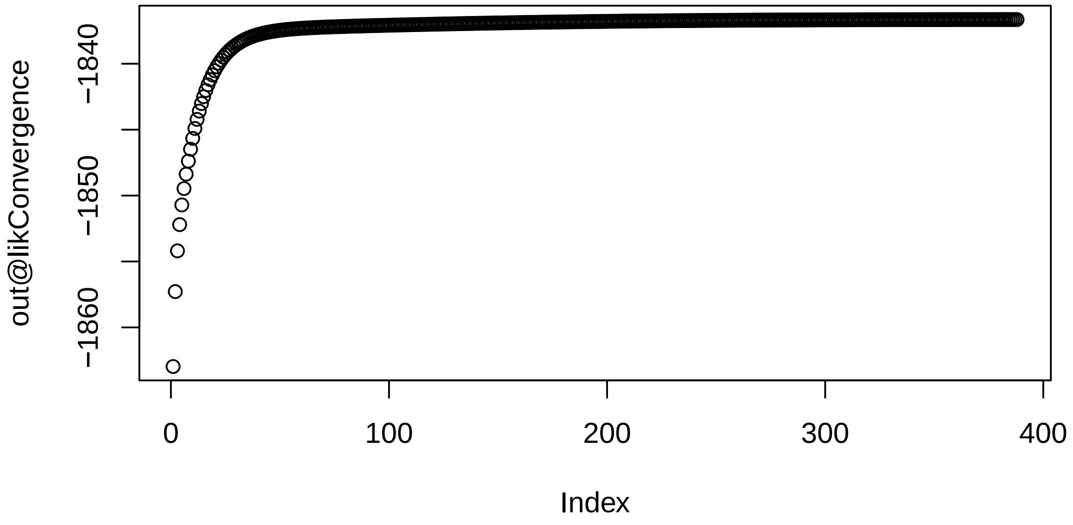


Figure 9 Log-likelihood values are shown on the y-axis for each iteration. The index of the iteration is presented on the x-axis.

generating a plot, adding points to a plot, printing the results or showing the Bayesian information criterion (BIC) values can be applied to the and output applied in the usual fashion. Even if the BIC function was not explicitly defined in the R package it is possible to use it since it is a generic function relying on the logLik generic function that is defined for the S4 classes fitmixout and mixpara in this R package. For each class, generic functions need to be defined to handle the class correctly. Defining the most common generic functions, as mentioned above, opens a wide range of tools applicable to the newly created objects.

As an example, the code below generates a histogram for this case and overlays the fitted skew-t mixture that makes a visual inspection of the goodness of fit possible.

```
hist(out@y,freq = F)
plot(out,add=T,col="red")
```

In Figure 10, we see that the histogram and the fitted density are matching well. Note that the graphical representation is only available for one and two-dimensional data.

The empirical covariance matrix can be used to derive standard errors of the estimates.

```
round(out@empcov,3)
```

```
##        [,1]   [,2]   [,3]   [,4]
## [1,]  0.001  0.000 -0.001  0.086
## [2,]  0.000  0.001  0.000 -0.013
## [3,] -0.001  0.000  0.003 -0.260
## [4,]  0.086 -0.013 -0.260 39.553
```

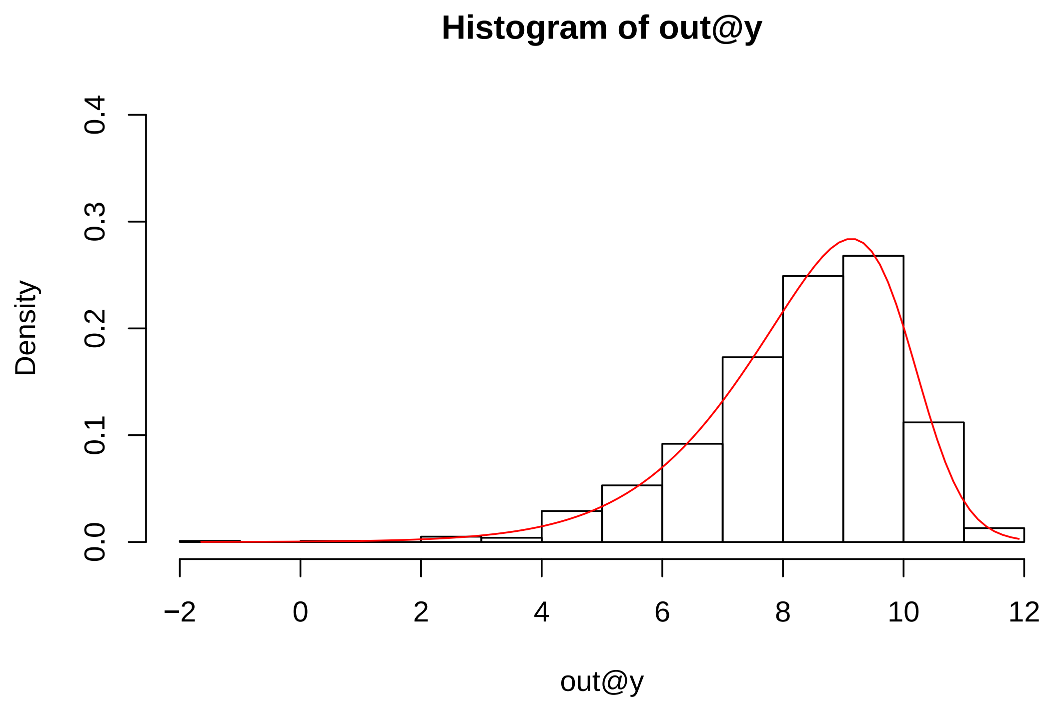


Figure 10 The histogram of the observed data with the overlaid pdf function estimated from the data with the `fitmixst4` package is shown.

The diagonal elements of the empirical covariance reflect the variance of the corresponding parameter. Hence, the variance of the estimate for mu is 0.001. Similar to the other parameters the variance for Sigma is 0.001, for delta 0.003 and nu 39.553. The other elements in the matrix reflect the covariance between these parameters. With more iterations, the variance of the parameters would shrink further since the empirical covariance depends on the number of iterations. We can see that the parameter nu has the highest variance, which means we would require more iterations for better precision of the estimates.

4.1.4 Summary

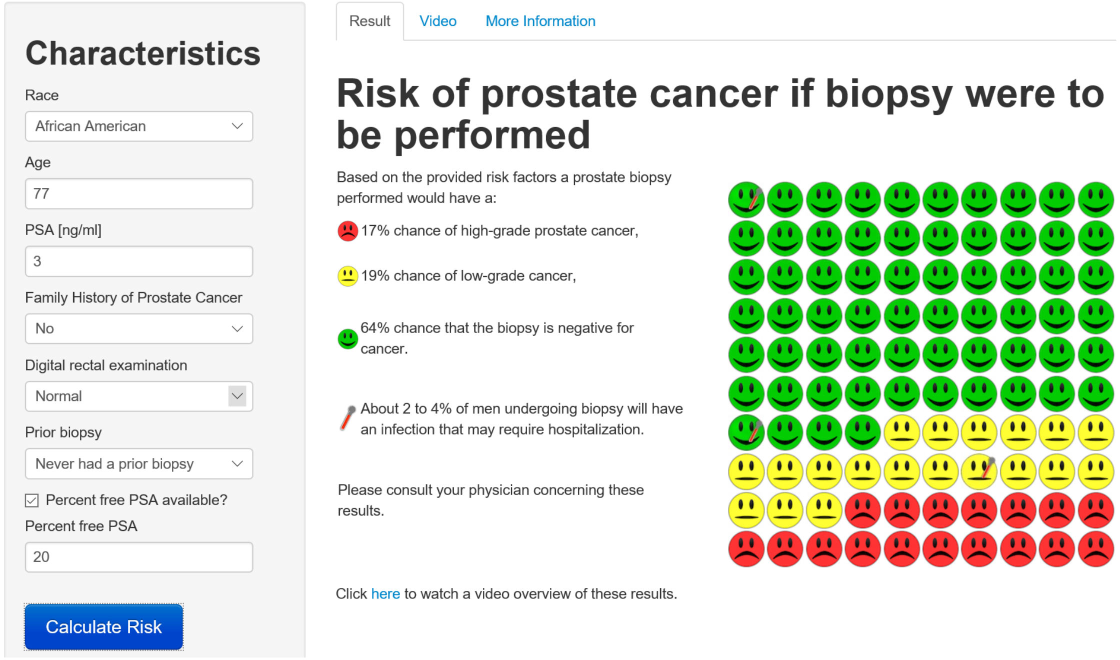
In summary, the `fitmixst4` package offers a fast and user-friendly experience fitting and visualizing multivariate skew-t distributions. Besides, the R package incorporates some convenience features. For example, if the user interrupts the `fitmixst` function, it still finishes the last iteration and saves the output for a later restart. All core routines are implemented using C++, optimizing the speed of computation.

4.2 Risk calculator web application

4.2.1 Introduction

The risk calculator web application was built using the R package `shiny` (Chang et al. 2019). The risk calculator named Prostate Cancer Prevention Trial Risk Calculator (PCPTRC) 2.0 based on a multinomial logistic model with updates are presented in Section 5.1 based on

Ankerst et al. (2014). This section covers the structure of risk calculator and how the `shiny` app was used to bring the model to an accessible online format for patients and doctors to use (see Figure 11).



Characteristics

Race: African American

Age: 77

PSA [ng/ml]: 3

Family History of Prostate Cancer: No

Digital rectal examination: Normal

Prior biopsy: Never had a prior biopsy

☑ Percent free PSA available?

Percent free PSA: 20

Risk of prostate cancer if biopsy were to be performed

Based on the provided risk factors a prostate biopsy performed would have a:

☹ 17% chance of high-grade prostate cancer,

😐 19% chance of low-grade cancer,

☺ 64% chance that the biopsy is negative for cancer.

/ About 2 to 4% of men undergoing biopsy will have an infection that may require hospitalization.

Please consult your physician concerning these results.

Click here to watch a video overview of these results.

Figure 11 The user interface of the risk calculator.

The original risk calculator was created in Java. Later it was utterly newly programmed and extended using R with the package `shiny` in August 2014. The code of the risk calculator is stored in a repository with over 150 commits, several features and development branches that were later merged into the master branch of the repository. The repository was used to keep the documentation of all changes and updates. Additionally, it enables multiple users to access and work on the risk calculator. A complete translation into Spanish, as well as a more intuitive user interface making it possible to interact with several risk calculators through one user interface, was implemented. Google analytics was added to the website to show the country of users currently accessing the risk calculator.

The R package `shiny` was designed as an alternative for building web applications that would be more tenable to statisticians compared to the typically used JavaScript programming language used by computer scientists. The online motto of shiny is "Turn your analyses into interactive web applications" (RStudio, Inc 2019) and is described as a fast-bidirectional communication between the web browser and R using web socket forms the basis of each web application. The web socket is a protocol providing communication over a single connection. The default User Interface (UI) elements utilizes Twitter Bootstrap (Twitter, Inc 2019). Twitter Bootstrap is an open source framework to create responsive web pages. With `shiny`, any web application can use the power of R and all its packages.

Usually, every `shiny` application is structured in two parts, the server.R and the ui.R. The server.R is the back end of the web application responsible for all calculations and reactive behavior, while the ui.R creates the UI. The ui.R part of the application generates all UI elements, such as input fields, buttons, and drop-down menus. Hence, every time an input changes the reactive binding invokes the calculations and generates outputs or graphs using the server.R. This enables users to create interactive applications quickly while still having the full power of R behind the scenes for computation.

Initially, the shiny web application was hosted on a shiny server installation. Later the web application was moved to the hosting of www.shinyapps.io and is available at
http://riskcalc.org/PCPTRC/.

4.2.2 Structure of the web application

To give an overview of the structure, we start with the features of the web application. Figure 12 depicts the directory tree of the `shiny` application showing the main elements.

```
Risk Calculator PCPTRC 2.0
 ├── server.R
 ├── ui.R
 ├── /data/
 │    ├── riskcalc.css
 │    ├── SNP.RData
 │    ├── english.R
 │    └── spanish.R
 ├── /www/
 │    ├── green.jpg
 │    ├── orange.jpg
 │    └── red.jpg
 ├── global.R
 └── riskcalc.R
```

Figure 12 Directory tree of the shiny PCPTRC 2.0 web application.

As stated previously, the server.R module contains the typical R programming for fitting the risk model, and the ui.R the UI elements. For this calculator, these are mainly drop-down menus and numeric fields required for input to the risk model, including prompting for the input risk factors from the user. As covered in Section 5.1 and seen in Figure 11, these include PSA, DRE, race, age, family, and prior biopsy history. Optional inclusion of the biomarker percent free PSA is available, detailed family history and single nucleotide polymorphism (SNP) can be enabled or disabled via additional tick boxes. The module ends with a Calculate risk button, which initiates the calculation of the risk based on the current input parameters.

Since the web application is available in English and as well in Spanish, the translations are stored in english.R and spanish.R, respectively. The SNP.RData file stores the names of the different SNPs required for the calculator. Per design by clinicians, smiley emoji are used to visualize the output risk and available in different colors as pictures (i.e., green.jpg). The riskcalc.css is a cascading style sheet that describes the presentation of the webpage. The global.R file loads general settings, while the riskcalc.R file contains the core algorithm deriving the risk for different parameters. The two tree nods www and data are folders that contain the respective pictures, scripts or data.

4.2.3 Application flow

The application flow is presented in Figure 13. The page initializes with a disclaimer and the option for the user to switch to Spanish. After accepting the disclaimer, the web application shows the UI of the risk calculator. Now the user can enter the required information into the UI. At this initial entry stage, the user can decide if additional input such as percent free PSA are available. After pressing the Calculate Risk button, the web application presents the resulting risks for the entered data. The results page allows the user to select other input parameters as shown in Figure 13. For example, if the input race was Caucasian, the option exists for detailed history or single-nucleotide polymorphisms (SNPs) to be included. The other input parameters remain unchanged as the new risk factors are entered, allowing for a quick recalculation.

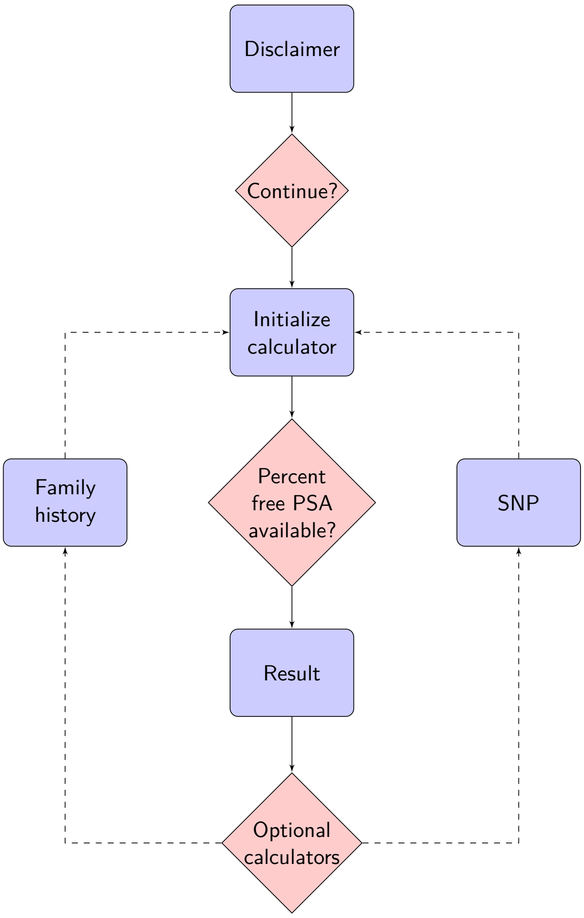


Figure 13 Schematic representation of the steps using the PCPTRC 2.0 web application.

Reactive binding is created on the Calculate risk button to restrict the calculation of the risk to the moment when the user clicks the button. Once an event invokes the binding, the risk calculation function is called and returns the risk for the three groups: low-grade, high-grade, and no cancer. Based on the results, the smiley graphic is creating using R, combining the right number of emoji for each color. The numeric result, as well as the graphical representation, is returned from the server.R to the ui.R.

Conditional panels are used to allow the user to enter optional inputs, while upon entry of variables range checks with error messages prohibiting calculation for an entry being out of range. The optional Spanish interpretation makes the implementation more difficult since all input elements need to be available in two languages and still function correctly. Here we created a separate R script containing all text used in the calculator and the corresponding Spanish translation. With the help of a global variable, each drop-down menu, button text, and text on the web page is either set to Spanish or English. All reactive bindings need to foresee the possibility to handle Spanish and English that everything works correctly after switching the language. So many problems appeared during the implementation, and all issues need to be solved case by case. Furthermore, extensive testing of all options and situations was required to ensure a working risk calculator. In order to ensure that the risk calculator is always online, the application was hosted on two different servers. Hence, only the domain name system (DNS) needs to be adapted, pointing to the running server. This flexibility enables versatile and fast handling of server issues. The web application was stored in a repository as mentioned earlier and can be uploaded by several people to have quick reaction times.

4.2.4 Summary

The quite complicated calculation for prostate cancer risk is implemented in an easy to use web application that is freely available under the R open software philosophy. This web application should help users to make an informed decision together with their physician, thereby transporting scientific results more readily to the community. The R package `shiny` makes the transition from an R function to an easy to use web application possible.

5 Applications

In this chapter, we show how mixture models can be used for classification of continuous biomarkers into groups, such as diseased versus not-diseased, via density ratios. In the first example, an approach is presented on how to update existing risk prediction models by including new risk factors using external data. The work done for the publication Ankerst et al. (2014) is the basis for the following example. Parts of the paper with a more detailed elaboration of the results will be presented. The focus of this work was to update an existing risk prediction tool by incorporating new biomarkers to improve the precision of the prediction.

The second example deals with forest management, where we assess several competition indices in the context of tree mortality.

5.1 Prostate cancer prediction

5.1.1 Introduction

Prostate-specific antigen (PSA) screening is highly controversial, recommended to be offered by several organizations, but recommended against by the US Preventive Services Task Force (Moyer 2012; Carter et al. 2013). A low-grade prostate cancer diagnosis, in most cases, is not helpful to the patient as, even on active surveillance, outcomes are exceptionally-good (Klotz et al. 2009). It is most likely that the patient who benefits from a biopsy (and subsequent therapy) is the one with high-grade cancer, as was noted in the US randomized trial of observation versus surgery (Wilt et al. 2012). Thus, it is the relative frequency of the three outcomes (negative, low-grade, and high-grade cancer) that determines whether a biopsy is indicated for an individual patient. Ideally, with variables associated with risk, the optimal circumstance is the most accurate prediction of these three outcomes that the clinician seeks today. Our goal in this study was to model the utility of [-2]proPSA and percent free PSA, both being widely-available and Food and Drug Administration (FDA) registered biomarkers. To determine their efficiency as new biomarkers for the current prediction of prostate cancer risk in the updated Prostate Cancer Prevention Trial risk calculator (PCPTRC) 2.0.

5.1.2 Material and methods

The San Antonio center for Biomarkers Of Risk (SABOR) is a Clinical and Epidemiological Center of the Early Detection Research Network (EDRN) of the National Cancer Institute. SABOR comprises a 3,930-member multi-ethnic, multi-racial cohort recruited from San Antonio and South Texas from 2000 to 2012 with up to 13 years of follow-up. Subjects were initially followed annually with PSA, and digital rectal examination (DRE) performed annually. In 2010, annual DRE was eliminated, and PSA measurements changed to biannually in men with a PSA < 1.0 ng/mL, to improve focus on patients with the highest risk of cancer development. A separate EDRN reference biopsy data set was used for validation of the model; these biopsies included patients referred to Beth Israel Deaconess Medical Center, the University of Michigan, and the University of Texas Health Science Center at San Antonio for suspicion of prostate cancer, generally based on PSA or DRE findings (Sokoll et al. 2010).

Pre-biopsy patient characteristics, including PSA, percent free PSA, and [-2]proPSA, were compared between the three San Antonio Biomarkers Of Risk (SABOR) groups (no cancer, low- and high-grade cancer) using the chi-square test for categorical outcomes and Kruskal-Wallis test for continuous measures. Overlaid densities of PSA, percent free PSA, and [-2]proPSA across the three outcome groups were visually assessed. Additionally, scattergrams of percent free PSA and [-2]proPSA were overlaid across the three outcome groups. The same procedures were followed on the EDRN validation set. Density ratios comparing the conditional distributions of percent free PSA and [-2]proPSA were computed as previously performed for the original PCPTRC (Vickers et al. 2010; Ankerst et al. 2012). Joint distributions of percent free PSA and [-2]proPSA were modeled separately in the three outcome groups using the robust regression based on a normal, skew-t, and t distribution, whereas the skew-t and t distribution had six degrees of freedom. The Bayes theorem was the basis for the updating method, and the performance of the different distributions will be compared in the results. More details are presented in Section 2.3. PSA was the only statistically significant predictor and is included in all regressions. The density ratios were multiplied by the prior odds of each cancer outcome (low- or high-grade cancer compared to no cancer, respectively) provided by PCPTRC 2.0 to offer updated risks. These updated risks are depending on percent free PSA and [-2]proPSA in addition to the standard risk factors included in PCPTRC 2.0, i.e., PSA, DRE, African American race, first-degree family history of prostate cancer, age, and history of a prior negative biopsy.

Comparisons of the new risk calculator incorporating percent free PSA and [-2]proPSA to the PCPTRC 2.0 were performed using the independent biopsy data set from the EDRN using area underneath the receiver-operating characteristic curves (AUC) for predicting prostate cancer (low- and high-grade) versus no cancer and for predicting high-grade prostate cancer versus low-grade prostate cancer or no cancer. The AUC measures the discrimination power of a risk that is calculated as the area under the receiver operating curve (ROC) (Murphy 1973; Agresti 2007). The ROC plots the true positive rate (sensitivity) versus the false positive rate (1-specificity) for all possible thresholds of a risk tool that could be used to classify a tree as likely to experience mortality. A perfect prediction would have an AUC of 1, while a non-informative prediction an AUC of 0.5. Tests comparing the AUC of the upgraded calculator to the PCPTRC 2.0 were performed using non-parametric U-statistic procedures. Clinical net benefit curves were compared between upgraded and PCPTRC 2.0 risks (Vickers and Elkin 2006). These curves were calculated as the proportion of true positive counts minus a weight multiplied by the proportion of false-positive counts. The weight, $[-threshold/(1-threshold)]$, varies by the threshold of the risk tool used to refer to biopsy and reflects the differential assignment by the clinician of the benefits and harms of true and false positives, respectively. At the threshold of 0.5 the harm of a false positive count is equal to the benefit of a true positive count (weight $= 1$); for thresholds less than 0.5 (weight < 1) the harm of a false positive is less than the benefit of a true positive, and for thresholds > 0.5

(weight > 1) the harm of a false positive exceeds the benefit of a true positive.

5.1.3 Results

From the SABOR biopsy cohort, four biopsies that were positive for cancer but had no Gleason score were censored, leaving 547 patients for analysis. Among these, 96 (17.6%) were diagnosed with high-grade prostate cancer on biopsy, 199 (36.4%) with low-grade prostate cancer, and 252 (46.1%) with no cancer (Table 2).

	No cancer	Low-grade	High-grade
	$n = 252$ (46.1%)	$n = 199$ (36.4%)	$n = 96$ (17.6%)
Age at biopsy			
Mean (SD)	64.2 (8.6)	65.0 (7.9)	64.3 (9.1)
Range	44.97 – 83.55	45.99 – 88.55	44.42 – 84.82
Race			
White	171 (49.1)	118 (33.9)	59 (17.0)
African American	35 (45.5)	30 (39.0)	12 (15.6)
Other	46 (37.7)	51 (41.8)	25 (20.5)
Prior biopsy			
Never	208 (48.0)	148 (34.2)	77 (17.8)
At least one	44 (38.6)	51 (44.7)	19 (16.7)
Digital rectal exam			
Normal	243 (55.2)	141 (32.0)	56 (12.7)
Abnormal	7 (8.0)	46 (52.9)	34 (39.1)
Family history			
No	224 (50.5)	147 (33.1)	73 (16.4)
Yes	28 (27.2)	52 (50.5)	23 (22.3)
PSA (ng/ml)			
Mean (SD)	1.5 (1.3)	3.8 (2.9)	15.7 (78.4)
Range	0.1 – 8.4	0.3 – 29.3	0 – 779
$[-2]$proPSA (pg/ml)			
Mean (SD)	1.5 (1.3)	3.8 (2.9)	15.7 (78.4)
Range	0.8 – 39.0	2.2 – 70.1	3.1 – 726.0
percent free PSA			
Mean (SD)	32.0 (11.6)	22.3 (11.5)	17.8 (9.7)
Range	6.6 – 73.0	5.6 – 72.0	4.4 – 48.8

Table 2 Characteristics of the $n = 547$ SABOR (training set) participants used in the analysis. For categorical outcomes, percents in each outcome group are provided across the rows.

DRE findings, family history, PSA, [-2]proPSA, and percent free PSA differed significantly among the three outcome groups ($p < 0.05$). PSA and [-2]proPSA increased, and percent free PSA decreased monotonically from the non-cancer to high-grade cancer groups. The joint distribution of percent free PSA and [-2]proPSA in SABOR showed considerable overlap among the three outcome groups (no cancer, low- and high-grade cancer, Figure 14). In all three outcome groups, percent free PSA decreased, and [-2]proPSA increased with increasing PSA (all p-values < 0.0001). Formulas for PCPTRC 2.0 risks and risks to incorporating percent free PSA and [-2]proPSA are provided on the PCPTRC 2.0 website (http://riskcalc.org/PCPTRC/).

Characteristics of the 575 EDRN biopsies used for validation are given in Table 3 and Figure 14. This cohort had more high-grade cancers and non-cancer diagnoses than SABOR, likely as it was more of a referral population than the population-based SABOR study. Age, DRE, PSA, [-2]proPSA, and percent free PSA differed significantly across the three outcome groups with trends similar to those in SABOR. For each of the three outcome groups (no cancer, low-grade, and high-grade) separately, different models using different assumptions of the error term were fit to the SABOR data. Bayesian information criterions (BICs) were compared across different models, which used different covariates as well as different error distributions. Note, the BIC for the skew-t distribution was calculated following Frühwirth-Schnatter and Pyne (2010).

	No cancer	Low-grade	High-grade
	$n = 324$ (56.3%)	$n = 110$ (19.1%)	$n = 141$ (24.6%)
Age at biopsy			
Mean (SD)	60.5 (7.8)	61.7 (8.5)	64.8 (9.7)
Range	42 – 80	46 – 83	41 – 93
Race			
White	276 (56.1)	92 (18.7)	124 (25.5)
African American	24 (53.3)	11 (24.4)	10 (22.2)
Other	24 (63.2)	7 (18.4)	7 (18.4)
Prior biopsy			
Never	324 (56.3)	110 (19.1)	141 (24.5)
At least one	0 (0)	0 (0)	0 (0)
Digital rectal exam			
Normal	261 (57.9)	96 (21.3)	94 (20.8)
Abnormal	59 (49.2)	14 (11.7)	47 (39.2)
Family history			
No	233 (57.4)	75 (18.5)	98 (24.1)
Yes	75 (53.6)	27 (19.3)	38 (27.1)
PSA (ng/ml)			
Mean (SD)	4.5 (3.1)	6.0 (9.2)	14.2 (33.8)
Range	0.3 – 18.2	0.7 – 94.1	1.0 – 310.6
[−2]proPSA (pg/ml)			
Mean (SD)	12.1 (10.4)	14.5 (14.7)	46.9 (169.2)
Range	2.2 – 133.9	4.4 – 145.0	3.9 – 1831.7
percent free PSA			
Mean (SD)	23.4 (10.6)	20.0 (11.6)	15.3 (9.0)
Range	6.4 – 64.8	5.9 – 77.0	3.7 – 78.9

Table 3 Characteristics of the $n = 575$ EDRN (validation set) participants used in the analysis. For categorical outcomes, percents in each outcome group are provided across the rows.

Both markers, percent free PSA and [-2]proPSA were assessed first separately and afterward in a multivariate model accounting for the correlation between the two markers. The assessment of the markers separately will be only briefly reported and discussed since the steps are like the multivariate approach, which was used finally.

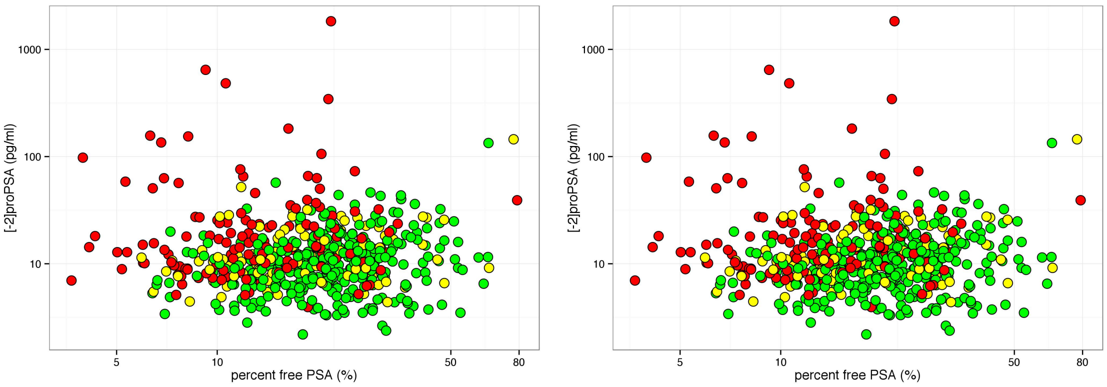


Figure 14 Scattergram of SABOR (left) and EDRN (right) showing the three outcomes and their corresponding [-2]proPSA and percent free PSA values, respectively.

5.1.4 Updating risk calculator - univariate models

A multinomial model builds the original risk calculator. So, three probabilities for three binary variables are estimated. That means that we need to update all three probabilities with the help of the information on the new biomarker. To discuss this method directly using the example, we will consider the three binary variables no cancer, low-grade, and high-grade in the following. The probability for each patient should still sum up to one overall three categories.

Different types of regression models were used to update the risk calculator. Linear regression, as well as more advanced models with different error distributions incorporating the linear predictor in the location parameter, were assessed. Note that for the skew-normal and skew-t distribution, the errors are not centered around 0. As explained in the theory part of this work in Section 2.4, mode and mean do not coincide with the skew-t and skew-normal distribution. The models were fit using maximum likelihood optimization. Four different error distributions have been considered: normal, t, skew-normal, and skew-t, where the degrees of freedom for t and skew-t distribution were fixed. Several baseline characteristics, such as $\log_2$ PSA and age, have been examined as covariates. Only $\log_2$ PSA was regarded as relevant, comparing different models with the likelihood ratio tests on training and validation set. Therefore, the following models will only include $\log_2$ PSA as a covariate. The BICs on the training set for each model were reported. The performance was assessed afterward for each of the four models on the validation set.

It is important to note that for each high-grade, low-grade, and no cancer, a different number of observations was available for each group. So, the model for high-grade cancer is based on 96 observations, 199 observations for low-grade cancer, and 252 observations for no cancer, respectively.

Table 4 shows an overview of the results of the different regression types for each group and biomarker. We can see that across all groups and for both biomarkers, the regression with a t distribution is working well since the BIC values are always among the lowest values across the different error distributions. Overall, the differences are quite small, considering the

Biomarker	Group	Error dist.	BIC	σ^2	λ	β_0	β_1
pct free PSA	High-grade	normal	201.93	0.50	-	4.41	-0.20
		t, $\nu = 6$	203.60	0.37	-	4.49	-0.23
		skew-normal	206.37	0.78	-0.69	4.78	-0.20
		skew-t, $\nu = 6$	208.03	0.36	0.15	4.36	-0.23
	Low-grade	normal	338.11	0.34	-	4.94	-0.39
		t, $\nu = 6$	345.55	0.26	-	4.97	-0.39
		skew-normal	342.15	0.75	-1.29	5.42	-0.39
		skew-t, $\nu = 6$	349.67	0.22	-0.32	5.25	-0.40
	No cancer	normal	389.38	0.27	-	4.93	-0.23
		t, $\nu = 6$	374.14	0.17	-	4.95	-0.21
		skew-normal	376.92	0.74	-2.23	5.46	-0.22
		skew-t, $\nu = 6$	372.88	0.11	-0.41	5.30	-0.21
[-2]proPSA	High-grade	normal	232.61	0.68	-	2.69	0.66
		t, $\nu = 6$	213.81	0.35	-	2.52	0.70
		skew-normal	222.65	1.19	2.76	1.77	0.68
		skew-t, $\nu = 6$	214.34	0.21	0.62	1.99	0.71
	Low-grade	normal	415.30	0.51	-	2.75	0.49
		t, $\nu = 6$	411.65	0.36	-	2.79	0.47
		skew-normal	419.45	0.87	-1.04	3.25	0.49
		skew-t, $\nu = 6$	415.29	0.28	-0.43	3.17	0.47
	No cancer	normal	492.52	0.40	-	2.82	0.55
		t, $\nu = 6$	497.35	0.30	-	2.83	0.55
		skew-normal	496.79	0.78	-1.11	3.29	0.55
		skew-t, $\nu = 6$	502.39	0.28	-0.20	3.01	0.55

Table 4 Overview results of normal, t, skew-normal, and skew-t regression for high-grade, low-grade, and no cancer for the biomarker percent free PSA and [-2]proPSA. For each model, the coefficients, as well as BICs, are shown.

BICs. It is also interesting to note that for all models, the estimated regression coefficient for $\log_2$ PSA has very similar values. For both biomarkers, the skewness estimates were always more pronounced for the high-grade group than for low-grade and no cancer. In general, the skewness parameter λ was different for the skew-normal and skew-t distribution. One reason for this might be that setting the degrees of freedom to six yields fatter tails. Hence, the skewness of the distribution could arise from data in the tails of the distribution. It seems that the skew-normal distribution might be too sensitive to skewed tails in the data.

Since the BICs are very similar across all models for each group and biomarker, it will be interesting to see the performance differences on the validation set. The more complicated models incorporating skewness parameters might tend to overfit on the training set and, therefore, could have worse performance on the validation set.

The performance of the prediction models was assessed on the EDRN validation data set. Calibration plots were evaluated and are only shown for the multivariate model. The prediction model tends to underestimate the risk on the validation set. The predicted risk is lower than the observed risk for high-grade as well as for cancer (combining high and low-grade). A reason for the lower risk might be that the EDRN data have a higher prevalence of cancer compared to the SABOR data. For the assessment with the measured AUC, the data of the training set are dichotomized in high-grade versus everything else, and cancer versus no cancer. The risk percentages from the risk calculator have been summed up accordingly for the respective groups to match the dichotomized groups.

New information	Calculator	AUC high-grade	AUC cancer
-	PCPTRC 2.0	0.73	0.68
	PSA	0.70	0.66
	$[-2]$proPSA	0.69	0.65
	percent free PSA	0.72	0.69
pct free PSA	Normal update	0.77	0.73
	Skew-normal update	0.77	0.72
	t update	0.77	0.72
	Skew-t update	0.77	0.72
[-2]proPSA	Normal update	0.71	0.63
	Skew-normal	0.68	0.65
	t update	0.72	0.65
	Skew-t update	0.70	0.65

Table 5 AUCs for continuous biomarkers on the original scale, PCPTRC2.0, and updated versions of the risk calculator for high-grade vs. everything else and cancer (high and low-grade cancer) vs. no cancer.

Overall, we see that according to the AUC, the PCPTRC 2.0 is performing well on the validation set. [-2]proPSA has the lowest AUC values, whereas percent free PSA has the highest AUC values of the continuous biomarkers on the original scale. The updated risk calculator with information from [-2]proPSA shows worse performance on the validation set. It seems including this marker yields no additional information for the risk prediction model. However, including percent free PSA improves the overall performance for all types of models. The AUC values are very similar across all models with simple regression being the model with the highest AUC value for cancer vs. no cancer.

5.1.5 Updating risk calculator - multivariate models

To incorporate the correlation structure between the two biomarkers [-2]proPSA and percent free PSA, multivariate models will be used. Utilizing the dependencies between the two biomarkers might result in a better performance of the risk prediction. Like the univariate case, the estimation results with BICs are reported for all groups and models. The scale matrix is symmetric, and therefore only three entries are shown in Table 6.

The performance of the updated risk calculator is evaluated on the EDRN data (validation set). The calibration curves (figure 15) show that all models underestimate the risk on the EDRN set for high-grade cancer. Similar to the original calibration curves, we can explain the underestimation of the risk by the unusually high prevalence of high-grade cancer observations in the EDRN data. Furthermore, we see that only a few data points are available for risk higher than 0.2. Overall the calibration curve for cancer vs. no cancer fits well for all versions of the risk calculator.

For overall cancer prediction (low- and high-grade versus no cancer), the improvement was

Grp.	Reg.	BIC	σ^2_{11}	σ^2_{12}	σ^2_{22}	λ_1	λ_2	β_{10}	β_{11}	β_{20}	β_{21}
H	normal	432.3	0.50	-0.15	0.69	-	-	4.41	-0.20	2.69	0.66
	t	383.0	0.34	0.19	0.35	-	-	4.55	-0.26	2.49	0.71
	skew-n	409.8	0.51	0.07	1.23	-1.45	3.68	4.58	-0.21	1.87	0.69
	skew-t	387.6	0.32	0.19	0.23	0.24	0.58	4.34	-0.26	2.00	0.71
L	normal	683.9	0.34	0.09	0.51	-	-	4.94	-0.39	2.76	0.48
	t,	668.5	0.27	0.20	0.36	-	-	4.99	-0.41	-0.41	0.47
	skew-n	687.8	0.55	0.35	0.57	-1.44	0.21	5.41	-0.39	-0.39	0.48
	skew-t	676.7	0.23	0.19	0.30	-0.3	-0.37	5.24	-0.41	-0.41	0.47
N	normal	774.7	0.27	0.08	0.40	-	-	4.93	-0.23	-0.23	0.56
	t	727.2	0.17	0.15	0.29	-	-	4.95	-0.22	-0.22	0.57
	skew-n	761.7	0.56	0.44	0.59	-2.22	-0.2	5.47	-0.22	-0.22	0.56
	skew-t	734.5	0.14	0.14	0.25	-0.25	-0.33	5.16	-0.22	-0.22	0.56

Table 6 Overview results of multivariate normal, t ($\nu = 6$), skew-normal and skew-t ($\nu = 6$) regression for high-grade, low-grade and no cancer for the biomarker combination of percent free PSA and [-2]proPSA. For each model, the coefficients, as well as BICs, are shown.

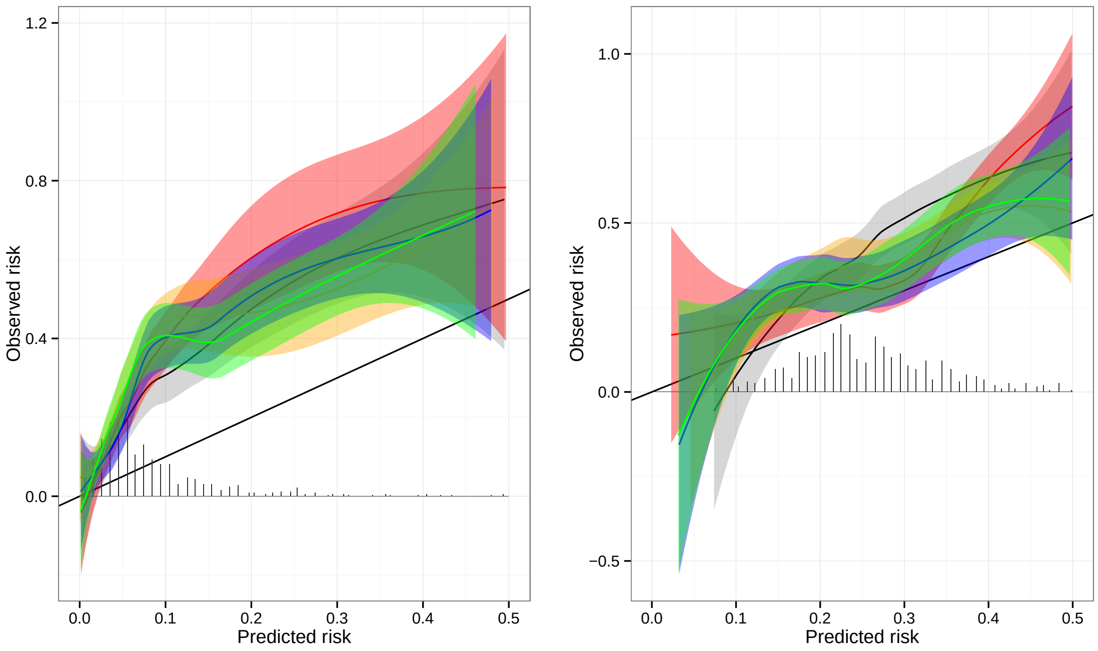


Figure 15 Calibration plot for high-grade and low-grade vs. no cancer (left), and cancer vs. no cancer (right) on the EDRN validation set for PCPTRC 2.0 (black), normal update (red), skew-normal update (orange), t update (blue) and skew-t update (green). The calibration plot is showing observed and predicted risk with rug plot of the available data. The colored calibration lines were estimated using locally weighted scatterplot smoothing with confidence corresponding intervals.

minimal with the addition of percent free PSA and [-2]proPSA (AUC $= 0.72$ versus 0.68 for the PCPTRC 2.0, respectively). For a threshold of prostate cancer risk of 34.3% (associated with 80% specificity), the sensitivity of the upgraded calculator for detecting prostate cancer was 54.6%. The net benefit of using the PCPTRC 2.0 with the two new biomarkers to determine overall prostate cancer risk greatly exceeded that of the PCPTRC 2.0 (Figure 16). Results incorporating percent free PSA alone were similar to the outcomes with both new biomarkers. For the prediction of high-grade disease, the PCPTRC 2.0 with the two new biomarkers

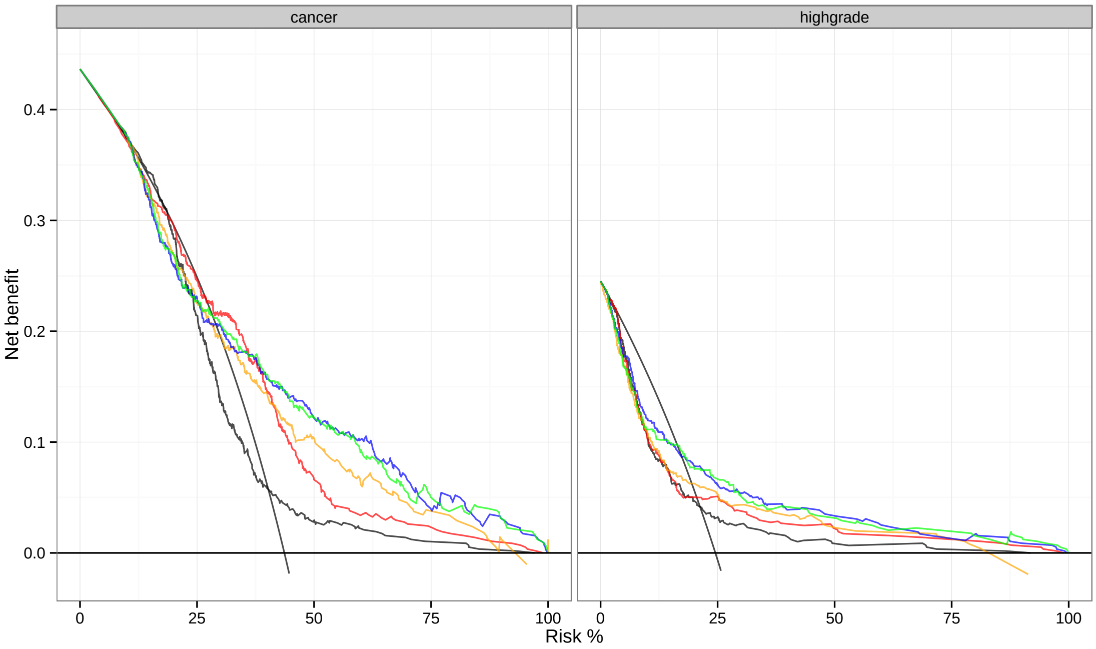


Figure 16 Clinical net benefit for cancer versus no cancer on the EDRN data (i.e. pooling high-grade and low-grade cancer) on the left panel, clinical net benefit for high-grade vs. other on the right panel.

Calculator	AUC high-grade	AUC cancer
Normal update	0.79	0.74
Skew-normal	0.79	0.72
t update	0.80	0.72
skew-t update	0.81	0.73

Table 7 AUCs for updated versions of the risk calculator comparing high-grade versus everything else and cancer (high and low-grade cancer) versus no cancer.

significantly improved the operating characteristics of the PCPTRC 2.0 (AUC $= 0.80$ versus 0.73, respectively, p-value = 0.001). Sensitivities of the upgraded risk calculator remained low (for a threshold of risk for high-grade cancer to prompt a biopsy of 7.7%, associated with 80% specificity, the sensitivity was only 64.5%). Net benefit curves showed that there would be a benefit to using the risk calculators for referring patients to biopsy for high-grade prostate cancer risks exceeding 20% (Figure 16); i.e., it would only be helpful for clinicians to recommend testing one or both additional biomarkers if a 20% or higher risk threshold of high-grade prostate cancer is used to prompt a biopsy. The PCPTRC 2.0 with percent free PSA performed as well as the calculator including [-2]proPSA.

5.1.6 Discussion

We and others have demonstrated the benefit of the use of a composite tool to assess a man's risk of prostate cancer instead of using current measures in a dichotomous fashion (Thompson et al. 2006). For example, a PSA of 3.8 in a 55-year-old Caucasian man with a prior negative biopsy and no family history of prostate cancer has an 11% risk of low-grade prostate cancer and a 3% risk of high-grade cancer. On the other hand, the same PSA in a very healthy 73-year-old African American man with a family history of prostate cancer and no prior biopsy has risks of 22% for high-grade and 22% for low-grade cancer, respectively. PSA alone should not be used to determine cancer risk or, more importantly, the risk of high-grade cancer. Similarly, a 55-year-old Caucasian man with a prostate nodule who has no other risk factors and is subsequently found to have a PSA of 0.4 has risks of low-grade cancer of 8% and only a $< 1\%$ risk of high-grade cancer.

Other biomarkers have been developed to estimate more accurately the patient's risk of cancer. Both percent free PSA and [-2]proPSA are widely available for this purpose. In this study, we explored whether these biomarkers should be incorporated into clinical decision-making. From our initial evaluation in the SABOR cohort with validation in the EDRN cohort, we found that the biomarkers do indeed improve diagnosis but only in a limited range of risks.

As can be seen from Figure 16, the net benefit curves only separate at risk of 25%. This separation indicates that if a physician uses a 25% or higher prostate cancer risk to recommend a prostate biopsy (a method strongly supported, rather than the inaccurate use of PSA alone), the addition of the biomarkers tested in this study will improve the recommendation for biopsy. Similarly, at about 8% risk of high-grade prostate cancer as a threshold to recommend a biopsy, the new biomarkers can be helpful for clinical recommendations. As crucial from this analysis was the observation that the PCPTRC 2.0 including only percent free PSA as an additional biomarker performed as well as the inclusion of percent free PSA and [-2]proPSA. We have previously made a similar observation.

5.1.7 Conclusion

Percent free PSA and [-2]proPSA improve the performance of the PCPTRC 2.0. At levels of risk of 25% for low-grade cancer and 8% of high-grade cancer and higher, clinicians should consider the incorporation of these biomarkers in the patient evaluation before recommendation of prostate biopsy. The above-discussed results are now available at http://riskcalc.org/PCPTRC/. Users of the risk calculator can now additionally set their percent free PSA value and get a more precise prediction of their prostate cancer risk. The user interface focuses on simplicity. The calculations and derivation of the model are hidden, and only the relevant information is shown. The results are presented in a comfortable and straightforward to understand way. This way of presentation makes it feasible for everybody to use the risk calculator. The risk calculator should be only used as a tool that gives additional information based on previous medical results. It is important to note that the risk calculator should not replace but only support the physician in charge.

5.2 Tree-mortality prediction

5.2.1 Introduction

Tree-mortality is an essential outcome for surveillance in forest management. Many complex, deterministic and stochastic models have been developed that can simulate tree-based growth and mortality according to proven scientific relationships, thus facilitating prediction of tree-mortality (Pretzsch 1992; Monserud and Sterba 1999; Pretzsch et al. 2002; Fortin et al. 2008; Kiernan et al. 2009). In practice, several methodologic issues can arise in this setting. The observational period of individual trees, as well as the time between follow-up, can vary, leading to unbalanced data. The model needs to account for the dependence structure arising from multiple observations per tree, and in addition, there may be a high number of missing values.

Predicting tree mortality is one of the missions with great potential in forest science and has been tackled with various statistical methods, with logistic regression, the most common approach (Fabrika et al. 2018; Hartmann et al. 2018; Salas-Eljatib et al. 2018; Bravo et al. 2019). Logistic regression has the advantages of being easy to use with interpretable coefficients expressed as odds ratios, as well as being available in nearly every statistical package. To handle complex forestry data, however, where there is dependence among plots and regions as well as non-linear predictor effects, adaptations of standard logistic regression is needed. Using data from the Chair for Forest Growth and Yield at the Technische Universität München, Böck et al. (2014) described an approach for modeling tree mortality using random-effects incorporating splines to achieve more flexibility. The same data set will be used in the following, applying a different methodology for predicting tree mortality.

This application outlines an alternative approach to logistic regression for predicting tree mortality that is closely related to traditional classification and discrimination ideas. Multivariate mixture models are fit using the EM algorithm described in Section 3.3 to the multi-dimensional mortality risk factors separately among the strata of trees that experienced mortality and those that did not. Visualization shows collapsed clusters among the multivariate distributions, thence the modified ECME algorithm in Section 3.4 is implemented to handle these. Combining the multivariable risk factor distributions with prevalence estimates of mortality results in calibrated mortality risk models, which are then compared to logistic regression via validation comprising separate test and training sets.

5.2.2 Material and methods

The Chair for Forest Growth and Yield at the Technische Universität München collected the data reported here (Böck 2014; Böck et al. 2014), which can be summarized as follows. Data were obtained between 1954 until 2007 from beech trees in 60 plots at 11 test sites in a Bavarian long-term forest research plot network. The observational periods ranged from 3 to 28 years, and individual trees were observed between one to seven observational periods. In total $21,051$ single tree observation periods comprising $9,292$ beech trees were available for the analysis. Only 604 periods (2.9%) were associated with mortality. Five tree characteristics measured at the beginning of each observation period were modeled, representing the size, height, and quality of the tree location, as well as criteria for the competition concerning light and space.

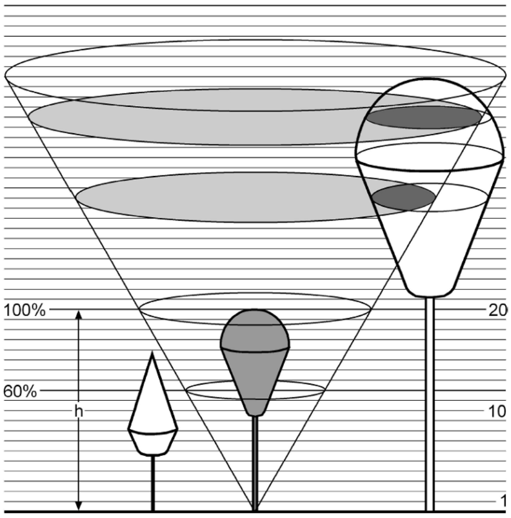


Figure 17 The principle for determining vertical competition profiles. Any tree intersecting the cone around a tree (shaded in gray) is considered a competitor. The competition indices quantify the light competition, competition within the same species or compared to conifers, and the amount of over-shading by other trees. Higher values represent more competition from overshading. Printed with permission from Böck et al. (2014).

Diameter at Breast Height (DBH) measures the diameter of the tree at 1.3 meters distance from the ground (approximately breast height). The next characteristic, height, measures the tree height in meters. The consolidation of confidence intervals from the height and the relative distance to the other trees results in competition indices, illustrated in Figure 17. KKL quantifies the total light competition by neighboring trees. The vertical competition profile for trees of the same species, by beech trees in this study, is quantified by CIIntra, while CIConifer measures the competition from conifer trees. CIOvershade is a measurement for the extent of overshading by other trees. More precise definitions and units are provided in Pretzsch (2001) and Böck (2014). Only the five factors DBH, KKL, CIIntra, CIOvershade, and CIConifer, will be considered as these were the most influential for tree mortality, as shown by Böck et al. (2014).

Following visualization of the data, multivariate mixtures of skew-t distributions were fit to characteristics of trees separately for those that did and did not experience mortality at the end of the observation period. All ten pairs of characteristics were modeled using bivariate skew-t mixtures. One to three-component mixture models were fit separately to trees that did and did not experience mortality. Additionally, five-dimensional skew-t mixtures with one and two components were fit to all five characteristics for comparison to the bivariate models. For validation, data were split into training and validation sets with a ratio of 2:1 and with the same proportion of mortality events in each set. Hence, the models were fit on $2/3$ of the data resulting in $13,373$ tree periods in the non-mortality subset and 378 tree periods in the death subset. The number of mixture components was determined using the BIC, which was preferred over the AIC to limit the number of variables and mixture components in the model. For the prediction, the following density ratio (as in Section 2.27) weighted by the prevalence of trees that did and did not experience mortality was calculated as

$$\frac{P(Y=1)}{P(Y=0)}\frac{f_{Y=0}(x)}{f_{Y=1}(x)},$$

where Y is the mortality status of the trees, and $f_{Y=0}(x)$ and $f_{Y=1}(x)$ are the fitted multivariate skew-t densities of the tree characteristics.

Marden (1998) introduced quantile-quantile (QQ) plots for bivariate data arising from two-dimensional distributions, which were used here to assess the model fit of the bivariate skew-t mixtures. The idea of bivariate QQ plots is to find an empirical quantile that minimizes the distance to the theoretical quantile. Hence, for each observation, the distance to the theoretical quantile is calculated and visualized as an arrow in the graph. Arrows indicate how far observations from the empirical distribution deviate from their theoretical counterparts. Hence, if the graph shows no arrows, the empirical distribution perfectly matches the theoretical one.

As the reference model, the well-established logistic regression was applied to the data, including all five risk factors as predictors in the model with no interactions. Additionally, a five-variate skew-t mixture was fit to risk factors to form the likelihood ratio, as for the bivariate skew-t mixtures. The AUC, introduced in the prostate cancer application in Section 5.1.2, was used as a performance measure for prediction on the validation set, which consisted of $7,074$ non-mortality and 276 mortality periods.

5.2.3 Results

Pairwise scatterplots of the five tree characteristics in Figure 17 show that for lower values of DBH, higher values of competition indices could be expected, with negative correlations of -0.523 for KKL, -0.41 for CIIntra and -0.675 for CIOvershade.

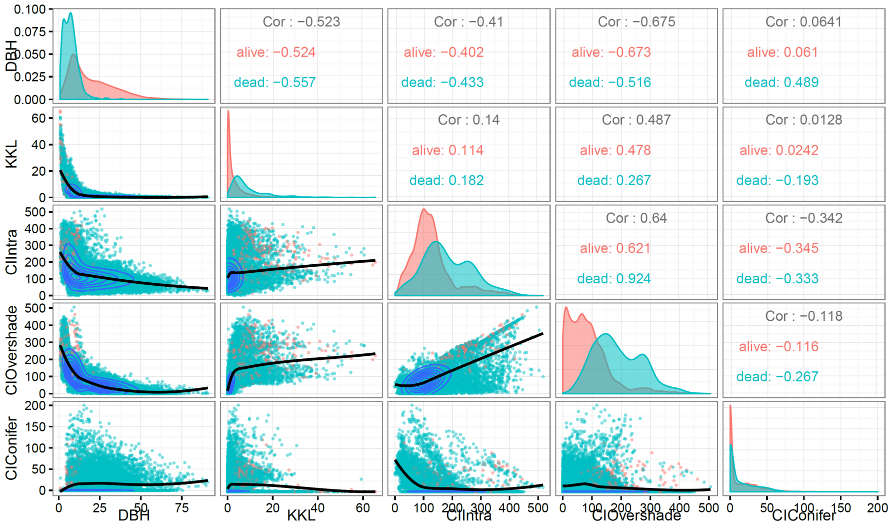


Figure 18 Scatterplots with overlaid contours of the fitted two-component skew-t mixture models and local linear regression curves in the lower matrix, density plots by mortality status in the diagonal, and pairwise correlations in the upper matrix.

Correlations were similar among the groups of trees experiencing and not experiencing mortality.

The scatterplot of DBH with CIConifer has a pattern that differs from the other three competition indices. Here, it is more difficult to detect a definite pattern. There is a high number of CIConifer values equal to zero for almost the complete support of DBH. The correlation is close to zero in the non-mortality subgroup and has a positive correlation of 0.489 in the group experiencing mortality. Based on this result, we expect a predictive potential for these characteristics since the bivariate distribution differs across the subgroups.

Models	Comp.	Non-mortality	Mortality
DBH and KKL	1	164891.48	7006.68
DBH and KKL	2	148357.33	**6587.52**
DBH and KKL	3	**145247.63**	6604.73
DBH and CIIntra	1	254062.03	10017.53
DBH and CIIntra	2	250212.18	**9935.43**
DBH and CIIntra	3	**248979.51**	9990.71
DBH and CIOvershade	1	247408.33	9931.70
DBH and CIOvershade	2	239236.03	**9820.78**
DBH and CIOvershade	3	**236550.10**	9840.10
DBH and CIConifer	1	206945.25	7603.81
DBH and CIConifer	2	189317.39	**3892.77**
DBH and CIConifer	3	**137346.67**	3946.21
KKL and CIIntra	1	219777.35	10703.32
KKL and CIIntra	2	217621.17	10722.99
KKL and CIIntra	3	**156223.71**	**10684.89**
KKL and CIOvershade	1	212337.78	10598.23
KKL and CIOvershade	2	206082.52	**10540.80**
KKL and CIOvershade	3	**202629.49**	10580.86
KKL and CIConifer	1	171690.79	8415.56
KKL and CIConifer	2	162195.65	**4711.79**
KKL and CIConifer	3	**156223.71**	4760.99
CIIntra and CIOvershade	1	305826.85	12341.13
CIIntra and CIOvershade	2	302462.31	**12085.99**
CIIntra and CIOvershade	3	**291562.90**	12096.96
CIIntra and CIConifer	1	48992.03	2606.78
CIIntra and CIConifer	2	38212.19	**2309.37**
CIIntra and CIConifer	3	**35101.23**	2368.23
CIOvershade and CIConifer	1	262207.59	11431.92
CIOvershade and CIConifer	2	194421.07	**7675.39**
CIOvershade and CIConifer	3	**192460.28**	7701.68
DBH, KKL, CIIntra, CIOvershade and CIConifer	1	567586.60	15383.69
DBH, KKL, CIIntra, CIOvershade and CIConifer	2	509007.90	14859.38
DBH, KKL, CIIntra, CIOvershade and CIConifer	3	**496700.80**	**13950.34**

Table 8 BICs of bivariate and five-dimensional skew-t mixtures with up to three components fit separately for trees that experienced and did not experience mortality. The lowest BICs are indicated with bold.

The correlation of KKL with CIIntra was positive and lower than the correlation of KKL with

ClOvershade. In the scatterplot of KKL and ClConifer, no clear trend could be seen, which a correlation of almost zero confirmed. The highest correlation occurred between ClIntra and ClOvershade among mortality periods (0.924).

For all pairwise scatterplots with ClConifer, no clear pattern was visible, which could be because ClConifer had an unusually high number of zeros. Correlations of the competition indices with ClConifer were rather low, with the highest in absolute value for ClIntra, which was negative (-0.342). In the density plot of DBH, trees with higher DBH tended to be in the non-mortality subgroup, whereas the tendency was the opposite for KKL, ClIntra, and ClOvershade. For ClConifer, densities of both subgroups were similar.

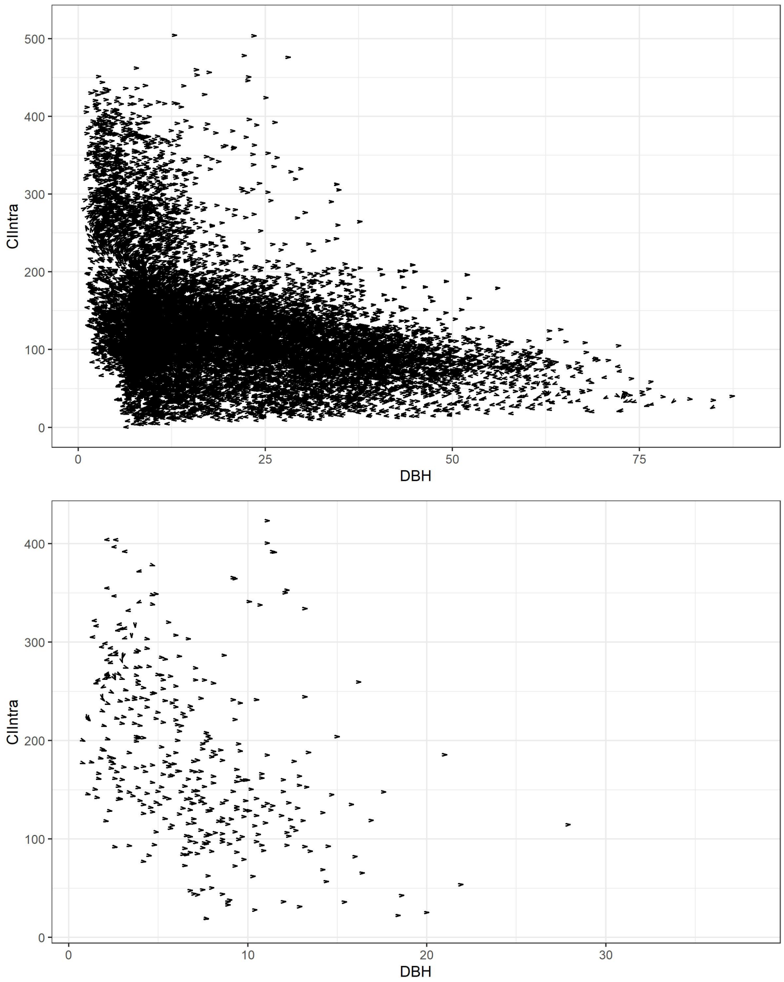


Figure 19 QQ-plot for DBH versus ClIntra for the observation periods where individual trees did not (upper) and did (lower) experience mortality. From each data point, an arrow to the theoretical quantile from the fitted skew-t distribution is added.

The marginal density plots in Figure 18 indicated that the risk factors were not normally distributed. Some, such as ClOvershade, had bimodal forms. Most of the data were left-skewed, and ClConifer was zero-inflated. Based on this examination, multivariate mixtures of skew-t distributions were reasonable.

Table 8 shows that for all bivariate mixtures, three components for the non-mortality subgroup versus two components for the mortality subgroup were chosen. This could be attributed to the much larger sample size of non-mortality tree periods $(20,447$ versus $604)$, which supported a bigger model with more parameters. As expected from Figure 18, collapsed clusters were detected for CIConifer and KKL when using skew-t mixtures with two or three components. Hence, the density ratios were calculated using three-components for trees not experiencing mortality versus two components for trees experiencing mortality.

Bivariate models	AUC with 95% confidence interval
DBH and KKL	0.838 (0.818, 0.858)
DBH and CIIntra	0.836 (0.814, 0.859)
DBH and CIOvershade	0.855 (0.835, 0.875)
DBH and CIConifer	0.844 (0.825, 0.863)
KKL and CIIntra	0.854 (0.834, 0.874)
KKL and CIOvershade	0.868 (0.849, 0.887)
KKL and CIConifer	0.728 (0.698, 0.758)
CIIntra and CIOvershade	0.852 (0.829, 0.874)
CIIntra and CIConifer	0.615 (0.581, 0.649)
CIOvershade and CIConifer	**0.871 (0.852, 0.890)**
DBH, KKL, CIIntra, CIOvershade and CIConifer	0.870 (0.852, 0.888)
Logistic regression	0.862 (0.843, 0.880)

Table 9 Mixture models with optimal number of components were fit on the training set and used to predict mortality status with performance measured by the AUC.

Visual inspection of the arrows in Figure 19 showed that the theoretical quantiles of the selected models matched quite well for trees not experiencing and experiencing mortality.

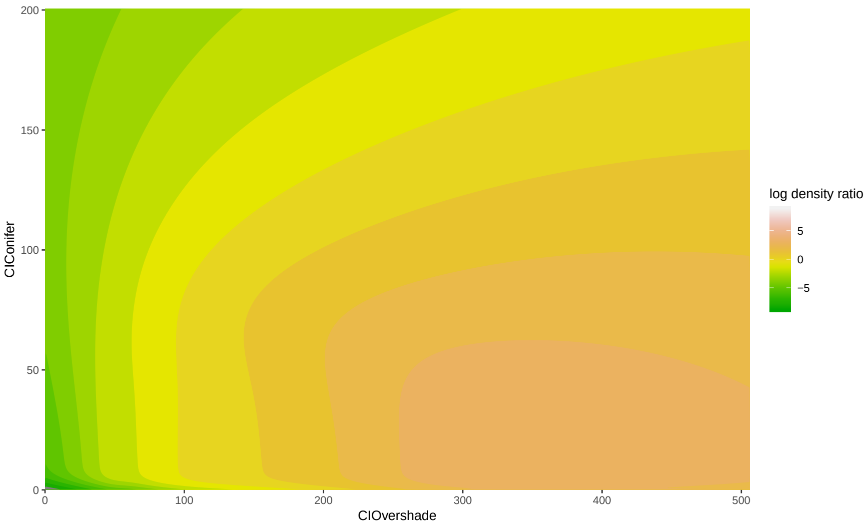


Figure 20 Logarithm of density ratios for bivariate skew-t mixture models of CIOvershade and CIConifer fit separately to mortality and non-mortality tree observation periods. The green color indicates a low risk of mortality while red a high risk.

The AUCs in Table 9 showed that for all bivariate models, the discrimination was quite high on the validation set, ranging from a low of 0.615 for CIIntra - CIConifer to a high of 0.871 for CIOvershade - CIConifer. The skew-t mixture model including all five tree characteristics had an AUC only marginally lower at 0.870 and the logistic regression model using the five predictors had an AUC of 0.862, which was not far behind.

Figure 20 shows a heat map of the log density ratio that can be used as a tool in forest management to visually differentiate which type of trees are at risk of mortality. The graph shows that trees with CIOvershade in the range of 300 to 500 and small CIConifer values were at higher risk of experiencing mortality.

5.2.4 Discussion

The tree mortality prediction method based on multivariate skew-t mixtures outperformed logistic regression slightly, even when only two out of five characteristics were used instead of all five as in logistic regression. One reason for this might be that since there were so few trees with mortality (2.9%), the multivariate skew-t models could better tailor the model to the mortality population. It is known that logistic regression and machine learning methods, in general, can experience diminished performance in the face of highly unbalanced data (King and Zeng 2001; Puhr et al. 2017; Maalouf et al. 2018). Logistic regression and other machine learning methods are popular because they are easy to use. Lu et al. (2013) has compared logistic regression to classification approaches based on multivariate models as performed here, and found that such models do not always outperform logistic regression. Similarly, a systematic review of clinical prediction models by Jie et al. (2019) showed no performance benefit in favor of machine learning methods over logistic regression, concluding that they may not be worth the computational effort. By creating an R package, we have relieved the computational burden for fitting skew-t mixture models. However, fitting the five-variate models to the large sample of trees not experiencing mortality ($n = 13,373$) took one week on a laptop. In contrast to prior studies, we have provided an application where skew-t mixtures improved prediction compared to logistic regression, albeit not by a large magnitude.

6 Discussion

This thesis has examined and extended a flexible parametric approach, multivariate mixtures of skew-t distributions, to model multi-dimensional data for utilization in prediction modeling or for augmenting prediction models in the life sciences. The historically proven EM algorithm was modified to facilitate fast identification and fitting for potentially moderate-dimensional data comprising independent and identically distributed data. Chapter 5 comprised two applications; the first applied skew-t mixtures to bivariate observations of serum markers for prostate cancer detection and the second, a five-dimensional set of forest mortality risk factors. Both applications demanded robust models since the respective data sets contained outliers. As the introduction covered the vast background literature on finite mixtures and the EM algorithm, the focus turns in this section to alternative approaches, to the problems faced in this thesis, as well as the outlook to even more challenging data applications, including non-independent and big data.

6.1 Non-parametric alternatives

There are many classical non-parametric approaches for modeling multi-dimensional data. Kernel density methods, in their purest form the histogram, enable the estimation of densities (Pearson 1894; Venables and Ripley 2013; Scott 2015). Histograms are the graphical representation of binned data, showing the frequency in each bin. Histograms are usually not used for higher-dimensional data because the visualization is not possible, and binning each dimension separately can result in sparse multi-dimensional bins. Kernel density methods use multivariate kernels, including covariance structures, to estimate the density of data. For kernel density methods, the convergence rate of the mean squared error (MSE) of the kernel density estimator decreases as the dimension decreases, and the number of observations increases (Duong and Hazelton 2005; Silverman 2018). This phenomenon is called the curse of dimensionality in the literature (Chang and Sangrey 2019). To increase the rate of convergence the choice of bandwidth is crucial, and difficulties might occur depending on the smoothness of the underlying distribution.

One interesting approach that can handle similar problems as those in this work is the multivariate density estimation with Bayesian partitioning proposed by Lu et al. (2013). It works by splitting the space into an unknown number of disjoint regions (Holmes et al. 1999). The assumptions for this method are that in each partition, the data are exchangeable and follow some simple distribution. Usually, MCMC methods are used to obtain the distribution of the partition; they consecutively average across samples to obtain a smooth prediction surface for the multivariate distribution. To combat the discreteness and subsequent lack of convergence issues, Lu et al. (2013) adopted a Bayesian approach, defining new priors for which the posterior distribution of the kernel partitions are analytically available, thus making estimation computational feasible. Hence, with this algorithm to estimate posterior distributions, the efficiency for determining multivariate densities could be improved compared to classical kernel density techniques. Like our approach, theirs too considered the application of classification. They introduced class-specific densities estimated using cross-validation on a training set and found their method to be comparable, but not superior to other standard classifiers, such as support vector machines, classification, and regression trees.

Since we discussed finite mixture models in this work, it is natural also to have a look at infinite mixture models, where data come from a mixture of an infinite number of distributions (El-Arini 2008). One of the most popular is the Dirichlet mixture model, which belongs to the field of Bayesian non-parametric distributions and traces its introduction to Ferguson (1973) and Antoniak (1974). The Dirichlet process, as described by West and Escobar (1993) and Escobar and West (1995), can be understood as a two-parameter distribution over distributions, enabling a way to draw random samples. Usually, $X \sim DP(X_0, \alpha)$ denotes the Dirichlet process, where X_0 is the base probability distribution and α the scaling parameter, which essentially tells how similar the base distribution is to a distribution drawn from the specific Dirichlet process. The basic idea behind the Dirichlet mixture models is that realizations are from a continuous distribution, for example, a Gaussian, where no two values would be the same, i.e., the probability that any two samples are equal is zero (Blackwell and MacQueen 1973). Dirichlet mixtures consist of a countably infinite number of point masses. For fitting a Dirichlet mixture, each observation in a cluster comes from a distribution that is drawn from a Dirichlet process. Published applications of infinite mixture models use very similar algorithms for fitting as skew-t mixtures. The algorithms often apply MCMC and EM methodology. The main advantage of Dirichlet mixtures is the automatic determination of the number of clusters (Rasmussen 2000). One of the disadvantages compared to finite mixtures is that mixture components do not have an interpretation. All pros and cons when comparing parametric and non-parametric models apply. In this thesis, skew-t mixtures were preferred because of the lack of interpretability of clusters and their visualization.

Beyond classification, the regression potential of skew-t distributions comes from their ability to account for non-normal data. Quantile regression is a further option to tackle this problem. In contrast to linear regression considering the mean, quantile regression estimates the quantiles such as the median for the response variables. The initial idea of quantile regression was based only on the median. Formally, quantile regression was introduced by Koenker and Bassett Jr (1978) to complement linear regression. Since then much literature discussing properties and extensions has been published, including Stigler (1984) and Koenker and Hallock (2001), with applications in many fields, such as in climate change (Buchinsky 1994; Yu et al. 2003; Matiu et al. 2016).

The starting point of quantile regression is that the sample median minimizes the expected sum of the absolute error between observations and the population median. Thus, the median is the solution to the minimization problem that minimizes the expected sum of absolute errors. With this thought in mind, any quantile τ can be written as an optimization problem, with the goal to minimize a sum of asymmetrical weighted absolute errors depending on τ. A linear combination of the predictor variables forms the solution of the minimization problem. Quantile regression can be extended to not only perform optimization with regards to one quantile, but all quantiles simultaneously (Hao et al. 2007; Koenker et al. 2017). The set of all quantiles of the response conditional on the predictor variables reproduces the conditional distribution of the response. Quantile regression can be interpreted very similarly to linear regression. One unit change in the predictor quantifies the estimated change in a quantile of the response variable. Quantile regression overcomes several problems that appear in linear regression, such as violation of homoscedasticity of the error terms, or sensitivity to outliers. Quantile regression does not only focus on the mean, while disregarding the tails of the distribution. However, there are still difficulties when applying this method, including computational time, implementation of statistical inference, or estimation of standard errors. However, the R package `quantreg` (Koenker 2019) is continuously improving and providing new features. For the applications in Section 5.2, skew-t mixtures were used to account for the different clusters, such as particularly the collapsed clusters, in the data.

6.2 Reversible-jump Markov chain Monte Carlo

As discussed in Section 3.3 Frühwirth-Schnatter and Pyne (2010), who performed estimation of skew-t mixtures with MCMC, chose the optimal number of mixture components with the help of the BIC. We followed their example and also used the BIC to select the number of mixture components. Additionally, we provided an example of how to use MCMC to estimate skew-t mixtures for the parameterization used in this work. An alternative to choosing the optimal number of components would be Reversible-Jump Markov Chain Monte Carlo (RJMCMC) as an extension to MCMC (Green 1995; Richardson and Green 1997; Green and Hastie 2009). RJMCMC, which does not fix the dimension of the vector representing the states in the Markov chain, adds the possibility of joint inference across dimensions. One typical example of joint inference across dimensions is variable selection. For statistical modeling, variable selection changes the dimension of the model space since a parameter disappears when the variable is not in the model. The RJMCMC framework, regarded as trans-dimensional modeling, opens a wide range of applications, including incorporating variable selection seamlessly into the fitting process. Many different applications have implemented RJMCMC including Johnson et al. (2003), Sisson (2005), Li et al. (2012), and Papastamoulis et al. (2017). But in one of the first applications, Richardson and Green (1997) showed how to apply RJMCMC to mixture models. The method allows the estimation of the optimal number of components simultaneously to the estimation of the model parameters. The optimal parameters of the model are not only searched in the subspace of a model with a fixed number of components but estimating the optimal number of components during the parameter estimation. The fitting process is a two-step procedure. Conditionally, on the number of components, the optimal parameters are estimated using standard MCMC. Conditional on the estimates from the previous step, the dimension can be changed. Birth and death processes are responsible for adding or dropping dimensions, or in this case, the mixture components.

6.3 Statistical learning approaches to classification

In Section 2.3, we explained the connection of density ratios and linear discrimination in the case of assuming normal distributions for univariate outcomes with equal variance and quadratic discrimination for unequal variances, with the latter resulting in more instability of predictions. Further relaxations to skew-t or skew-t mixture distributions helped to enhance the flexibility even further but led to increased instability and potential overfitting on the training set, which deteriorated out-of-sample prediction. This instability was heightened as the dimensionality of the risk factor space increased, following the curse of dimensionality. Logistic regression has often been favored to discrimination approaches for classification since it does not model the high-dimensional risk factor space but instead treats the factors as fixed covariates with no distribution. With the coefficients having a natural interpretation as odds ratios of risk factors in the model, logistic regression with its many extensions offers simple usage. Furthermore, it can handle very complex data with the incorporation of splines to deal with anomalies, for example, for zero-inflated data.

The forest example in Section 5.2 assessed the same data as Böck et al. (2014), who observed that logistic regression with random effects and B-splines performed well in terms of the AUC. The discrimination performance was very similar to the density ratio approach presented in this work. In practice, the density ratio approach adds additional information, since not only the mortality status is estimated, but also characteristics of trees that do and do not experience mortality are individually modeled. Modeling of characteristics provides more in-depth insights into possible mechanisms for forestry experts, as shown in Figure 20. In addition, it could provide improvement in other data sets with more abundant covariate distributions. Many machine learning approaches have recently gained popularity, such as artificial neural networks and deep learning (Ripley 1993; Cheng and Titterington 1994). These approaches will provide alternatives to logistic regression, particularly when there are high-dimensional predictors (Dechter 1986; Schmidhuber 2016).

6.4 Methods for incorporating new factors into existing risk tools

Risk prediction tools are usually built on a large cohort study with several thousand participants to be representative of the population. For example, the PCPTRC 2.0 discussed in Section 5.1 made use of 3,900 members with up to 13 years of follow-up time. Since practitioners in the clinic frequently apply them, clinical practice guidelines incorporate the latest risk prediction models (Freedman et al. 2005).

A constantly evolving clinical landscape evolves in the challenge of how to update the numerous risk prediction models for various indications, such as breast (Gail and Mai 2010), colorectal (Freedman et al. 2009) and prostate (Ankerst et al. 2014; Grill et al. 2015a,b). Already many years earlier, Beven and Binley (1992) had discussed the future of distributed models considering model calibration and uncertainty prediction. Since then, many frequentist and Bayesian approaches emerged for updating risk prediction tools and incorporating new risk factors (Gail and Costantino 2001; Steyerberg 2008; Moons et al. 2012; Grill et al. 2017). Our likelihood ratio approach to incorporating new risk factors followed Ankerst et al. (2014) and Grill et al. (2015a,b).

6.5 The R shiny application

The prostate cancer application of Section 5.1 used the R Shiny package to bring the results of the prediction model to the users (Chang et al. 2019). The R package `shiny` was created in 2012 and provides an excellent framework for building web applications while running R sessions in the background to generates graphs and calculate results (Chang et al. 2019). Since then, many extensions and improvements to the package have been published, making use of the latest web technologies. Many statistical publications also provide shiny applications for the developed model to additionally provide an interactive tool usable for daily life (Jahanshiri and Shariff 2014; McMurdie and Holmes 2014; Nisa et al. 2014; Spitzer et al. 2014; Metsalu and Vilo 2015; Dunning et al. 2017; Francis 2017; Nelson et al. 2017). The prostate cancer application provides a simple user interface available in Spanish or English. Depending on specific input, additional input options might appear or disappear, which helps the user to focus on the relevant information on the screen. Comparing the prostate cancer application to other risk prediction calculators from the Clevland Clinic www.riskcalc.org website shows that usability is one of the key features for accessibility to broad audiences. The prostate cancer application provides tooltips for all input parameters to explain in simple words the required information.

6.6 Big messy data

Nowadays, data come in large streaming forms, which makes the quick fitting of parametric models by slow iterative algorithms, such as the EM algorithm, infeasible (Walker 2014). The term big data is one of today's buzzwords that describes data that is considered to be big regarding volume, variety, or velocity (Laney 2001). Processing big data is a challenge that needs to be handled in the future to deal with the massive volume of data. One problem with this type of data is that often only a little part is of interest. The vast quantity of data does not necessarily mean an increased information content in the data. Data can have redundancies or be permeated with uninformative data. Later concepts describe data additionally by how much noise it contains (Goes 2014) or its value (Marr 2014). For example, looking at data collected by search engines, it is crucial to understand that only a fraction of the data is usable. Identifying and filtering out crawlers that populate the search data with background noise that is not meaningful and clogs the process becomes a critical aspect.

Currently, fitting the skew-t mixture models on a conventional laptop takes much time, especially for big data sets and high-dimensional models. Even with the R package proposed in this work, making use of efficient algorithms, fast implementation, and multi-processing technology, the roughly 21,000 observations used in the forestry application resulted in about 15 hours of computation time for a 5-dimensional model. The increase in computation time is not linear in the number of observations for the implemented algorithm, and thus it might prove difficult for high-performance computing to apply this method to big data. Thus, it is evident that the proposed methodology cannot be applied to real-time applications where nearly an instantaneous result is required.

With big data, nearest-neighbor or other efficient machine learning approaches become a more feasible choice. For example, risk models can be built directly on live Electronic Medical Record (EMR) databases and linked to data collected with smartphones. The messy part of big data often has to deal with missing data. Missing data can hinder the convergence of the ECME algorithm even when the assumption of missing at random holds. Therefore, in the applications of this thesis, the skew-t densities were only fit with the ECME algorithm on complete data. Obtaining complete data was either accomplished by reducing the data set to have complete entries or using data imputation. Nevertheless, when viewed as density estimators, skew-t mixtures could be automated and applied to a reduced collection of observations identified with specific properties, for example, nearest neighbors. A shortcoming would be the restriction to continuous data since often risk factors are discrete. Hence, further work on mixture models will be required to capitalize on the advantages of nonparametric and parametric models, and thereby open this class of models to a broader range of applications. With increasing computational power of servers and devices, it will be possible to overcome the issues concerning computational time and use the potential of more complicated models such as skew-t mixture models for real-time applications.

A Definitions

A.1 Functions

Definition A.1 (Gamma function). *The gamma function is defined as*

$$\Gamma(z) = \int_0^\infty x^{z-1} \exp -x \, dx,$$

where $z \in \mathbb{R}$.

Definition A.2 (Digamma function). *The digamma function is defined as*

$$\psi(x) = \frac{d}{dx} \ln\left(\Gamma(x)\right) = \frac{\Gamma'(x)}{\Gamma(x)},$$

where $x \in \mathbb{R}$.

A.2 Univariate distributions

Definition A.3 (Gamma distribution). *Let $\alpha, \beta \geq 0$. A random variable G is said to be gamma distributed with shape parameter α and scale parameter β, denoted by $G \sim \mathcal{G}(\alpha, \beta)$, if its probability density function is*

$$f(g \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} g^{\alpha-1} e^{-\beta g} \, 1_{\{g \geq 0\}}, \; g \in \mathbb{R},$$

where $1_A = 1$ if A holds, 0 otherwise.

The expectation of a gamma distributed random variable G is $\mathrm{E}(G) = \frac{\alpha}{\beta}$, but the gamma distribution is typically right-skewed.

Definition A.4 (Normal distribution). *Let $\mu \in \mathbb{R}$ and $\sigma^2 \in (0, \infty)$. A random variable Z is said to be normally distributed with mean μ and variance σ^2, denoted by $Z \sim \mathcal{N}(\mu, \sigma^2)$ if its probability density function is*

$$f(z \mid \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2\sigma^2}}, \; z \in \mathbb{R}.$$

Definition A.5 (Student t distribution). *Let $\nu \in (0, \infty)$. A random variable T is said to be t distributed with degrees of freedom ν, denoted by $T \sim \mathcal{T}(\nu)$ if its probability density function is*

$$f(t \mid \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\nu\pi}} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \; t \in \mathbb{R}.$$

The basic t distribution is symmetric about 0 with thicker tails than the normal distribution.

As the degrees of freedom $\nu \to \infty$ the t density function approaches to the probability density function of a random variable $Z \sim \mathcal{N}(0,1)$.

Definition A.6 (Half-normal distribution)**.** *Let $Z \sim \mathcal{N}(0,\sigma^2)$ and $X = |Z|$. Then X is said to be half-normal distributed, denoted by $X \sim \mathcal{HN}(0,\sigma^2)$, with density*

$$f(x \mid 0,\sigma^2) = \frac{\sqrt{2}}{\sigma\sqrt{\pi}} e^{-\frac{x^2}{2\sigma^2}}\, 1_{\{x \geq 0\}},\ x \in \mathbb{R}.$$

The expectation and variance of a half-normal distributed random variable X are $\mathrm{E}(X) = \sigma\sqrt{\frac{2}{\pi}}$ and $\mathrm{Var}(X) = \sigma^2\left(1 - \frac{2}{\pi}\right)$, respectively. This definition of the univariate half-normal distribution will be later needed for the stochastic representation of the skew-normal distribution. The following definition is a generalization of the half-normal distribution to allow truncation at an arbitrary a.

Definition A.7 (Truncated normal distribution)**.** *Let $Z \sim \mathcal{N}(\mu,\sigma^2)$ and $X = Z1_{\{Z \geq a\}}$. Then X is said to be truncated normal distributed, denoted by $X \sim \mathcal{TN}(a,\mu,\sigma^2)$, with density*

$$f(x \mid a,\mu,\sigma^2) = \frac{f_Z(x \mid \mu,\sigma^2)}{1 - F_Z(a \mid \mu,\sigma^2)}\, 1_{\{x \geq a\}},\ x \in \mathbb{R},$$

where $f_Z(\cdot)$ is the density and $F_Z(\cdot)$ is the cumulative distribution function of the t distribution.

A.3 Multivariate distributions

Definition A.8 (Multivariate normal distribution)**.** *Let $\boldsymbol{\mu} \in \mathbb{R}^p$ and positive definite matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$. A random vector $\boldsymbol{Z}$ is said to be multivariate normal distributed with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$, denoted by $\boldsymbol{Z} \sim \mathcal{MVN}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ if its probability density function is*

$$f(\boldsymbol{z} \mid \boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}(2\pi)^{\frac{p}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})\right),\ \boldsymbol{z} \in \mathbb{R}^p,$$

where $|\cdot|$ denotes the determinant of the matrix $\boldsymbol{\Sigma}$.

Definition A.9 (Multivariate half-normal distribution)**.** *Let $\boldsymbol{Z} \sim \mathcal{MVN}(\boldsymbol{0},\boldsymbol{I}_p)$ and $\boldsymbol{X} = |\boldsymbol{Z}|$, where $|\cdot|$ denotes the componentwise absolute value of a multivariate random vector. Then $\boldsymbol{X}$ is said to be half-normal distributed, denoted by $\boldsymbol{X} \sim \mathcal{MHN}(\boldsymbol{0},\boldsymbol{I}_p)$, with density*

$$f(\boldsymbol{x} \mid \boldsymbol{0},\boldsymbol{I}_p) = \frac{2^{\frac{p}{2}}}{\pi^{\frac{p}{2}}} \exp\left(-\frac{1}{2}\boldsymbol{x}\boldsymbol{x}^\top\right)\, 1_{\{\boldsymbol{x} \geq \boldsymbol{0}\}},\ \boldsymbol{x} \in \mathbb{R}^p.$$

The expectation and variance of a multivariate half-normal distributed random variable $\boldsymbol{X}$ are $\mathrm{E}(\boldsymbol{X}) = \sqrt{\frac{2}{\pi}}$ and $\mathrm{Var}(\boldsymbol{X}) = \left(1 - \frac{2}{\pi}\right)\boldsymbol{I}_p$, respectively. This definition of the univariate half-normal distribution will be later needed for the stochastic representation of the skew-normal distribution.

Definition A.10 (Multivariate t distribution)**.** *Let $\nu \in (0, \infty)$, $\boldsymbol{\mu} \in \mathbb{R}^p$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$ a positive definite matrix. A random variable $\boldsymbol{T}$ is said to be multivariate t distributed with mean $\boldsymbol{\mu}$, scale $\boldsymbol{\Sigma}$ and degrees of freedom ν, denoted by $\boldsymbol{T} \sim \mathcal{MVT}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ if its probability density function is*

$$f(\boldsymbol{t} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})(\nu\pi)^{\frac{p}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \left(1 + \frac{1}{\nu}(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right)^{-\frac{\nu+p}{2}}, \ \boldsymbol{t} \in \mathbb{R}^p.$$

The multivariate t distribution is symmetric about $\boldsymbol{\mu}$ with thicker tails than the normal distribution. As the degrees of freedom $\nu \to \infty$ the t density function approaches to the probability density function of a random variable $\boldsymbol{Z} \sim \mathcal{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Definition A.11 (Truncated multivariate distribution)**.** *A p-dimensional random vector $\boldsymbol{X}$ is said to follow a p-variate truncated distribution if*

(i) $\boldsymbol{X} \in \mathbb{A} = \left\{\boldsymbol{x} = (x_1, \dots, x_p)^\top \mid x_1 \geq a_1, \dots x_p \geq a_p\right\}$*, i.e. $\boldsymbol{X}$ lies in the left truncated hyperplane region $\mathbb{A}$.*

(ii) the p-dimensional random vector $\boldsymbol{X} = (X_1, \dots, X_p)^\top$ has the density distribution

$$f(\boldsymbol{x} \mid \boldsymbol{\Theta}; \mathbb{A}) = \frac{f(\boldsymbol{x} \mid \boldsymbol{\Theta})}{\prod_{r=1}^{p} \int_{a_r}^{\infty} f(\boldsymbol{x} \mid \boldsymbol{\Theta}) d\boldsymbol{x}} \mathbf{1}_{\mathbb{A}},$$

where $\boldsymbol{\Theta}$ are the parameters of the distribution, $\mathbf{1}_{\mathbb{A}} = \mathbf{1}$ if $\mathbb{A}$ holds, $\mathbf{0}$ otherwise and $\prod_{r=1}^{p} \int_{a_r}^{\infty} f(\boldsymbol{x}) d\boldsymbol{x} = \int_{a_1}^{\infty} \dots \int_{a_p}^{\infty} f(\boldsymbol{x}) dx_p \dots dx_1$ an abbreviational notation of multiple integrals.

B Calculations

B.1 Proof of Lemma 2.3

Proof.

$$\begin{aligned}
&\int_0^\infty f_{Z|G=g}\left(x \,\middle|\, \mu, \frac{\sigma^2}{g}\right)\; f_G\left(g \,\middle|\, \frac{\nu}{2}, \frac{\nu}{2}\right)\; dg \\
&\quad= \int_0^\infty \frac{\sqrt{g}}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2 g}{2\sigma^2}} \left(\frac{\nu}{2}\right)^{\frac{\nu}{2}} \frac{1}{\Gamma\left(\frac{\nu}{2}\right)} g^{\frac{\nu}{2}-1} e^{-\frac{\nu}{2}g}\; dg \\
&\quad= \frac{1}{\sigma\sqrt{2\pi}} \frac{1}{\Gamma\left(\frac{\nu}{2}\right)} \left(\frac{\nu}{2}\right)^{\frac{\nu}{2}} \int_0^\infty g^{\frac{\nu+1}{2}-1} e^{-\left(\frac{(x-\mu)^2}{2\sigma^2}+\frac{\nu}{2}\right)g} dg \\
&\quad= \frac{1}{\sigma\sqrt{2\pi}} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \left(\frac{\nu}{2}\right)^{\frac{\nu}{2}} \left(\frac{\nu}{2}\left(1+\frac{(x-\mu)^2}{\nu\sigma^2}\right)\right)^{-\frac{\nu+1}{2}},
\end{aligned}$$

where the integral equals the normalizing constant of a $\mathcal{G}\left(\frac{\nu+1}{2}, \frac{(x-\mu)^2}{2\sigma^2}+\frac{\nu}{2}\right)$ distribution,

$$= \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sigma\sqrt{\pi\nu}} \left(1+\frac{(x-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}},$$

the $\mathcal{T}(\mu, \sigma^2, \nu)$ density function. □

B.2 Proof of Lemma 2.4

Proof.

$$F_T(t \mid 0, \sigma^2 s^2, \nu) = \int_{-\infty}^{t} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sigma s\sqrt{\pi\nu}} \left(1+\frac{x^2}{\nu(\sigma^2 s^2)}\right)^{-\frac{\nu+1}{2}} dx.$$

Using the substitution $z = \frac{x}{s}$ with $dx = sdz$ it follows that

$$= \int_{-\infty}^{\frac{t}{s}} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sigma\sqrt{\pi\nu}} \left(1+\frac{\frac{z^2}{\sigma^2}}{\nu}\right)^{-\frac{\nu+1}{2}} dz = F_T\left(\frac{t}{s} \,\middle|\, 0, \sigma^2, \nu\right).$$

□

B.3 Proof of Proposition 2.8

Proof. Let $X = \lambda Z + Y$, where $Z \sim \mathcal{HN}(0,1)$. Then conditional on $Z = z$, X follows a normal distribution with mean $\lambda z + \mu$ and variance σ^2. The density of X can be expressed as

$$\begin{aligned}
f_X(x) &= \int_{-\infty}^{\infty} f_{X|Z=z}(x \mid \lambda z + \mu, \sigma^2) f_Z(z \mid 0,1) dz \\
&= \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-(\lambda z+\mu))^2}{2\sigma^2}} \frac{2}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} 1_{\{z\geq 0\}}\, dz \\
&= \frac{1}{\sigma\pi} \int_0^{\infty} \exp\left(-\frac{\mu^2 - 2\mu x + x^2 + \lambda^2 z^2 + 2\lambda\mu z - 2\lambda z x + \sigma^2 z^2}{2\sigma^2} \right) dz \\
&= \frac{1}{\sigma\pi} \int_0^{\infty} \exp\left(-\frac{(x-\mu)^2 + z^2(\lambda^2+\sigma^2) + 2\lambda z(\mu - x))}{2\sigma^2} \right) dz \\
&= \frac{1}{\sigma\pi} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \int_0^{\infty} e^{-\frac{\lambda^2+\sigma^2}{2\sigma^2}\left(z^2 - 2z\frac{\lambda(x-\mu)}{\lambda^2+\sigma^2}\right)} dz \\
&= \frac{1}{\sigma\pi} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} e^{\frac{\lambda^2+\sigma^2}{2\sigma^2}\frac{\lambda^2(x-\mu)^2}{(\lambda^2+\sigma^2)^2}} \underbrace{\int_0^{\infty} e^{-\frac{\lambda^2+\sigma^2}{2\sigma^2}\left(z - \frac{\lambda(x-\mu)}{\lambda^2+\sigma^2}\right)^2} dz}_{(\star)} .
\end{aligned}$$

Let $u = z - \frac{\lambda(x-\mu)}{\lambda^2+\sigma^2}$. Then $du = dz$, and the integral $(\star)$ is

$$\begin{aligned}
&\int_{-\frac{\lambda(x-\mu)}{\lambda^2+\sigma^2}}^{\infty} e^{-\frac{\lambda^2+\sigma^2}{2\sigma^2}u^2} du = \int_{-\infty}^{\frac{\lambda(x-\mu)}{\lambda^2+\sigma^2}} e^{-\frac{\lambda^2+\sigma^2}{2\sigma^2}u^2} du \text{ (by symmetry)} \\
&= \sqrt{2\pi}\sqrt{\frac{\sigma^2}{\lambda^2+\sigma^2}} F_Z\left(\frac{\lambda(x-\mu)}{\lambda^2+\sigma^2} \,\middle|\, 0, \frac{\sigma^2}{\lambda^2+\sigma^2} \right).
\end{aligned}$$

Hence, inserting this for $(\star)$ yields

$$\begin{aligned}
&\frac{1}{\sigma\pi} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} e^{\frac{\lambda^2+\sigma^2}{2\sigma^2}\frac{\lambda^2}{(\lambda^2+\sigma^2)^2}(x-\mu)^2} \sqrt{2\pi}\sqrt{\frac{\sigma^2}{\lambda^2+\sigma^2}} F_Z\left(\frac{\lambda(x-\mu)}{\lambda^2+\sigma^2} \,\middle|\, 0, \frac{\sigma^2}{\lambda^2+\sigma^2} \right) \\
&\qquad = \sqrt{\frac{2}{\pi(\lambda^2+\sigma^2)}} e^{-\frac{1}{2(\lambda^2+\sigma^2)}(x-\mu)^2} F_Z\left(\frac{\lambda(x-\mu)}{\lambda^2+\sigma^2} \,\middle|\, 0, \frac{\sigma^2}{\lambda^2+\sigma^2} \right) \\
&\qquad = 2 f_Z(x \mid \mu, \lambda^2+\sigma^2) F_Z\left(\frac{\lambda(x-\mu)}{\lambda^2+\sigma^2} \,\middle|\, 0, \frac{\sigma^2}{\lambda^2+\sigma^2} \right).
\end{aligned}$$

□

B.4 Proof of Theorem 2.10

The following proposition is an extension from Lin (2010) that will be used to derive the probability density function of the skew-t distribution from Def. 2.9.

Proposition B.1. *Let $G \sim \mathcal{G}(\alpha, \beta)$, $\alpha, \beta > 0$. Then for all $a \in \mathbb{R}$,*

$$E_G\left(F_Z(a\sqrt{G} \mid 0, \sigma^2)\right) = F_T\left(a\sqrt{\frac{\alpha}{\beta}} \,\middle|\, 0, \sigma^2, 2\alpha\right),$$

where $F_Z(z \mid 0, \sigma^2)$ is the cdf of a random variable $Z \sim \mathcal{N}(0, \sigma^2)$, $F_T(t \mid 0, \sigma^2, 2\alpha)$ is the cdf of a random variable $T \sim \mathcal{T}(0, \sigma^2, 2\alpha)$, and Z and G are independent.

Proof. First note that

$$E_G(F_Z(a\sqrt{G} \mid 0, \sigma^2)) = E_G\left(P(Z \leq a\sqrt{G})\right) = E_G\left(P\left(\frac{Z}{\sqrt{G}}\sqrt{\frac{\alpha}{\beta}} \leq a\sqrt{\frac{\alpha}{\beta}}\right)\right).$$

Let $W = \frac{Z\sqrt{\alpha}}{\sqrt{G\beta}}$. Conditional on $G = g$, W follows a normal distribution with mean 0 and variance $\frac{\alpha}{\beta g}\sigma^2$. Writing the expectation as an integral yields

$$\begin{aligned}
&\int_{-\infty}^{\infty} \frac{\beta^\alpha}{\Gamma(\alpha)} g^{\alpha-1} e^{-\beta g}\, 1_{\{g \geq 0\}} \underbrace{\int_{-\infty}^{a\sqrt{\frac{\alpha}{\beta}}} \frac{\sqrt{\beta g}}{\sqrt{2\pi\alpha}\sigma} e^{-\frac{w^2\beta g}{2\alpha\sigma^2}} dw}_{P(W \leq a\sqrt{\alpha/\beta})}\, dg \\
&= \int_0^\infty \int_{-\infty}^{a\sqrt{\frac{\alpha}{\beta}}} \frac{\beta^\alpha}{\Gamma(\alpha)} g^{\alpha-1} e^{-\beta g} \frac{\sqrt{\beta g}}{\sqrt{2\pi\alpha}\sigma} e^{-\frac{w^2\beta g}{2\alpha\sigma^2}} dw\, dg.
\end{aligned}$$

Now changing the integration order using Fubini's theorem, it follows that

$$\begin{aligned}
&= \int_{-\infty}^{a\sqrt{\frac{\alpha}{\beta}}} \int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)} g^{\alpha-1} e^{-\beta g} \frac{\sqrt{\beta g}}{\sqrt{2\pi\alpha}\sigma} e^{-\frac{w^2\beta g}{2\alpha\sigma^2}} dg\, dw \\
&= \int_{-\infty}^{a\sqrt{\frac{\alpha}{\beta}}} \frac{\beta^{\frac{2\alpha+1}{2}}}{\Gamma(\alpha)\sqrt{2\pi\alpha}\sigma} \int_0^\infty g^{\frac{2\alpha+1}{2}-1} e^{-\left(\frac{w^2\beta}{2\alpha\sigma^2}+\beta\right)g} dg\, dw \\
&= \int_{-\infty}^{a\sqrt{\frac{\alpha}{\beta}}} \frac{\beta^{\frac{2\alpha+1}{2}}\Gamma\left(\frac{2\alpha+1}{2}\right)}{\Gamma(\alpha)\sqrt{2\pi\alpha}\sigma} \left(\frac{w^2\beta}{2\alpha\sigma^2}+\beta\right)^{-\frac{2\alpha+1}{2}} dw,
\end{aligned}$$

where the latter is the normalizing constant of a $\mathcal{G}\left(\frac{2\alpha+1}{2}, \frac{w^2\beta}{2\alpha\sigma^2}+\beta\right)$ distribution. The result equals

$$\int_{-\infty}^{a\sqrt{\frac{\alpha}{\beta}}} \frac{\Gamma\left(\frac{2\alpha+1}{2}\right)}{\Gamma(\alpha)\sigma\sqrt{\pi 2\alpha}} \left(\frac{w^2}{2\alpha\sigma^2}+1\right)^{-\frac{2\alpha+1}{2}} dw = F_T\left(a\sqrt{\frac{\alpha}{\beta}} \,\middle|\, 0, \sigma^2, 2\alpha\right).$$

□

Proof. Let X, Y and G be as in Def. 2.9. As a first step we look at the distribution of X conditional on $G = g$. According to Prop. 2.8

$$X \mid (G = g) = \mu + \frac{1}{\sqrt{g}}(\lambda Z_0 + Z_1) = \frac{\lambda}{\sqrt{g}} Z_0 + \underbrace{\mu + \frac{1}{\sqrt{g}} Z_1}_{\sim\mathcal{N}\left(\mu, \frac{1}{g}\sigma^2\right)},$$

where $Z_0 \sim \mathcal{HN}(0,1)$ and $Z_1 \sim \mathcal{N}(0,\sigma^2)$. The normality result follows, as for constants $a, b \in \mathbb{R}$ and $W \sim \mathcal{N}(0,\sigma^2)$, $a + bW \sim \mathcal{N}(a, b^2\sigma^2)$. By Prop. 2.8 it follows that $X \mid (G = g) \sim \mathcal{SN}\left(\mu, \frac{\sigma^2}{g}, \frac{\lambda}{\sqrt{g}}\right)$. We next use the definition of the skew-normal density to integrate out G:

$$f(x) = \int_0^\infty f_{X|G=g}\left(x \,\middle|\, \mu, \frac{\sigma^2}{g}, \frac{\lambda}{\sqrt{g}}\right) f_G\left(g \,\middle|\, \frac{\nu}{2}, \frac{\nu}{2}\right)\, dg. \tag{B.1}$$

Using the definition of the skew-normal distribution, it follows that

$$f(x) = \int_0^\infty 2 f_Z\left(x \,\middle|\, \mu, \frac{\sigma^2+\lambda^2}{g}\right) F_Z\left(\frac{\sqrt{g}\lambda(x-\mu)}{\sigma^2+\lambda^2} \,\middle|\, 0, \frac{\sigma^2}{\sigma^2+\lambda^2}\right) f_G\left(g \,\middle|\, \frac{\nu}{2}, \frac{\nu}{2}\right)\, dg. \tag{B.2}$$

Considering only the density function of the normal and gamma distribution ($f_Z(\cdot)$ and $f_G(\cdot)$) of the equation it follows that

$$\begin{aligned} f_Z\left(x \,\middle|\, \mu, \frac{\sigma^2+\lambda^2}{g}\right) f_G\left(g \,\middle|\, \frac{\nu}{2}, \frac{\nu}{2}\right) \\ &= \frac{\sqrt{g}}{\sqrt{2\pi(\sigma^2+\lambda^2)}} e^{-\frac{g}{2(\sigma^2+\lambda^2)}(x-\mu)^2} \left(\frac{\nu}{2}\right)^{\frac{\nu}{2}} \frac{1}{\Gamma\left(\frac{\nu}{2}\right)} g^{\frac{\nu}{2}-1} e^{-\frac{\nu}{2}g} \\ &= \frac{1}{\sqrt{2\pi(\sigma^2+\lambda^2)}} \left(\frac{\nu}{2}\right)^{\frac{\nu}{2}} \frac{1}{\Gamma\left(\frac{\nu}{2}\right)} g^{\frac{\nu+1}{2}-1} e^{-\left(\frac{(x-\mu)^2}{2(\sigma^2+\lambda^2)}+\frac{\nu}{2}\right)g}. \end{aligned}$$

In terms of g, this is proportional to a gamma distribution with $\alpha = \frac{\nu+1}{2}$ and $\beta = \frac{1}{2}\left(\frac{(x-\mu)^2}{\sigma^2+\lambda^2} + \nu\right)$. Hence we replace the last two terms with $\Gamma(\alpha)\beta^{-\alpha} f_G(g \mid \alpha, \beta)$ to obtain

$$\begin{aligned} &\frac{\Gamma\left(\frac{\nu+1}{2}\right)\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{2\pi(\sigma^2+\lambda^2)}} \left(\frac{\nu}{2}\left(\frac{(x-\mu)^2}{\nu(\sigma^2+\lambda^2)}+1\right)\right)^{-\frac{\nu+1}{2}} f_G\left(g \,\middle|\, \frac{\nu+1}{2}, \frac{1}{2}\left(\frac{(x-\mu)^2}{\sigma^2+\lambda^2}+\nu\right)\right) \\ &= \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{(\sigma^2+\lambda^2)\pi\nu}} \left(\frac{(x-\mu)^2}{\nu(\sigma^2+\lambda^2)}+1\right)^{-\frac{\nu+1}{2}} f_G\left(g \,\middle|\, \frac{\nu+1}{2}, \frac{1}{2}\left(\frac{(x-\mu)^2}{\sigma^2+\lambda^2}+\nu\right)\right) \\ &= f_T\left(x \mid \mu, \sigma^2+\lambda^2, \nu\right) f_G\left(g \,\middle|\, \frac{\nu+1}{2}, \frac{1}{2}\left(\frac{(x-\mu)^2}{\sigma^2+\lambda^2}+\nu\right)\right). \end{aligned}$$

Inserting this result in equation B.2 yields

$$\begin{aligned} f(x) = 2 f_T\left(x \mid \mu, \sigma^2+\lambda^2, \nu\right) \int_0^\infty f_G\left(g \,\middle|\, \frac{\nu+1}{2}, \frac{1}{2}\left(\frac{(x-\mu)^2}{\sigma^2+\lambda^2}+\nu\right)\right) \\ \cdot F_Z\left(\frac{\sqrt{g}\lambda(x-\mu)}{\sigma^2+\lambda^2} \,\middle|\, 0, \frac{\sigma^2}{\sigma^2+\lambda^2}\right) dg. \end{aligned}$$

Applying Prop. B.1 with $a = \frac{\lambda(x-\mu)}{\sigma^2+\lambda^2}$ to the integral yields:

$$f(x) = 2f_T\left(x \mid \mu, \sigma^2+\lambda^2, \nu\right) F_T\left(\frac{\lambda(x-\mu)}{\sigma^2+\lambda^2}\sqrt{\frac{\nu+1}{\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}}}\;\middle|\; 0, \frac{\sigma^2}{\sigma^2+\lambda^2}, \nu+1\right).$$

□

B.5 Proof of Proposition 2.19

Proof. Let $\boldsymbol{X} = \boldsymbol{\Lambda}\boldsymbol{Z} + \boldsymbol{Y}$, where $\boldsymbol{Z} \sim \mathcal{MHN}(\boldsymbol{0}, \boldsymbol{I}_p)$ and $Y \sim \mathcal{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The multiple integral is denoted in the following with the dimension of the integral in the subscript and using Euclidean space. Then conditional on $\boldsymbol{Z} = \boldsymbol{z}$, $\boldsymbol{X}$ follows a normal distribution with mean $\boldsymbol{\Lambda z} + \boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$. The density of $\boldsymbol{X}$ can be expressed as

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \int_{\mathbb{R}^p} f_{\boldsymbol{X}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{x} \mid \boldsymbol{\Lambda z} + \boldsymbol{\mu}, \boldsymbol{\Sigma}) f_{\boldsymbol{Z}}(\boldsymbol{z} \mid \boldsymbol{0}, \boldsymbol{I}_p) d\boldsymbol{z}.$$

Hence inserting the definitions of the pdfs yields

$$\begin{aligned}
&= \int_{\mathbb{R}^p} \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}(2\pi)^{\frac{p}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - (\boldsymbol{\Lambda z} + \boldsymbol{\mu}))^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - (\boldsymbol{\Lambda z} + \boldsymbol{\mu}))\right) \\
&\qquad \cdot \frac{2^p}{(2\pi)^{\frac{p}{2}}} \exp\left(-\frac{1}{2}\boldsymbol{z}^\top \boldsymbol{z}\right) 1_{\{\boldsymbol{z} \geq \boldsymbol{0}\}}\, d\boldsymbol{z} \\
&= \int_{\mathbb{R}^p_+} \frac{2^p}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}(2\pi)^p} \exp\Big(-\frac{1}{2}((\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) + (\boldsymbol{\Lambda z})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\Lambda z}) \\
&\qquad - (\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda z} - \boldsymbol{\Lambda z}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) + \boldsymbol{z}^\top \boldsymbol{z})\Big) d\boldsymbol{z} \\
&= \int_{\mathbb{R}^p_+} \exp\left(-\frac{1}{2}((\boldsymbol{\Lambda z})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\Lambda z}) - (\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda z} - \boldsymbol{\Lambda z}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) + \boldsymbol{z}^\top \boldsymbol{z})\right) d\boldsymbol{z} \\
&\qquad \cdot \frac{2^p}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}(2\pi)^p} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right).
\end{aligned}$$

Extending the expression results in

$$\begin{aligned}
&\int_{\mathbb{R}^p_+} \exp\left(-\frac{(\boldsymbol{z} - \boldsymbol{\Lambda}(\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))^\top (\boldsymbol{I}_p + \boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})(\boldsymbol{z} - \boldsymbol{\Lambda}(\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))}{2}\right) d\boldsymbol{z} \\
&\qquad \cdot \frac{2^p}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}(2\pi)^p} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \\
&\qquad \cdot \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top (\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)^{-1}\boldsymbol{\Lambda}(\boldsymbol{I}_p + \boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})\boldsymbol{\Lambda}(\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right).
\end{aligned}$$

The function in the multiple integral is a multivariate normal distribution $\mathcal{MVN}(\boldsymbol{\Lambda}(\boldsymbol{\Sigma} + \boldsymbol{\Lambda}^2)^{-1}(\boldsymbol{x} - \boldsymbol{\mu}), (\boldsymbol{I}_p + \boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})^{-1})$ without the normalizing constant. Hence it follows that

$$
\begin{aligned}
&= (2\pi)^{\frac{p}{2}} |(\boldsymbol{I}_p + \boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})^{-1}|^{\frac{1}{2}} F_{\boldsymbol{Z}}(\boldsymbol{\Lambda}(\boldsymbol{\Sigma} + \boldsymbol{\Lambda}^2)^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \mid \boldsymbol{0}, (\boldsymbol{I}_p + \boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})^{-1}) \\
&\qquad \cdot \frac{2^p}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}(2\pi)^p} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right) \\
&\qquad \cdot \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top (\boldsymbol{\Sigma} + \boldsymbol{\Lambda}^2)^{-1}\boldsymbol{\Lambda}(\boldsymbol{I}_p + \boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})\boldsymbol{\Lambda}(\boldsymbol{\Sigma} + \boldsymbol{\Lambda}^2)^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right) \\
&= \frac{2^p |(\boldsymbol{I}_p + \boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})^{-1}|^{\frac{1}{2}}}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}(2\pi)^{\frac{p}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right) \\
&\qquad \cdot \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Lambda}(\boldsymbol{\Sigma} + \boldsymbol{\Lambda}^2)^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}(\boldsymbol{x} - \boldsymbol{\mu})\right) \\
&\qquad \cdot F_{\boldsymbol{Z}}(\boldsymbol{\Lambda}(\boldsymbol{\Sigma} + \boldsymbol{\Lambda}^2)^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \mid \boldsymbol{0}, (\boldsymbol{I}_p + \boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})^{-1}) \\
&= \frac{2^p}{|\boldsymbol{I}_p + \boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}|^{\frac{1}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}} (2\pi)^{\frac{p}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top (\boldsymbol{\Sigma} + \boldsymbol{\Lambda}^2)^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right) \\
&\qquad \cdot F_{\boldsymbol{Z}}(\boldsymbol{\Lambda}(\boldsymbol{\Sigma} + \boldsymbol{\Lambda}^2)^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \mid \boldsymbol{0}, (\boldsymbol{I}_p + \boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})^{-1}) \\
&= \frac{2^p}{|\boldsymbol{\Sigma}^{-1}|^{\frac{1}{2}} |\boldsymbol{\Sigma} + \boldsymbol{\Lambda}^2|^{\frac{1}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}} (2\pi)^{\frac{p}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top (\boldsymbol{\Sigma} + \boldsymbol{\Lambda}^2)^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right) \\
&\qquad \cdot F_{\boldsymbol{Z}}(\boldsymbol{\Lambda}(\boldsymbol{\Sigma} + \boldsymbol{\Lambda}^2)^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \mid \boldsymbol{0}, (\boldsymbol{I}_p + \boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})^{-1}) \\
&= \frac{2^p}{|\boldsymbol{\Sigma} + \boldsymbol{\Lambda}^2|^{\frac{1}{2}} (2\pi)^{\frac{p}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top (\boldsymbol{\Sigma} + \boldsymbol{\Lambda}^2)^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right) \\
&\qquad \cdot F_{\boldsymbol{Z}}(\boldsymbol{\Lambda}(\boldsymbol{\Sigma} + \boldsymbol{\Lambda}^2)^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \mid \boldsymbol{0}, (\boldsymbol{I}_p + \boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})^{-1}) \\
&= 2^p f_{\boldsymbol{Z}}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma} + \boldsymbol{\Lambda}^2) F_{\boldsymbol{Z}}(\boldsymbol{\Lambda}(\boldsymbol{\Sigma} + \boldsymbol{\Lambda}^2)^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \mid \boldsymbol{0}, (\boldsymbol{I}_p + \boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})^{-1}).
\end{aligned}
$$

Now using the representation of Def. 2.18, $\boldsymbol{\Omega} = \boldsymbol{\Sigma} + \boldsymbol{\Lambda}^2$ and $\boldsymbol{\Delta} = \boldsymbol{I}_p - \boldsymbol{\Lambda}\boldsymbol{\Omega}^{-1}\boldsymbol{\Lambda}$

$$
= 2^p f_{\boldsymbol{Z}}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Omega}) F_{\boldsymbol{Z}}\left(\boldsymbol{\Lambda}\boldsymbol{\Omega}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \,\middle|\, \boldsymbol{0}, \boldsymbol{\Delta}\right),
$$

since $\boldsymbol{\Delta} = \boldsymbol{I}_p - \boldsymbol{\Lambda}\boldsymbol{\Omega}^{-1}\boldsymbol{\Lambda} \overset{(\star)}{=} (\boldsymbol{I}_p + \boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})^{-1}$. We proof $(\star)$ by showing that $\boldsymbol{I}_p = \boldsymbol{\Delta}\boldsymbol{\Delta}^{-1}$ with $\boldsymbol{\Delta}^{-1} = \boldsymbol{I}_p + \boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}$ and $\boldsymbol{\Delta} = \boldsymbol{I}_p - \boldsymbol{\Lambda}\boldsymbol{\Omega}^{-1}\boldsymbol{\Lambda}$. Hence we get

$$
\begin{aligned}
&(\boldsymbol{I}_p - \boldsymbol{\Lambda}\boldsymbol{\Omega}^{-1}\boldsymbol{\Lambda})(\boldsymbol{I}_p + \boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}) = \boldsymbol{I}_p - \boldsymbol{\Lambda}\boldsymbol{\Omega}^{-1}\boldsymbol{\Lambda} + \boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} - \boldsymbol{\Lambda}\boldsymbol{\Omega}^{-1}\boldsymbol{\Lambda}^2\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} \\
&= \boldsymbol{I}_p - \boldsymbol{\Lambda}\boldsymbol{\Omega}^{-1}(\boldsymbol{I}_p + \boldsymbol{\Lambda}^2\boldsymbol{\Sigma}^{-1})\boldsymbol{\Lambda} + \boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} = \boldsymbol{I}_p - \boldsymbol{\Lambda}\boldsymbol{\Omega}^{-1}(\boldsymbol{\Sigma} + \boldsymbol{\Lambda}^2)\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} + \boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} \\
&= \boldsymbol{I}_p - \boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} + \boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} = \boldsymbol{I}_p.
\end{aligned}
$$

□

B.6 Proof of Theorem 2.21

Proposition B.2. *Let $G \sim \mathcal{G}(\alpha, \beta)$, $\mathbf{\Sigma} \in \mathbb{R}^{p \times p}$ positive definite covariance matrix. Then for any $\boldsymbol{a} \in \mathbb{R}^p$,*

$$E_G\left(F_{\boldsymbol{Z}}(\boldsymbol{a}\sqrt{G} \mid \mathbf{0}, \mathbf{\Sigma})\right) = F_{\boldsymbol{T}}\left(\boldsymbol{a}\sqrt{\frac{\alpha}{\beta}} \,\middle|\, \mathbf{0}, \mathbf{\Sigma}, 2\alpha\right),$$

where $F_{\boldsymbol{Z}}(\boldsymbol{z} \mid \mathbf{0}, \mathbf{\Sigma})$ is the cdf of a random variable $\boldsymbol{Z} \sim \mathcal{MVN}(\mathbf{0}, \mathbf{\Sigma})$, $F_{\boldsymbol{T}}(\boldsymbol{t} \mid \mathbf{0}, \mathbf{\Sigma}, 2\alpha)$ is the cdf of a random variable $\boldsymbol{T} \sim \mathcal{MVT}(\mathbf{0}, \mathbf{\Sigma}, 2\alpha)$, and $\boldsymbol{Z}$ and G are independent.

Proof. The proof follows the same steps as the univariate case, i.e. the Proof of Theorem 2.10. First note that

$$E_G(F_{\boldsymbol{Z}}(\boldsymbol{a}\sqrt{G} \mid \mathbf{0}, \mathbf{\Sigma})) = E_G\left(P(\boldsymbol{Z} \leq \boldsymbol{a}\sqrt{G})\right) = E_G\left(P\left(\frac{\boldsymbol{Z}}{\sqrt{G}}\sqrt{\frac{\alpha}{\beta}} \leq \boldsymbol{a}\sqrt{\frac{\alpha}{\beta}}\right)\right).$$

Let $\boldsymbol{W} = \frac{\boldsymbol{Z}\sqrt{\alpha}}{\sqrt{G\beta}}$. Conditional on $G = g$, W follows a normal distribution with mean $\mathbf{0}$ and variance $\frac{\alpha}{\beta g}\mathbf{\Sigma}$. Writing the expectation as an integral with $\boldsymbol{a} = (a_1, \ldots, a_p)^\top$ yields

$$\begin{aligned}
&\int_{-\infty}^{\infty} \frac{\beta^\alpha}{\Gamma(\alpha)} g^{\alpha-1} e^{-\beta g}\, 1_{\{g \geq 0\}} \prod_{i=1}^{p} \int_{-\infty}^{a_i\sqrt{\frac{\alpha}{\beta}}} \frac{\sqrt{\beta g}}{(2\pi\alpha)^{\frac{p}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{\beta g}{2\alpha}\boldsymbol{w}^\top \mathbf{\Sigma}^{-1}\boldsymbol{w}\right) d\boldsymbol{w}\, dg \\
&= \int_0^{\infty} \prod_{i=1}^{p} \int_{-\infty}^{a_i\sqrt{\frac{\alpha}{\beta}}} \frac{\beta^\alpha}{\Gamma(\alpha)} g^{\alpha-1} e^{-\beta g} \frac{\sqrt{\beta g}}{(2\pi\alpha)^{\frac{p}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{\beta g}{2\alpha}\boldsymbol{w}^\top \mathbf{\Sigma}^{-1}\boldsymbol{w}\right) d\boldsymbol{w}\, dg.
\end{aligned}$$

Now changing the integration order using Fubini's theorem, it follows that

$$\begin{aligned}
&= \prod_{i=1}^{p} \int_{-\infty}^{a_i\sqrt{\frac{\alpha}{\beta}}} \int_0^{\infty} \frac{\beta^\alpha}{\Gamma(\alpha)} g^{\alpha-1} e^{-\beta g} \frac{\sqrt{\beta g}}{(2\pi\alpha)^{\frac{p}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{\beta g}{2\alpha}\boldsymbol{w}^\top \mathbf{\Sigma}^{-1}\boldsymbol{w}\right) dg\, d\boldsymbol{w} \\
&= \prod_{i=1}^{p} \int_{-\infty}^{a_i\sqrt{\frac{\alpha}{\beta}}} \frac{\beta^{\frac{2\alpha+1}{2}}}{\Gamma(\alpha)(2\pi\alpha)^{\frac{p}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}} \int_0^{\infty} g^{\frac{2\alpha+1}{2}-1} \exp\left(-\left(\frac{\beta}{2\alpha}\boldsymbol{w}^\top \mathbf{\Sigma}^{-1}\boldsymbol{w} + \beta\right) g\right) dg\, d\boldsymbol{w} \\
&= \prod_{i=1}^{p} \int_{-\infty}^{a_i\sqrt{\frac{\alpha}{\beta}}} \frac{\beta^{\frac{2\alpha+1}{2}}\Gamma\left(\frac{2\alpha+1}{2}\right)}{\Gamma(\alpha)(2\pi\alpha)^{\frac{p}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}} \left(\frac{\beta}{2\alpha}\boldsymbol{w}^\top \mathbf{\Sigma}^{-1}\boldsymbol{w} + \beta\right)^{-\frac{2\alpha+1}{2}} d\boldsymbol{w},
\end{aligned}$$

where the latter is the normalizing constant of a $\mathcal{G}\left(\frac{2\alpha+1}{2}, \frac{\beta g}{2\alpha}\boldsymbol{w}^\top \mathbf{\Sigma}^{-1}\boldsymbol{w} + \beta\right)$ distribution. The result equals

$$\begin{aligned}
&= \prod_{i=1}^{p} \int_{-\infty}^{a_i\sqrt{\frac{\alpha}{\beta}}} \frac{\Gamma\left(\frac{2\alpha+1}{2}\right)}{\Gamma(\alpha)|\mathbf{\Sigma}|^{\frac{1}{2}}(\pi 2\alpha)^{\frac{p}{2}}} \left(\frac{1}{2\alpha}\boldsymbol{w}^\top \mathbf{\Sigma}^{-1}\boldsymbol{w} + 1\right)^{-\frac{2\alpha+1}{2}} d\boldsymbol{w} \\
&= F_{\boldsymbol{T}}\left(\boldsymbol{a}\sqrt{\frac{\alpha}{\beta}} \,\middle|\, \mathbf{0}, \mathbf{\Sigma}, 2\alpha\right).
\end{aligned}$$

□

Proof. Let $\boldsymbol{X}, \boldsymbol{Y}$ and G be as in Def. 2.20. As a first step we look at the distribution of $\boldsymbol{X}$ conditional on $G = g$. According to Prop. 2.19

$$\boldsymbol{X} \mid (G = g) = \boldsymbol{\mu} + \frac{1}{\sqrt{g}}(\boldsymbol{\lambda} \boldsymbol{Z}_0 + \boldsymbol{Z}_1) = \frac{\boldsymbol{\lambda}}{\sqrt{g}} \boldsymbol{Z}_0 + \underbrace{\boldsymbol{\mu} + \frac{1}{\sqrt{g}} \boldsymbol{Z}_1}_{\sim \mathcal{MVN}\left(\boldsymbol{\mu}, \frac{1}{g}\boldsymbol{\Sigma}\right)} ,$$

where $\boldsymbol{Z}_0 \sim \mathcal{MHN}(\boldsymbol{0}, \boldsymbol{I}_p)$ and $\boldsymbol{Z}_1 \sim \mathcal{MVN}(\boldsymbol{0}, \boldsymbol{\Sigma})$. The normal result follows, since for constants $\boldsymbol{a} \in \mathbb{R}^p$, $b \in \mathbb{R}$ and $\boldsymbol{W} \sim \mathcal{MVN}(\boldsymbol{0}, \boldsymbol{\Sigma})$, $\boldsymbol{a} + b\boldsymbol{W} \sim \mathcal{MVN}(\boldsymbol{a}, b^2\boldsymbol{\Sigma})$. By Prop. 2.19 it follows that $\boldsymbol{X} \mid (G = g) \sim \mathcal{MSN}\left(\boldsymbol{\mu}, \frac{1}{g}\boldsymbol{\Sigma}, \frac{1}{\sqrt{g}}\boldsymbol{\lambda}\right)$. We next use the definition of the skew-normal density to integrate out G:

$$f(\boldsymbol{x}) = \int_0^\infty f_{\boldsymbol{X}|G=g}\left(\boldsymbol{x} \,\middle|\, \boldsymbol{\mu}, \frac{1}{g}\boldsymbol{\Sigma}, \frac{1}{\sqrt{g}}\boldsymbol{\lambda}\right) f_G\left(g \,\middle|\, \frac{\nu}{2}, \frac{\nu}{2}\right) \, dg.$$

Using Def. 2.18, it follows that

$$\begin{aligned} f(\boldsymbol{x}) = \int_0^\infty & 2^p f_{\boldsymbol{Z}}\left(\boldsymbol{x} \,\middle|\, \boldsymbol{\mu}, \frac{1}{g}(\boldsymbol{\Sigma} + \boldsymbol{\Lambda}^2)\right) \\ & \cdot F_{\boldsymbol{Z}}\left(\sqrt{g}\boldsymbol{\Lambda}(\boldsymbol{\Sigma} + \boldsymbol{\Lambda}^2)^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \,\middle|\, \boldsymbol{0}, \boldsymbol{\Delta}\right) f_G\left(g \,\middle|\, \frac{\nu}{2}, \frac{\nu}{2}\right) \, dg, \end{aligned}$$

where $\boldsymbol{\Delta} = \boldsymbol{I}_p - \frac{1}{g}\boldsymbol{\Lambda}(\boldsymbol{\Sigma} + \boldsymbol{\Lambda}^2)^{-1}\boldsymbol{\Lambda}$. Considering only the density function of the multivariate normal and gamma distributions ($f_{\boldsymbol{Z}}(\cdot)$ and $f_G(\cdot)$) and that $|\alpha \boldsymbol{A}| = \alpha^p |\boldsymbol{A}|$ for $\alpha \in \mathbb{R}$, $\boldsymbol{A} \in \mathbb{R}^p$, it follows that

$$\begin{aligned} & f_{\boldsymbol{Z}}\left(\boldsymbol{x} \,\middle|\, \boldsymbol{\mu}, \frac{1}{g}(\boldsymbol{\Sigma} + \boldsymbol{\Lambda}^2)\right) f_G\left(g \,\middle|\, \frac{\nu}{2}, \frac{\nu}{2}\right) \\ & = \frac{g^{\frac{p}{2}}}{(2\pi)^{\frac{p}{2}} |\boldsymbol{\Sigma} + \boldsymbol{\Lambda}^2|^{\frac{1}{2}}} \exp\left(-g(\boldsymbol{x} - \boldsymbol{\mu})^\top (\boldsymbol{\Sigma} + \boldsymbol{\Lambda}^2)^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right) \left(\frac{\nu}{2}\right)^{\frac{\nu}{2}} \frac{1}{\Gamma\left(\frac{\nu}{2}\right)} g^{\frac{\nu}{2}-1} e^{-\frac{\nu}{2}g} \\ & = \frac{\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}}{(2\pi)^{\frac{p}{2}} |\boldsymbol{\Sigma} + \boldsymbol{\Lambda}^2|^{\frac{1}{2}}} \frac{1}{\Gamma\left(\frac{\nu}{2}\right)} g^{\frac{\nu+p}{2}-1} \exp\left(-\left(\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top (\boldsymbol{\Sigma} + \boldsymbol{\Lambda}^2)^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) + \frac{\nu}{2}\right) g\right). \end{aligned}$$

In terms of g, this is proportional to a gamma distribution with $\alpha = \frac{\nu+p}{2}$ and $\beta = \frac{q(\boldsymbol{x})+\nu}{2}$ with $q(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{\mu})^\top (\boldsymbol{\Sigma} + \boldsymbol{\Lambda}^2)^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$. Hence, we replace the last two terms with $\Gamma(\alpha)\beta^{-\alpha} f_G(g \mid \alpha, \beta)$ to obtain

$$\begin{aligned} & \frac{\Gamma\left(\frac{\nu+p}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(2\pi)^{\frac{p}{2}} |\boldsymbol{\Sigma} + \boldsymbol{\Lambda}^2|^{\frac{1}{2}}} \left(\frac{\nu}{2}\right)^{\frac{\nu}{2}} \left(\frac{\nu}{2}\left(\frac{q(\boldsymbol{x})}{\nu} + 1\right)\right)^{-\frac{\nu+p}{2}} f_G\left(g \,\middle|\, \frac{\nu+p}{2}, \frac{q(\boldsymbol{x})+\nu}{2}\right) \\ & = \frac{\Gamma\left(\frac{\nu+p}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\pi\nu)^{\frac{p}{2}} |\boldsymbol{\Sigma} + \boldsymbol{\Lambda}^2|^{\frac{1}{2}}} \left(q(\boldsymbol{x}) + 1\right)^{-\frac{\nu+p}{2}} f_G\left(g \,\middle|\, \frac{\nu+p}{2}, \frac{q(\boldsymbol{x})+\nu}{2}\right) \\ & = f_{\boldsymbol{T}}\left(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma} + \boldsymbol{\Lambda}^2, \nu\right) f_G\left(g \,\middle|\, \frac{\nu+p}{2}, \frac{q(\boldsymbol{x})+\nu}{2}\right). \end{aligned}$$

Inserting this result in the original equation yields

$$f(\boldsymbol{x}) = 2^p f_{\boldsymbol{T}}\left(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma} + \boldsymbol{\Lambda}, \nu\right) \int_0^\infty f_G\left(g \,\middle|\, \frac{\nu+p}{2}, \frac{q(\boldsymbol{x})+\nu}{2}\right) \\ \cdot F_{\boldsymbol{Z}}\left(\sqrt{g}\boldsymbol{\Lambda}(\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \,\middle|\, 0, \boldsymbol{\Delta}\right) dg.$$

Applying Prop. B.2 with $\boldsymbol{a} = \boldsymbol{\Lambda}(\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$ to the integral yields:

$$f(\boldsymbol{x}) = 2^p f_{\boldsymbol{T}}\left(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2, \nu\right) F_{\boldsymbol{T}}\left(\boldsymbol{\Lambda}(\boldsymbol{\Sigma}+\boldsymbol{\Lambda}^2)(\boldsymbol{x}-\boldsymbol{\mu})\sqrt{\frac{\nu+p}{\nu+q(\boldsymbol{x})}} \,\middle|\, \boldsymbol{0}, \boldsymbol{\Delta}, \nu+1\right).$$

□

B.7 Proof of Proposition 2.23

Proof. a) We apply the law of iterated expectation to the representation in Lemma 2.22. From the proof of the univariate case in Prop. 2.12 we know that

$$\mathrm{E}(\boldsymbol{Z}) = \sqrt{\frac{\nu}{\pi}}\frac{\Gamma\left(\frac{\nu-1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\boldsymbol{1}, \text{ where } \boldsymbol{1} \text{ is the } p\text{-dimensional vector of ones.}$$

$\mathrm{E}(\boldsymbol{X} \mid \boldsymbol{Z} = \boldsymbol{z}, G = g) = \boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{z}$ does not depend on g. Therefore, $\mathrm{E}(\boldsymbol{X} \mid \boldsymbol{Z} = \boldsymbol{z}, G = g) = \mathrm{E}(\boldsymbol{X} \mid \boldsymbol{Z} = \boldsymbol{z})$ and

$$\mathrm{E}(\boldsymbol{X}) = \mathrm{E}(\mathrm{E}(\boldsymbol{X} \mid \boldsymbol{Z})) = \mathrm{E}(\boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{Z}) = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathrm{E}(\boldsymbol{Z}) = \boldsymbol{\mu} + \sqrt{\frac{\nu}{\pi}}\frac{\Gamma\left(\frac{\nu-1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\boldsymbol{\lambda}.$$

b) Using again the hierarchical representation it follows from the law of total variance that

$$\begin{aligned}
\mathrm{Var}(\boldsymbol{X}) &= \mathrm{E}(\mathrm{Var}(\boldsymbol{X} \mid \boldsymbol{Z}, G)) + \mathrm{Var}(\mathrm{E}(\boldsymbol{X} \mid \boldsymbol{Z}, G)) \\
&= \mathrm{E}\left(\frac{1}{G}\boldsymbol{\Sigma}\right) + \mathrm{Var}(\boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{Z}) = \mathrm{E}\left(\frac{1}{G}\right)\boldsymbol{\Sigma} + \mathrm{Var}(\boldsymbol{\Lambda}\boldsymbol{Z}) \\
&= \mathrm{E}\left(\frac{1}{G}\right)\boldsymbol{\Sigma} + \boldsymbol{\Lambda}(\mathrm{E}(\mathrm{Var}(\boldsymbol{Z} \mid G)) + \mathrm{Var}(\mathrm{E}(\boldsymbol{Z} \mid G)))\boldsymbol{\Lambda}^\top \\
&= \mathrm{E}\left(\frac{1}{G}\right)\boldsymbol{\Sigma} + \boldsymbol{\Lambda}\left(\left(1 - \frac{2}{\pi}\right)\boldsymbol{I}_p\,\mathrm{E}\left(\frac{1}{G}\right) + \mathrm{Var}\left(\sqrt{\frac{1}{G}}\sqrt{\frac{2}{\pi}}\boldsymbol{1}\right)\right)\boldsymbol{\Lambda}^\top.
\end{aligned}$$

Using that the variance can be expressed in terms of the second moment and the squared expectation it follows that

$$= \mathrm{E}\left(\frac{1}{G}\right)\left(\boldsymbol{\Sigma} + \left(1 - \frac{2}{\pi}\right)\boldsymbol{\Lambda}^2\right) + \boldsymbol{\Lambda}\frac{2}{\pi}\mathrm{E}\left(\frac{1}{G}\boldsymbol{1}_p\right)\boldsymbol{\Lambda}^\top - \boldsymbol{\Lambda}\mathrm{E}\left(\frac{1}{\sqrt{G}}\sqrt{\frac{2}{\pi}}\right)^2 \boldsymbol{1}_p\,\boldsymbol{\Lambda}^\top,$$

where $\mathbf{1}_p$ is the $p \times p$ dimensional matrix with only ones. For the second expectation we can use the proof of a), which yields

$$= \mathrm{E}\left(\frac{1}{G}\right)\left(\boldsymbol{\Sigma} + \left(1 - \frac{2}{\pi}\right)\boldsymbol{\Lambda}^2\right) + \frac{2}{\pi}\mathrm{E}\left(\frac{1}{G}\right)\boldsymbol{\lambda}\boldsymbol{\lambda}^\top - \left(\sqrt{\frac{\nu}{\pi}}\frac{\Gamma\left(\frac{\nu-1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\right)^2 \boldsymbol{\lambda}\boldsymbol{\lambda}^\top,$$

since $\boldsymbol{\Lambda}\mathbf{1}_p\boldsymbol{\Lambda}^\top = \boldsymbol{\lambda}\boldsymbol{\lambda}^\top$. Using that $\mathrm{E}\left(\frac{1}{G}\right) = \frac{\nu}{\nu-2}$, which was shown in the univariate case it follows that

$$\begin{aligned}
&\frac{\nu}{\nu-2}\left(\boldsymbol{\Sigma} + \left(1 - \frac{2}{\pi}\right)\boldsymbol{\Lambda}^2\right) + \frac{2}{\pi}\frac{\nu}{\nu-2}\boldsymbol{\lambda}\boldsymbol{\lambda}^\top - \left(\sqrt{\frac{\nu}{\pi}}\frac{\Gamma\left(\frac{\nu-1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\right)^2 \boldsymbol{\lambda}\boldsymbol{\lambda}^\top \\
&= \frac{\nu}{\nu-2}\left(\boldsymbol{\Sigma} + \boldsymbol{\Lambda}^2\left(1 - \frac{2}{\pi}\right)\right) + \frac{2}{\pi}\left(\frac{\nu}{\nu-2} - \left(\frac{\Gamma\left(\frac{\nu-1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\right)^2 \frac{\nu}{2}\right)\boldsymbol{\lambda}\boldsymbol{\lambda}^\top.
\end{aligned}$$

□

B.8 Proof of Theorem 3.1

Proof. First note that

$$\begin{aligned}
\ln(p(\boldsymbol{\theta} \mid \boldsymbol{Y})) &= \ln\left(\frac{p(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{Z})p(\boldsymbol{Z} \mid \boldsymbol{Y})}{p(\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{\theta})}\right) \\
&= \ln(p(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{Z})) - \ln(p(\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{\theta})) + \ln(\boldsymbol{Z} \mid \boldsymbol{Y})
\end{aligned}$$

For an arbitrary $\boldsymbol{\phi}$ integrating both sides of the equation with respect to $p(\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{\phi})$ yields

$$\begin{aligned}
\int_{\boldsymbol{Z}} \ln(p(\boldsymbol{\theta} \mid \boldsymbol{Y}))p(\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{\phi})d\boldsymbol{Z} = &\int_{\boldsymbol{Z}} \ln(p(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{Z}))p(\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{\phi})d\boldsymbol{Z} \\
&- \int_{\boldsymbol{Z}} \ln(p(\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{\theta}))p(\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{\phi})d\boldsymbol{Z} \\
&+ \int_{\boldsymbol{Z}} \ln(\boldsymbol{Z} \mid \boldsymbol{Y})p(\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{\phi})d\boldsymbol{Z}.
\end{aligned}$$

Rearrangement of integrals and addition of definitions yields

$$\begin{aligned}
\ln(p(\boldsymbol{\theta} \mid \boldsymbol{Y}))\underbrace{\int_{\boldsymbol{Z}} p(\boldsymbol{Z} \mid \boldsymbol{Y}, \phi)d\boldsymbol{Z}}_{=1} = &\underbrace{\int_{\boldsymbol{Z}} \ln(p(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{Z}))p(\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{\phi})d\boldsymbol{Z}}_{Q(\boldsymbol{\theta},\boldsymbol{\phi})} \\
&- \underbrace{\int_{\boldsymbol{Z}} \ln(p(\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{\theta}))p(\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{\phi})d\boldsymbol{Z}}_{=H(\boldsymbol{\theta},\boldsymbol{\phi}))} \\
&+ \underbrace{\int_{\boldsymbol{Z}} \ln(\boldsymbol{Z} \mid \boldsymbol{Y})p(\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{\phi})d\boldsymbol{Z}.}_{=K(\boldsymbol{\phi}) \text{ independent of } \boldsymbol{\theta}}
\end{aligned}$$

The final is thus expressed as

$$\ln(p(\boldsymbol{\theta} \mid \boldsymbol{Y})) = Q(\boldsymbol{\theta}, \boldsymbol{\phi}) - H(\boldsymbol{\theta}, \boldsymbol{\phi})) + K(\boldsymbol{\phi}).$$

Note that $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)})$ is computed by the E-step of the EM algorithm. For EM iterates $\boldsymbol{\theta}^{(k)}$, $k \geq 1$ we have

$$\begin{aligned}\ln(p(\boldsymbol{\theta}^{(k+1)} \mid \boldsymbol{Y})) - \ln(p(\boldsymbol{\theta}^{(k)} \mid \boldsymbol{Y})) = &\underbrace{(Q(\boldsymbol{\theta}^{(k+1)}, \boldsymbol{\theta}^{(k)}) - Q(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k)}))}_{\geq 0 \text{ by M-step}} - (H(\boldsymbol{\theta}^{(k+1)}, \boldsymbol{\theta}^{(k)}) \\ &- H(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k)})) + \underbrace{(K(\boldsymbol{\theta}^{(k)}) - K(\boldsymbol{\theta}^{(k)}))}_{=0}.\end{aligned}$$

Furthermore it holds that

$$\begin{aligned}H(\boldsymbol{\theta}^{(k+1)}, \boldsymbol{\theta}^{(k)}) - H(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k)}) = &\int_{\boldsymbol{Z}} \ln\left(\frac{p(\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{\theta}^{(k+1)})}{p(\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{\theta}^{(k)})}\right) p(\boldsymbol{Z} \mid \boldsymbol{Y}, \theta^{(k)}) d\boldsymbol{Z} \\ \overset{\text{Jensen inequality}}{\leq} &\ln\left(\int_{\boldsymbol{Z}} \frac{p(\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{\theta}^{(k+1)})}{p(\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{\theta}^{(k)})} p(\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{\theta}^{(k)}) d\boldsymbol{Z}\right) \\ = &\ln(1) = 0.\end{aligned}$$

Recall Jensen's inequality that $\mathrm{E}(g(\boldsymbol{X})) \leq g(\mathrm{E}(\boldsymbol{X}))$ if g concave and $\mathrm{E}(|\boldsymbol{X}|) < \infty, \mathrm{E}(g(|\boldsymbol{X}|)) < \infty$. Therefore, $\Rightarrow \ln(p(\boldsymbol{\theta}^{(k+1)} \mid \boldsymbol{Y})) - \ln(p(\boldsymbol{\theta}^{(k)} \mid \boldsymbol{Y})) \geq 0$. □

B.9 Proof of Lemma 3.2

First we calculate the maximum for the parameter μ.

This $\dfrac{dQ(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})}{d\mu^{(k)}}$ equals

$$
\begin{aligned}
&= -\frac{d}{d\mu^{(k)}}\left(\frac{\sum_{i=1}^{n}(x_i^2 - 2x_i\mu^{(k)} + (\mu^{(k)})^2)\mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)})}{2((\sigma^{(k)})^2 + (\lambda^{(k)})^2)}\right) \\
&- \frac{d}{d\mu^{(k)}}\left(\frac{(\sigma^{(k)})^2 + (\lambda^{(k)})^2}{2(\sigma^{(k)})^2}\sum_{i=1}^{n}\mathrm{E}\left(g_i\left(z_i - \frac{\lambda^{(k)}(x_i - \mu^{(k)})}{(\sigma^{(k)})^2 + (\lambda^{(k)})^2}\right)^2 \,\middle|\, x_i, \boldsymbol{\theta}^{(k)}\right)\right) \\
&= \frac{\sum_{i=1}^{n} x_i\mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)}) - \mu^{(k)}\sum_{i=1}^{n}\mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)})}{(\sigma^{(k)})^2 + (\lambda^{(k)})^2} \\
&- \frac{d}{d\mu^{(k)}}\left(\frac{(\sigma^{(k)})^2 + (\lambda^{(k)})^2}{2(\sigma^{(k)})^2}\left(\sum_{i=1}^{n}\mathrm{E}(g_iz_i^2 \mid x_i, \boldsymbol{\theta}^{(k)})\right.\right. \\
&\left.\left.-2\sum_{i=1}^{n}\mathrm{E}\left(g_iz_i\frac{\lambda^{(k)}(x_i - \mu^{(k)})}{(\sigma^{(k)})^2 + (\lambda^{(k)})^2} \,\middle|\, x_i, \boldsymbol{\theta}^{(k)}\right) + \sum_{i=1}^{n}\mathrm{E}\left(g_i\left(\frac{\lambda^{(k)}(x_i - \mu^{(k)})}{(\sigma^{(k)})^2 + (\lambda^{(k)})^2}\right)^2 \,\middle|\, x_i, \boldsymbol{\theta}^{(k)}\right)\right)\right) \\
&= \frac{\sum_{i=1}^{n} x_i\mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)}) - \mu^{(k)}\sum_{i=1}^{n}\mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)})}{(\sigma^{(k)})^2 + (\lambda^{(k)})^2} \\
&- \frac{(\sigma^{(k)})^2 + (\lambda^{(k)})^2}{2(\sigma^{(k)})^2}\left(\frac{-2\lambda^{(k)}\sum_{i=1}^{n}\mathrm{E}(g_iz_i \mid x_i, \boldsymbol{\theta}^{(k)})}{(\sigma^{(k)})^2 + (\lambda^{(k)})^2} + \right. \\
&\left.\frac{2(\lambda^{(k)})^2\mu^{(k)}\sum_{i=1}^{n}\mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)}) - 2(\lambda^{(k)})^2\sum_{i=1}^{n} x_iE(g_i \mid x_i, \boldsymbol{\theta}^{(k)})}{((\sigma^{(k)})^2 + (\lambda^{(k)})^2)^2}\right) \\
&= \frac{(\sigma^{(k)})^2\sum_{i=1}^{n} x_i\mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)}) - \mu^{(k)}(\sigma^{(k)})^2\sum_{i=1}^{n}\mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)})}{(\sigma^{(k)})^2((\sigma^{(k)})^2 + (\lambda^{(k)})^2)} \\
&+ \frac{-\lambda^{(k)}((\sigma^{(k)})^2 + (\lambda^{(k)})^2)\sum_{i=1}^{n}\mathrm{E}(g_iz_i \mid x_i, \boldsymbol{\theta}^{(k)}) - (\lambda^{(k)})^2\mu^{(k)}\sum_{i=1}^{n}\mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)})}{(\sigma^{(k)})^2((\sigma^{(k)})^2 + (\lambda^{(k)})^2)} \\
&+ \frac{(\lambda^{(k)})^2\sum_{i=1}^{n} x_i\mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)})}{(\sigma^{(k)})^2((\sigma^{(k)})^2 + (\lambda^{(k)})^2)} \\
&= \frac{((\sigma^{(k)})^2 + (\lambda^{(k)})^2)\sum_{i=1}^{n} x_i\mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)}) - \mu^{(k)}((\sigma^{(k)})^2 + (\lambda^{(k)})^2)\sum_{i=1}^{n}\mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)})}{(\sigma^{(k)})^2((\sigma^{(k)})^2 + (\lambda^{(k)})^2)} \\
&- \frac{\lambda^{(k)}((\sigma^{(k)})^2 + (\lambda^{(k)})^2)\sum_{i=1}^{n}\mathrm{E}(g_iz_i \mid x_i, \boldsymbol{\theta}^{(k)})}{(\sigma^{(k)})^2((\sigma^{(k)})^2 + (\lambda^{(k)})^2)} \\
&= \frac{\sum_{i=1}^{n} x_i\mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)}) - \mu^{(k)}\sum_{i=1}^{n}\mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)}) - \lambda^{(k)}\sum_{i=1}^{n}\mathrm{E}(g_iz_i \mid x_i, \boldsymbol{\theta}^{(k)})}{(\sigma^{(k)})^2}.
\end{aligned}
$$

We set the derivative to zero and solve the equation for μ to achieve the maximum. Hence it follows that

$$
\Rightarrow \mu^{(k+1)} = \frac{\sum_{i=1}^{n} x_i\mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)}) - \lambda^{(k)}\sum_{i=1}^{n}\mathrm{E}(g_iz_i \mid x_i, \boldsymbol{\theta}^{(k)})}{\sum_{i=1}^{n}\mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)})}.
$$

In the next step we calculate the derivative

$\frac{dQ(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})}{d\nu^{(k)}}$, which equals
$$\frac{d}{d\nu^{(k)}}\left(-n\left[\frac{\nu^{(k)}}{2}\ln\left(\frac{\nu^{(k)}}{2}\right)-\ln\left(\sigma\pi\Gamma\left(\frac{\nu^{(k)}}{2}\right)\right)\right]+\frac{\nu^{(k)}}{2}\sum_{i=1}^{n}\mathrm{E}(\ln(g_i) \mid x_i,\boldsymbol{\theta}^{(k)})\right.$$
$$\left.-\frac{\nu^{(k)}}{2}\sum_{i=1}^{n}\mathrm{E}(g_i \mid x_i,\boldsymbol{\theta}^{(k)})\right).$$

Utilizing that $\frac{d}{dx}\ln(\Gamma(x)) = \psi(x)$ yields

$$=\frac{n}{2}\left(\ln\left(\frac{\nu^{(k)}}{2}\right)+1-\psi\left(\frac{\nu^{(k)}}{2}\right)\right)+\frac{1}{2}\left(\sum_{i=1}^{n}\mathrm{E}(\ln(g_i) \mid x_i,\boldsymbol{\theta}^{(k)})-\sum_{i=1}^{n}\mathrm{E}(g_i \mid x_i,\boldsymbol{\theta}^{(k)})\right).$$

Setting the equation zero it follows that

$$\Rightarrow \ln\left(\frac{\nu^{(k)}}{2}\right)-\psi\left(\frac{\nu^{(k)}}{2}\right)+1+\frac{\sum_{i=1}^{n}\mathrm{E}(\ln(g_i) \mid x_i,\boldsymbol{\theta}^{(k)})-\sum_{i=1}^{n}\mathrm{E}(g_i \mid x_i,\boldsymbol{\theta}^{(k)})}{n}=0.$$

This equation cannot be solved explicitly for ν. In the EM algorithm we calculate ν using a root search algorithm. Calculating the derivative of $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})$ with respect to λ yields

$$\begin{aligned}
&\frac{dQ(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})}{d\lambda^{(k)}} = \\
&= \frac{d}{d\lambda^{(k)}}\left(-\frac{\sum_{i=1}^{n}(x_i-\mu^{(k)})^2\mathrm{E}(g_i \mid x_i,\boldsymbol{\theta}^{(k)})}{2((\sigma^{(k)})^2+(\lambda^{(k)})^2)}\right. \\
&\left.-\frac{(\sigma^{(k)})^2+(\lambda^{(k)})^2}{2(\sigma^{(k)})^2}\sum_{i=1}^{n}\mathrm{E}\left(g_i\left(z_i-\frac{\lambda^{(k)}(x_i-\mu^{(k)})}{(\sigma^{(k)})^2+(\lambda^{(k)})^2}\right)^2 \,\middle|\, x_i,\boldsymbol{\theta}^{(k)}\right)\right) \\
&= \frac{\lambda^{(k)}\sum_{i=1}^{n}(x_i-\mu^{(k)})^2\mathrm{E}(g_i \mid x_i,\boldsymbol{\theta}^{(k)})}{((\sigma^{(k)})^2+(\lambda^{(k)})^2)^2} \\
&- \frac{d}{d\lambda^{(k)}}\left(\frac{((\sigma^{(k)})^2+(\lambda^{(k)})^2)\sum_{i=1}^{n}\mathrm{E}(g_i z_i^2 \mid x_i,\boldsymbol{\theta}^{(k)})}{2(\sigma^{(k)})^2}\right) \\
&+ \frac{d}{d\lambda^{(k)}}\left(\frac{\lambda^{(k)}\sum_{i=1}^{n}(x_i-\mu^{(k)})\mathrm{E}(g_i z_i \mid x_i,\boldsymbol{\theta}^{(k)})}{(\sigma^{(k)})^2}\right) \\
&- \frac{d}{d\lambda^{(k)}}\left(\frac{(\lambda^{(k)})^2\sum_{i=1}^{n}(x_i-\mu^{(k)})^2\mathrm{E}(g_i \mid x_i,\boldsymbol{\theta}^{(k)})}{2(\sigma^{(k)})^2((\sigma^{(k)})^2+(\lambda^{(k)})^2)}\right)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\lambda^{(k)} \sum_{i=1}^{n}(x_i - \mu^{(k)})^2 \mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)})}{((\sigma^{(k)})^2 + (\lambda^{(k)})^2)^2} - \frac{\lambda^{(k)} \sum_{i=1}^{n} \mathrm{E}(g_i z_i^2 \mid x_i, \boldsymbol{\theta}^{(k)})}{(\sigma^{(k)})^2} \\
&+ \frac{\sum_{i=1}^{n}(x_i - \mu^{(k)}) \mathrm{E}(g_i z_i \mid x_i, \boldsymbol{\theta}^{(k)})}{(\sigma^{(k)})^2} \\
&- \frac{\lambda^{(k)} \sum_{i=1}^{n}(x_i - \mu^{(k)})^2 \mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)})}{((\sigma^{(k)})^2 + (\lambda^{(k)})^2)^2} \\
&= \frac{\sum_{i=1}^{n}(x_i - \mu^{(k)}) \mathrm{E}(g_i z_i \mid x_i, \boldsymbol{\theta}^{(k)})}{(\sigma^{(k)})^2} - \frac{\lambda^{(k)} \sum_{i=1}^{n} \mathrm{E}(g_i z_i^2 \mid x_i, \boldsymbol{\theta}^{(k)})}{(\sigma^{(k)})^2}.
\end{aligned}
$$

Setting the derivative zero and solving the equation for λ yields

$$
\Rightarrow \lambda^{(k+1)} = \frac{\sum_{i=1}^{n}(x_i - \mu^{(k+1)}) \mathrm{E}(g_i z_i \mid x_i, \boldsymbol{\theta}^{(k)})}{\sum_{i=1}^{n} \mathrm{E}(g_i z_i^2 \mid x_i, \boldsymbol{\theta}^{(k)})}.
$$

For the EM algorithm we first need to calculated $\mu^{(k+1)}$ and then insert it in the equation for $\lambda^{(k)}$. The derivative of $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})$ with respect to σ^2 equals

$$
\begin{aligned}
&\frac{dQ(\theta \mid \theta^{(k)})}{d\sigma^{(k)}} = 0 \\
&= -\frac{d}{d\sigma^{(k)}} \left(n \ln \sigma^{(k)} \right) - \frac{d}{d\sigma^{(k)}} \left(\frac{\sum_{i=1}^{n}(x_i - \mu^{(k)})^2 \mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)})}{2((\sigma^{(k)})^2 + (\lambda^{(k)})^2)} \right) \\
&- \frac{d}{d\sigma^{(k)}} \left(\frac{(\sigma^{(k)})^2 + (\lambda^{(k)})^2}{2(\sigma^{(k)})^2} \sum_{i=1}^{n} \mathrm{E} \left(g_i \left(z_i - \frac{\lambda^{(k)}(x_i - \mu^{(k)})}{(\sigma^{(k)})^2 + (\lambda^{(k)})^2} \right)^2 \,\middle|\, x, \theta^{(k)} \right) \right) \\
&= -\frac{n}{\sigma^{(k)}} + \frac{\sigma^{(k)} \sum_{i=1}^{n}(x_i - \mu^{(k)})^2 \mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)})}{((\sigma^{(k)})^2 + (\lambda^{(k)})^2)^2} \\
&- \frac{d}{d\sigma^{(k)}} \left(\frac{((\sigma^{(k)})^2 + (\lambda^{(k)})^2) \sum_{i=1}^{n} \mathrm{E}(g_i z_i^2 \mid x_i, \boldsymbol{\theta}^{(k)})}{2(\sigma^{(k)})^2} \right) \\
&+ \frac{d}{d\sigma^{(k)}} \left(\frac{\lambda^{(k)} \sum_{i=1}^{n}(x_i - \mu^{(k)}) \mathrm{E}(g_i z_i \mid x_i, \boldsymbol{\theta}^{(k)})}{(\sigma^{(k)})^2} \right) \\
&- \frac{d}{d\sigma^{(k)}} \left(\frac{(\lambda^{(k)})^2 \sum_{i=1}^{n}(x_i - \mu^{(k)})^2 \mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)})}{2(\sigma^{(k)})^2((\sigma^{(k)})^2 + (\lambda^{(k)})^2)} \right) \\
&= -\frac{n}{\sigma^{(k)}} + \frac{\sigma^{(k)} \sum_{i=1}^{n}(x_i - \mu^{(k)})^2 \mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)})}{((\sigma^{(k)})^2 + (\lambda^{(k)})^2)^2} + \frac{(\lambda^{(k)})^2 \sum_{i=1}^{n} \mathrm{E}(g_i z_i^2 \mid x_i, \boldsymbol{\theta}^{(k)})}{(\sigma^{(k)})^3} \\
&- \frac{2\lambda^{(k)} \sum_{i=1}^{n}(x_i - \mu^{(k)}) \mathrm{E}(g_i z_i \mid x_i, \boldsymbol{\theta}^{(k)})}{(\sigma^{(k)})^3}) \\
&- \frac{(\lambda^{(k)})^2((2\sigma^{(k)})^2 + (\lambda^{(k)})^2) \sum_{i=1}^{n}(x_i - \mu^{(k)})^2 \mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)})}{(\sigma^{(k)})^3((\sigma^{(k)})^2 + (\lambda^{(k)})^2)^2}.
\end{aligned}
$$

Simplifying yields

$$
\begin{aligned}
&-\frac{n}{\sigma^{(k)}}+\frac{\left[(\sigma^{(k)})^4-(\lambda^{(k)})^2((2\sigma^{(k)})^2+(\lambda^{(k)})^2)\right]\sum_{i=1}^{n}(x_i-\mu^{(k)})^2\mathrm{E}(g_i\mid x_i,\boldsymbol{\theta}^{(k)})}{(\sigma^{(k)})^3((\sigma^{(k)})^2+(\lambda^{(k)})^2)^2}\\
&+\frac{(\lambda^{(k)})^2\sum_{i=1}^{n}\mathrm{E}(g_iz_i^2\mid x_i,\boldsymbol{\theta}^{(k)})}{(\sigma^{(k)})^3}-\frac{2\lambda^{(k)}\sum_{i=1}^{n}(x_i-\mu^{(k)})\mathrm{E}(g_iz_i\mid x_i,\boldsymbol{\theta}^{(k)})}{(\sigma^{(k)})^3}\\
&=-\frac{n}{\sigma^{(k)}}+\frac{\sum_{i=1}^{n}(x_i-\mu^{(k)})^2\mathrm{E}(g_i\mid x_i,\boldsymbol{\theta}^{(k)})+(\lambda^{(k)})^2\sum_{i=1}^{n}\mathrm{E}(g_iz_i^2\mid x_i,\boldsymbol{\theta}^{(k)})}{(\sigma^{(k)})^3}\\
&-\frac{2\lambda^{(k)}\sum_{i=1}^{n}(x_i-\mu^{(k)})\mathrm{E}(g_iz_i\mid x_i,\boldsymbol{\theta}^{(k)})}{(\sigma^{(k)})^3}.
\end{aligned}
$$

Setting the derivative zero and solving it for σ^2 yields

$$
\begin{aligned}
\Rightarrow (\sigma^{(k+1)})^2 &= \frac{\sum_{i=1}^{n}(x_i-\mu^{(k+1)})^2\mathrm{E}(g_i\mid x_i,\boldsymbol{\theta}^{(k)})+(\lambda^{(k+1)})^2\sum_{i=1}^{n}\mathrm{E}(g_iz_i^2\mid x_i,\boldsymbol{\theta}^{(k)})}{n}\\
&-\frac{2\lambda^{(k+1)}\sum_{i=1}^{n}(x_i-\mu^{(k+1)})\mathrm{E}(g_iz_i\mid x_i,\boldsymbol{\theta}^{(k)})}{n}.
\end{aligned}
$$

B.10 Proof of Lemma 3.3

Proof. Since the distribution function of the t distribution has no closed form it needs to be expressed as integral over the density function. We need to calculate

$$
\frac{d}{d\nu}\left(F_T\left(\frac{\lambda(x-\mu)}{\sigma^2+\lambda^2}\sqrt{\frac{\nu+1}{\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}}}\;\middle|\;0,\frac{\sigma^2}{\sigma^2+\lambda^2},\nu+1\right)\right)
$$

and again swap integration and derivation by applying the Leibniz rule. It is important that there is no ν in the bounds of the integral. So as a first step the integral is reparametrized by changing the scale parameter of the cdf and the upper bound of the integral using Lemma 2.4.

$$
\frac{d}{d\nu}\left(F_T\left(\frac{\lambda(x-\mu)}{\sigma^2+\lambda^2}\;\middle|\;0,\frac{\sigma^2\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)}{(\sigma^2+\lambda^2)(\nu+1)},\nu+1\right)\right)
$$

As next step, the definition of the cdf as integral of the pdf is used to rewrite the formula and results in

$$
\frac{d}{d\nu}\left(\int_{-\infty}^{\frac{\lambda(x-\mu)}{\sigma^2+\lambda^2}}\frac{\Gamma\left(\frac{\nu+2}{2}\right)}{\Gamma\left(\frac{\nu+1}{2}\right)\sqrt{\pi(\nu+1)}}\sqrt{\frac{(\sigma^2+\lambda^2)(\nu+1)}{\sigma^2\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)}}\left(1+\frac{t^2}{\frac{\sigma^2\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)(\nu+1)}{(\sigma^2+\lambda^2)(\nu+1)}}\right)^{-\frac{\nu+2}{2}}dt\right).
$$

Rewriting of some parts of the equation yields a simplified representation to

$$= \frac{d}{d\nu}\left(\int_{-\infty}^{\frac{\lambda(x-\mu)}{\sigma^2+\lambda^2}} \frac{\Gamma\left(\frac{\nu+2}{2}\right)\sqrt{\sigma^2+\lambda^2}}{\Gamma\left(\frac{\nu+1}{2}\right)\sigma\sqrt{\pi\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)}}\left(1+\frac{t^2(\sigma^2+\lambda^2)}{\sigma^2\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)}\right)^{-\frac{\nu+2}{2}} dt\right).$$

Using Leibniz rule allows to change the order of the integral and the derivative yields

$$= \int_{-\infty}^{\frac{\lambda(x-\mu)}{\sigma^2+\lambda^2}} \frac{d}{d\nu}\left(\frac{\Gamma\left(\frac{\nu+2}{2}\right)\sqrt{\sigma^2+\lambda^2}}{\Gamma\left(\frac{\nu+1}{2}\right)\sigma\sqrt{\pi\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)}}\left(1+\frac{t^2(\sigma^2+\lambda^2)}{\sigma^2\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)}\right)^{-\frac{\nu+2}{2}}\right) dt.$$

Using the product rule it follows that

$$\begin{aligned} &= \frac{\sqrt{\sigma^2+\lambda^2}}{\sigma\sqrt{\pi}} \int_{-\infty}^{\frac{\lambda(x-\mu)}{\sigma^2+\lambda^2}} \underbrace{\frac{d}{d\nu}\left(\frac{\Gamma\left(\frac{\nu+2}{2}\right)}{\Gamma\left(\frac{\nu+1}{2}\right)\sqrt{\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}}}\right)}_{(\star)} \left(1+\frac{t^2(\sigma^2+\lambda^2)}{\sigma^2\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)}\right)^{-\frac{\nu+2}{2}} \\ &+ \frac{\Gamma\left(\frac{\nu+2}{2}\right)}{\Gamma\left(\frac{\nu+1}{2}\right)\sqrt{\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}}} \underbrace{\frac{d}{d\nu}\left(\left(1+\frac{t^2(\sigma^2+\lambda^2)}{\sigma^2\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)}\right)^{-\frac{\nu+2}{2}}\right)}_{(\star\star)} dt. \end{aligned}$$

In the next step, the derivatives which are needed for the equation above are calculated separately. First, the derivative of $(\star)$ is calculated as

$$\frac{d}{d\nu}\left(\frac{\Gamma\left(\frac{\nu+2}{2}\right)}{\Gamma\left(\frac{\nu+1}{2}\right)\sqrt{\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}}}\right).$$

Using the quotient rule yields

$$= \frac{\Gamma\left(\frac{\nu+1}{2}\right)\sqrt{\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}}\frac{d}{d\nu}\left(\Gamma\left(\frac{\nu+2}{2}\right)\right) - \Gamma\left(\frac{\nu+2}{2}\right)\frac{d}{d\nu}\left(\Gamma\left(\frac{\nu+1}{2}\right)\sqrt{\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}}\right)}{\Gamma\left(\frac{\nu+1}{2}\right)^2\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)}.$$

Using that the derivative of the $\Gamma\left(\frac{\nu+1}{2}\right) = \Gamma\left(\frac{\nu+1}{2}\right)\psi\left(\frac{\nu+1}{2}\right)\frac{1}{2}$ and $\Gamma\left(\frac{\nu+2}{2}\right) = \Gamma\left(\frac{\nu+2}{2}\right)\psi\left(\frac{\nu+2}{2}\right)\frac{1}{2}$ yields

$$\begin{aligned}
&= \frac{\Gamma\left(\frac{\nu+1}{2}\right)\sqrt{\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}}\Gamma\left(\frac{\nu+2}{2}\right)\psi\left(\frac{\nu+2}{2}\right)\frac{1}{2}}{\Gamma\left(\frac{\nu+1}{2}\right)^2\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)} \\
&- \frac{\frac{1}{2}\Gamma\left(\frac{\nu+2}{2}\right)\Gamma\left(\frac{\nu+1}{2}\right)\left(\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)^{-\frac{1}{2}}+\psi\left(\frac{\nu+1}{2}\right)\sqrt{\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}}\right)}{\Gamma\left(\frac{\nu+1}{2}\right)^2\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)} \\
&= \frac{\frac{1}{2}\Gamma\left(\frac{\nu+2}{2}\right)\Gamma\left(\frac{\nu+1}{2}\right)\left(\sqrt{\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}}\psi\left(\frac{\nu+2}{2}\right)-\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)^{-\frac{1}{2}}-\psi\left(\frac{\nu+1}{2}\right)\sqrt{\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}}\right)}{\Gamma\left(\frac{\nu+1}{2}\right)^2\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)} \\
&= \frac{\frac{1}{2}\Gamma\left(\frac{\nu+2}{2}\right)\sqrt{\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}}\left(\psi\left(\frac{\nu+2}{2}\right)-\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)^{-1}-\psi\left(\frac{\nu+1}{2}\right)\right)}{\Gamma\left(\frac{\nu+1}{2}\right)\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)} \\
&= \frac{\frac{1}{2}\Gamma\left(\frac{\nu+2}{2}\right)\left(\psi\left(\frac{\nu+2}{2}\right)-\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)^{-1}-\psi\left(\frac{\nu+1}{2}\right)\right)}{\Gamma\left(\frac{\nu+1}{2}\right)\sqrt{\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}}}.
\end{aligned}$$

Now taking the derivative of the second part of the equation $(\star\star)$

$$\frac{d}{d\nu}\left(\left(1+\frac{t^2(\sigma^2+\lambda^2)}{\sigma^2\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)}\right)^{-\frac{\nu+2}{2}}\right).$$

Using the chain rule and using that $\frac{d}{dv}u^v = u^v\ln(u)$ yields

$$\begin{aligned}
&= -\frac{\nu+2}{2}\left(1+\frac{t^2(\sigma^2+\lambda^2)}{\sigma^2\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)}\right)^{-\frac{\nu+2}{2}-1}\frac{d}{d\nu}\left(1+\frac{t^2(\sigma^2+\lambda^2)}{\sigma^2\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)}\right) \\
&- \frac{1}{2}\left(1+\frac{t^2(\sigma^2+\lambda^2)}{\sigma^2\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)}\right)^{-\frac{\nu+2}{2}}\ln\left(1+\frac{t^2(\sigma^2+\lambda^2)}{\sigma^2\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)}\right).
\end{aligned}$$

Factoring out and simplifying results in

$$
\begin{aligned}
&= \frac{1}{2}\left(1+\frac{t^2(\sigma^2+\lambda^2)}{\sigma^2\left(\nu+\frac{(x-\mu)^2}{(\sigma^2+\lambda^2)}\right)}\right)^{-\frac{\nu+2}{2}} \left(\frac{\frac{t^2(\sigma^2+\lambda^2)(\nu+2)}{\sigma^2\left(\nu+\frac{(x-\mu)^2}{\lambda^2+\sigma^2}\right)^2}}{1+\frac{t^2(\sigma^2+\lambda^2)}{\sigma^2\left(\nu+\frac{(x-\mu)^2}{(\sigma^2+\lambda^2)}\right)}} - \ln\left(1+\frac{t^2(\sigma^2+\lambda^2)}{\sigma^2\left(\nu+\frac{(x-\mu)^2}{(\sigma^2+\lambda^2)}\right)}\right)\right)\\
&= \frac{1}{2}\left(1+\frac{t^2(\sigma^2+\lambda^2)}{\sigma^2\left(\nu+\frac{(x-\mu)^2}{(\sigma^2+\lambda^2)}\right)}\right)^{-\frac{\nu+2}{2}}\\
&\cdot\left(\frac{t^2\left(\frac{\sigma^2+\lambda^2}{\sigma^2}\right)(\nu+2)}{\left(\nu+\frac{(x-\mu)^2}{\lambda^2+\sigma^2}\right)\left(\nu+\frac{(x-\mu)}{\sigma^2+\lambda^2}+\frac{t^2(\sigma^2+\lambda^2)}{\sigma^2}\right)} - \ln\left(1+\frac{t^2(\sigma^2+\lambda^2)}{\sigma^2\left(\nu+\frac{(x-\mu)^2}{(\sigma^2+\lambda^2)}\right)}\right)\right).
\end{aligned}
$$

Now in the next step the calculated derivatives are inserted in the equation

$$
\begin{aligned}
&\frac{\sqrt{\sigma^2+\lambda^2}}{\sigma\sqrt{\pi}}\int_{-\infty}^{\frac{\lambda(x-\mu)}{\sigma^2+\lambda^2}} \frac{d}{d\nu}\left(\frac{\Gamma\left(\frac{\nu+2}{2}\right)}{\Gamma\left(\frac{\nu+1}{2}\right)\sqrt{\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}}}\right)\left(1+\frac{t^2(\sigma^2+\lambda^2)}{\sigma^2\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)}\right)^{-\frac{\nu+2}{2}}\\
&+\frac{\Gamma\left(\frac{\nu+2}{2}\right)}{\Gamma\left(\frac{\nu+1}{2}\right)\sqrt{\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}}}\frac{d}{d\nu}\left(\left(1+\frac{t^2(\sigma^2+\lambda^2)}{\sigma^2\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)}\right)^{-\frac{\nu+2}{2}}\right)dt.
\end{aligned}
$$

Inserting the derivative of the calculated parts yields

$$
\begin{aligned}
&= \frac{\sqrt{\sigma^2+\lambda^2}}{\sigma\sqrt{\pi}}\int_{-\infty}^{\frac{\lambda(x-\mu)}{\sigma^2+\lambda^2}}\left(1+\frac{t^2(\sigma^2+\lambda^2)}{\sigma^2\left(\nu+\frac{(x-\mu)^2}{(\sigma^2+\lambda^2)}\right)}\right)^{-\frac{\nu+2}{2}}\\
&\cdot\frac{\frac{1}{2}\Gamma\left(\frac{\nu+2}{2}\right)\left(\psi\left(\frac{\nu+2}{2}\right)-\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)^{-1}-\psi\left(\frac{\nu+1}{2}\right)\right)}{\Gamma\left(\frac{\nu+1}{2}\right)\sqrt{\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}}}\\
&+\frac{\Gamma\left(\frac{\nu+2}{2}\right)}{\Gamma\left(\frac{\nu+1}{2}\right)\sqrt{\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}}}\frac{1}{2}\left(1+\frac{t^2(\sigma^2+\lambda^2)}{\sigma^2\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)}\right)^{-\frac{\nu+2}{2}}\\
&\cdot\left(\frac{t^2\left(\frac{\sigma^2+\lambda^2}{\sigma^2}\right)(\nu+2)}{\left(\nu+\frac{(x-\mu)^2}{\lambda^2+\sigma^2}\right)\left(\nu+\frac{(x-\mu)}{\sigma^2+\lambda^2}+\frac{t^2(\sigma^2+\lambda^2)}{\sigma^2}\right)} - \ln\left(1+\frac{t^2(\sigma^2+\lambda^2)}{\sigma^2\left(\nu+\frac{(x-\mu)^2}{(\sigma^2+\lambda^2)}\right)}\right)\right)dt.
\end{aligned}
$$

After simplifying the equation it follows that

$$= \frac{1}{2} \int_{-\infty}^{\frac{\lambda(x-\mu)}{\sigma^2+\lambda^2}} \underbrace{\frac{\sqrt{\sigma^2+\lambda^2}}{\sigma\sqrt{\pi}} \frac{\Gamma\left(\frac{\nu+2}{2}\right)}{\Gamma\left(\frac{\nu+1}{2}\right)\sqrt{\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}}} \left(1+\frac{t^2(\sigma^2+\lambda^2)}{\sigma^2\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)}\right)^{-\frac{\nu+2}{2}}}_{=f_t\left(t,0,\frac{\sigma^2\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)}{(\sigma^2+\lambda^2)(\nu+1)},\nu+1\right)}$$

$$\cdot \left(\psi\left(\frac{\nu+2}{2}\right) - \psi\left(\frac{\nu+1}{2}\right) - \left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)^{-1}\right.$$

$$\left. + \frac{t^2\left(\frac{\sigma^2+\lambda^2}{\sigma^2}\right)(\nu+2)}{\left(\nu+\frac{(x-\mu)^2}{\lambda^2+\sigma^2}\right)\left(\nu+\frac{(x-\mu)}{\sigma^2+\lambda^2}+\frac{t^2(\sigma^2+\lambda^2)}{\sigma^2}\right)} - \ln\left(1+\frac{t^2(\sigma^2+\lambda^2)}{\sigma^2\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)}\right)\right) dt.$$

Inserting the definition of the extended t distribution yields the final form of the derivative

$$= \frac{1}{2} \int_{-\infty}^{\frac{\lambda(x-\mu)^2}{\sigma^2+\lambda^2}} f_T\left(t,0,\frac{\sigma^2\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)}{(\sigma^2+\lambda^2)(\nu+1)},\nu+1\right) \left(\psi\left(\frac{\nu}{2}+1\right) - \psi\left(\frac{\nu+1}{2}\right)\right.$$

$$- \left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)^{-1} + \frac{t^2\left(\frac{\sigma^2+\lambda^2}{\sigma^2}\right)(\nu+2)}{\left(\nu+\frac{(x-\mu)^2}{\lambda^2+\sigma^2}\right)\left(\nu+\frac{(x-\mu)}{\sigma^2+\lambda^2}+\frac{t^2(\sigma^2+\lambda^2)}{\sigma^2}\right)}$$

$$\left. - \ln\left(1+\frac{t^2(\sigma^2+\lambda^2)}{\sigma^2\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)}\right)\right) dt.$$

Defining $c(t)$ yields

$$= \frac{1}{2} \int_{-\infty}^{\frac{\lambda(x-\mu)^2}{\sigma^2+\lambda^2}} f_T\left(t,0,\frac{\sigma^2\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)}{(\sigma^2+\lambda^2)(\nu+1)},\nu+1\right) c(t) dt,$$

where

$$\begin{aligned} c(t) = & \psi\left(\frac{\nu}{2}+1\right) - \psi\left(\frac{\nu+1}{2}\right) - \left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)^{-1} \\ & + \frac{t^2\left(\frac{\sigma^2+\lambda^2}{\sigma^2}\right)(\nu+2)}{\left(\nu+\frac{(x-\mu)^2}{\lambda^2+\sigma^2}\right)\left(\nu+\frac{(x-\mu)}{\sigma^2+\lambda^2}+\frac{t^2(\sigma^2+\lambda^2)}{\sigma^2}\right)} \\ & - \ln\left(1+\frac{t^2(\sigma^2+\lambda^2)}{\sigma^2\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)}\right). \end{aligned}$$

□

B.11 Proof of Lemma 3.4

Proof. a)

$$\begin{aligned}
&\mathrm{E}(G \mid X = x)\\
&= \int_{-\infty}^{\infty} g f_{G|X=x}(g) dg\\
&= \int_0^{\infty} \frac{g^{\frac{\nu-1}{2}+1} \exp\left(-\frac{g}{2}\left(\nu + \frac{(x-\mu)^2}{(\sigma^2+\lambda^2)}\right)\right) F_Z\left(\frac{\sqrt{g}\lambda(x-\mu)}{\sigma^2+\lambda^2} \,\middle|\, 0, \frac{\sigma^2}{\sigma^2+\lambda^2}\right)}{\Gamma\left(\frac{\nu+1}{2}\right)\left(\frac{(x-\mu)^2}{2(\sigma^2+\lambda^2)} + \frac{\nu}{2}\right)^{-\frac{\nu+1}{2}} F_T\left(\frac{\lambda(x-\mu)}{\sigma^2+\lambda^2}\sqrt{\frac{\nu+1}{\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}}} \,\middle|\, 0, \frac{\sigma^2}{\sigma^2+\lambda^2}, \nu+1\right)} dg\\
&= \frac{\left(\frac{(x-\mu)^2}{2(\sigma^2+\lambda^2)} + \frac{\nu}{2}\right)^{\frac{\nu+1}{2}}}{\Gamma\left(\frac{\nu+1}{2}\right) F_T\left(\frac{\lambda(x-\mu)}{\sigma^2+\lambda^2}\sqrt{\frac{\nu+1}{\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}}} \,\middle|\, 0, \frac{\sigma^2}{\sigma^2+\lambda^2}, \nu+1\right)}\\
&\cdot \int_0^{\infty} g^{\frac{\nu+1}{2}} \exp\left(-\frac{g}{2}\left(\nu + \frac{(x-\mu)^2}{(\sigma^2+\lambda^2)}\right)\right) F_Z\left(\frac{\sqrt{g}\lambda(x-\mu)}{\sigma^2+\lambda^2} \,\middle|\, 0, \frac{\sigma^2}{\sigma^2+\lambda^2}\right) dg,
\end{aligned}$$

where the first part of the integral is the density function of $\mathcal{G}\left(\frac{\nu+3}{2}, \frac{\nu}{2} + \frac{(x-\mu^2)}{2(\sigma^2+\lambda^2)}\right)$ without the normalizing constant. Rewriting the equation above in terms of the gamma density and using that $\frac{\Gamma(\frac{\nu+3}{2})}{\Gamma(\frac{\nu+1}{2})} = \frac{\nu+1}{2}$, yields

$$\begin{aligned}
&= \frac{(v+1)\left(\frac{(x-\mu)^2}{2(\sigma^2+\lambda^2)} + \frac{\nu}{2}\right)^{-1}}{F_T\left(\frac{\lambda(x-\mu)}{\sigma^2+\lambda^2}\sqrt{\frac{\nu+1}{\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}}} \,\middle|\, 0, \frac{\sigma^2}{\sigma^2+\lambda^2}, \nu+1\right)}\\
&\cdot \int_0^{\infty} f_G\left(x \,\middle|\, \frac{\nu+3}{2}, \frac{\nu}{2} + \frac{(x-\mu^2)}{2(\sigma^2+\lambda^2)}\right)\\
&\cdot F_Z\left(\frac{\sqrt{g}\lambda(x-\mu)}{\sigma^2+\lambda^2} \,\middle|\, 0, \frac{\sigma^2}{\sigma^2+\lambda^2}\right) dg.
\end{aligned}$$

Applying Prop. B.1 to the integral it follows that

$$= \frac{(\nu+1) f_T\left(\frac{\lambda(x-\mu)}{\sigma^2+\lambda^2}\sqrt{\frac{\nu+3}{\nu+\frac{(x-\mu)^2}{(\sigma^2+\lambda^2)}}} \,\middle|\, 0, \frac{\sigma^2}{\sigma^2+\lambda^2}, \nu+3\right)}{\left(\nu + \frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right) F_T\left(\frac{\lambda(x-\mu)}{\sigma^2+\lambda^2}\sqrt{\frac{\nu+1}{\nu+\frac{(x-\mu)^2}{(\sigma^2+\lambda^2)}}} \,\middle|\, 0, \frac{\sigma^2}{\sigma^2+\lambda^2}, \nu+1\right)}.$$

b) Before we begin with the calculations necessary to prove this statement a brief outline. Since $f_{G|X=x}(g)$ with support $[0, \infty)$ is a probability density function, $\int_0^\infty f_{G|X=x}(g)\, dg = 1$. Applying the Leibniz rule to swap integral and derivative we obtain,

$$\frac{d}{d\nu}\int_0^\infty f_{G|X=x}(g)\,dg = \frac{d}{d\nu}1 = 0 = \int_0^\infty \frac{d}{d\nu}f_{G|X=x}(g)\,dg$$

Considering only $\frac{d}{d\nu}f_{G|X=x}(g)$ yields

$$\frac{d}{d\nu}\frac{g^{\frac{\nu-1}{2}}\exp\left(-\frac{g}{2}\left(\nu+\frac{(x-\mu)^2}{(\sigma^2+\lambda^2)}\right)\right)F_Z\left(\frac{\sqrt{g}\lambda(x-\mu)}{\sigma^2+\lambda^2}\,\middle|\,0,\frac{\sigma^2}{\sigma^2+\lambda^2}\right)}{\Gamma\left(\frac{\nu+1}{2}\right)\left(\frac{(x-\mu)^2}{2(\sigma^2+\lambda^2)}+\frac{\nu}{2}\right)^{-\frac{\nu+1}{2}}F_T\left(\frac{\lambda(x-\mu)}{\sigma^2+\lambda^2}\sqrt{\frac{\nu+1}{\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}}}\,\middle|\,0,\frac{\sigma^2}{\sigma^2+\lambda^2},\nu+1\right)}.$$

To simplify the equation we subsitute the nominator with $\frac{1}{K(\nu)}$, which yields

$$\begin{aligned}
&= F_Z\left(\frac{\sqrt{g}\lambda(x-\mu)}{\sigma^2+\lambda^2}\,\middle|\,0,\frac{\sigma^2}{\sigma^2+\lambda^2}\right)\frac{d}{d\nu}\left(K(\nu)\,g^{\frac{\nu-1}{2}}\exp\left(-\frac{g}{2}\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)\right)\right)\\
&= F_Z\left(\frac{\sqrt{g}\lambda(x-\mu)}{\sigma^2+\lambda^2}\,\middle|\,0,\frac{\sigma^2}{\sigma^2+\lambda^2}\right)\left[\frac{d}{d\nu}\left(K(\nu)\right)g^{\frac{\nu-1}{2}}\exp\left(-\frac{g}{2}\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)\right)\right.\\
&\left.+K(\nu)\frac{d}{d\nu}\left(g^{\frac{\nu-1}{2}}\exp\left(-\frac{g}{2}\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)\right)\right)\right].
\end{aligned}$$

In the next step we calculate the derivative of the second summand from the equation above. For the derivative it follows that

$$\begin{aligned}
&\frac{d}{d\nu}\left(g^{\frac{\nu-1}{2}}\exp\left(-\frac{g}{2}\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)\right)\right)\\
&= g^{\frac{\nu-1}{2}}\frac{d}{d\nu}\left(\exp\left(-\frac{g}{2}\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)\right)\right)+\frac{d}{d\nu}g^{\frac{\nu-1}{2}}\exp\left(-\frac{g}{2}\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)\right)\\
&= g^{\frac{\nu-1}{2}}\exp\left(-\frac{g}{2}\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)\right)\left(-\frac{g}{2}\right)+g^{\frac{\nu-1}{2}}\ln(g)\frac{1}{2}\exp\left(-\frac{g}{2}\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)\right)\\
&= g^{\frac{\nu-1}{2}}\exp\left(-\frac{g}{2}\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)\right)\left(\frac{\ln(g)}{2}-\frac{g}{2}\right).
\end{aligned}$$

With the calculated derivative and by using the Leibniz rule it follows that

$$\begin{aligned}
0 &= \int_0^\infty \frac{d}{d\nu}f_{G|X=x}(g)dg\\
&= \int_0^\infty F_Z\left(\frac{\sqrt{g}\lambda(x-\mu)}{\sigma^2+\lambda^2}\,\middle|\,0,\frac{\sigma^2}{\sigma^2+\lambda^2}\right)\left[\frac{d}{d\nu}\left(K(\nu)\right)g^{\frac{\nu-1}{2}}\exp\left(-\frac{g}{2}\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)\right)\right.\\
&\left.+K(\nu)g^{\frac{\nu-1}{2}}\exp\left(-\frac{g}{2}\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)\right)\left(\frac{\ln(g)}{2}-\frac{g}{2}\right)\right]dg.
\end{aligned}$$

Using the linearity of the integrals we can split it up. Note that $K(\nu)$ is not dependent on g. Hence it follows that

$$
\begin{aligned}
&= \frac{d}{d\nu}\left(K(\nu)\right)\int_0^\infty F_Z\left(\frac{\sqrt{g}\lambda(x-\mu)}{\sigma^2+\lambda^2}\,\middle|\, 0,\frac{\sigma^2}{\sigma^2+\lambda^2}\right) g^{\frac{\nu-1}{2}}\exp\left(-\frac{g}{2}\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)\right)dg\\
&+\frac{1}{2}\int_0^\infty F_Z\left(\frac{\sqrt{g}\lambda(x-\mu)}{\sigma^2+\lambda^2}\,\middle|\, 0,\frac{\sigma^2}{\sigma^2+\lambda^2}\right)\ln(g)K(\nu) g^{\frac{\nu-1}{2}}\exp\left(-\frac{g}{2}\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)\right)dg\\
&-\frac{1}{2}\int_0^\infty F_Z\left(\frac{\sqrt{g}\lambda(x-\mu)}{\sigma^2+\lambda^2}\,\middle|\, 0,\frac{\sigma^2}{\sigma^2+\lambda^2}\right) gK(\nu) g^{\frac{\nu-1}{2}}\exp\left(-\frac{g}{2}\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)\right)dg.
\end{aligned}
$$

The second an the third integral of the equation above equal $\mathrm{E}(\ln(G)\mid X=x)$ and $\mathrm{E}(G\mid X=x)$, respectivley. This yields

$$
\begin{aligned}
&= \frac{d}{d\nu}\left(K(\nu)\right)\int_0^\infty F_Z\left(\frac{\sqrt{g}\lambda(x-\mu)}{\sigma^2+\lambda^2}\,\middle|\, 0,\frac{\sigma^2}{\sigma^2+\lambda^2}\right) g^{\frac{\nu-1}{2}}\exp\left(-\frac{g}{2}\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)\right)dg\\
&+\frac{1}{2}\mathrm{E}(\ln(G)\mid X=x)-\frac{1}{2}\mathrm{E}(G\mid X=x).
\end{aligned}
$$

Since we want to calculate $\mathrm{E}(G\mid X=x)$, we solve the equation for $\mathrm{E}(G\mid X=x)$. Hence it follows that

$$
\begin{aligned}
&\mathrm{E}(\ln(G)\mid X=x)\\
&=\mathrm{E}(G\mid X=x)\\
&-2\frac{d}{d\nu}\left(K(\nu)\right)\int_0^\infty F_Z\left(\frac{\sqrt{g}\lambda(x-\mu)}{\sigma^2+\lambda^2}\,\middle|\, 0,\frac{\sigma^2}{\sigma^2+\lambda^2}\right) g^{\frac{\nu-1}{2}}\exp\left(-\frac{g}{2}\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)\right)dg.
\end{aligned}
$$

Using Prop. B.1 the integral can be written as $F_T\left(\frac{\lambda(x-\mu)}{\sigma^2+\lambda^2}\sqrt{\frac{\nu+1}{\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}}}\,\middle|\, 0,\frac{\sigma^2}{\sigma^2+\lambda^2},\nu+1\right)$ multiplied by the constants of the gamma distribution. This yields that

$$
\begin{aligned}
&\mathrm{E}(\ln(G)\mid X=x)\\
&=\mathrm{E}(G\mid X=x)\\
&-2\frac{d}{d\nu}\left(K(\nu)\right)\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\left(\frac{(x-\mu)^2}{2(\sigma^2+\lambda^2)}+\frac{\nu}{2}\right)^{\frac{\nu+1}{2}}}F_T\left(\frac{\lambda(x-\mu)}{\sigma^2+\lambda^2}\sqrt{\frac{\nu+1}{\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}}}\,\middle|\, 0,\frac{\sigma^2}{\sigma^2+\lambda^2},\nu+1\right).
\end{aligned}
$$

As next step we the derivative of $K(\nu)=\frac{\left(\frac{(x-\mu)^2}{2(\sigma^2+\lambda^2)}+\frac{\nu}{2}\right)^{\frac{\nu+1}{2}}}{\Gamma\left(\frac{\nu+1}{2}\right)F_T\left(\frac{\lambda(x-\mu)}{\sigma^2+\lambda^2}\sqrt{\frac{\nu+1}{\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}}}\,\middle|\, 0,\frac{\sigma^2}{\sigma^2+\lambda^2},\nu+1\right)}$ is calculated. The derivative $\frac{d}{d\nu}K(\nu)$ equals

$$\frac{d}{d\nu}\left(\frac{\left(\frac{(x-\mu)^2}{2(\sigma^2+\lambda^2)}+\frac{\nu}{2}\right)^{\frac{\nu+1}{2}}}{\Gamma\left(\frac{\nu+1}{2}\right)F_T\left(\frac{\lambda(x-\mu)}{\sigma^2+\lambda^2}\sqrt{\frac{\nu+1}{\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}}}\,\middle|\,0,\frac{\sigma^2}{\sigma^2+\lambda^2},\nu+1\right)}\right)$$

$$=\frac{d}{d\nu}\left(\frac{\left(\frac{(x-\mu)^2}{2(\sigma^2+\lambda^2)}+\frac{\nu}{2}\right)^{\frac{\nu+1}{2}}}{\Gamma\left(\frac{\nu+1}{2}\right)}\left(F_T\left(\frac{\lambda(x-\mu)}{\sigma^2+\lambda^2}\sqrt{\frac{\nu+1}{\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}}}\,\middle|\,0,\frac{\sigma^2}{\sigma^2+\lambda^2},\nu+1\right)\right)^{-1}\right)$$

Using the product rule for the derivative we get

$$=\frac{d}{d\nu}\left(\frac{\left(\frac{(x-\mu)^2}{2(\sigma^2+\lambda^2)}+\frac{\nu}{2}\right)^{\frac{\nu+1}{2}}}{\Gamma\left(\frac{\nu+1}{2}\right)}\right)\left(F_T\left(\frac{\lambda(x-\mu)}{\sigma^2+\lambda^2}\sqrt{\frac{\nu+1}{\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}}}\,\middle|\,0,\frac{\sigma^2}{\sigma^2+\lambda^2},\nu+1\right)\right)^{-1}$$

$$+\frac{\left(\frac{(x-\mu)^2}{2(\sigma^2+\lambda^2)}+\frac{\nu}{2}\right)^{\frac{\nu+1}{2}}}{\Gamma\left(\frac{\nu+1}{2}\right)}\frac{d}{d\nu}\left(\left(F_T\left(\frac{\lambda(x-\mu)}{\sigma^2+\lambda^2}\sqrt{\frac{\nu+1}{\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}}}\,\middle|\,0,\frac{\sigma^2}{\sigma^2+\lambda^2},\nu+1\right)\right)^{-1}\right).$$

In the next step we calculate the derivative $\frac{d}{d\nu}\left(\frac{\left(\frac{(x-\mu)^2}{2(\sigma^2+\lambda^2)}+\frac{\nu}{2}\right)^{\frac{\nu+1}{2}}}{\Gamma\left(\frac{\nu+1}{2}\right)}\right)$, which equals

$$\frac{\Gamma\left(\frac{\nu+1}{2}\right)\frac{d}{d\nu}\left(\left(\frac{(x-\mu)^2}{2(\sigma^2+\lambda^2)}+\frac{\nu}{2}\right)^{\frac{\nu+1}{2}}\right)-\left(\frac{(x-\mu)^2}{2(\sigma^2+\lambda^2)}+\frac{\nu}{2}\right)^{\frac{\nu+1}{2}}\frac{d}{d\nu}\left(\Gamma\left(\frac{\nu+1}{2}\right)\right)}{\left(\Gamma\left(\frac{\nu+1}{2}\right)\right)^2}.$$

Using that $\frac{d}{d\nu}\left(\Gamma\left(\frac{\nu+1}{2}\right)\right) = \Gamma\left(\frac{\nu+1}{2}\right)\psi\left(\frac{\nu+1}{2}\right)\frac{1}{2}$, where $\psi(x)$ denotes the digamma function $\psi(x) = \frac{d}{dx}\ln\Gamma(x) = \frac{\frac{d}{dx}\Gamma(x)}{\Gamma(x)}$ it follows that

$$
\begin{aligned}
&\frac{\Gamma(\frac{\nu+1}{2})\left(\frac{\nu+1}{2}\left(\frac{(x-\mu)^2}{2(\sigma^2+\lambda^2)}+\frac{\nu}{2}\right)^{\frac{\nu-1}{2}}\frac{1}{2}+\left(\frac{(x-\mu)^2}{2(\sigma^2+\lambda^2)}+\frac{\nu}{2}\right)^{\frac{\nu+1}{2}}\frac{1}{2}\ln\left(\frac{(x-\mu)^2}{2(\sigma^2+\lambda^2)}+\frac{\nu}{2}\right)\right)}{\left(\Gamma\left(\frac{\nu+1}{2}\right)\right)^2} \\
&-\frac{\left(\frac{(x-\mu)^2}{2(\sigma^2+\lambda^2)}+\frac{\nu}{2}\right)^{\frac{\nu+1}{2}}\Gamma\left(\frac{\nu+1}{2}\right)\psi\left(\frac{\nu+1}{2}\right)\frac{1}{2}}{\left(\Gamma\left(\frac{\nu+1}{2}\right)\right)^2} \\
&=\frac{\frac{1}{2}\left(\frac{(x-\mu)^2}{2(\sigma^2+\lambda^2)}+\frac{\nu}{2}\right)^{\frac{\nu+1}{2}}\left[\frac{\nu+1}{2}\left(\frac{(x-\mu)^2}{2(\sigma^2+\lambda^2)}+\frac{\nu}{2}\right)^{-1}+\ln\left(\frac{(x-\mu)^2}{2(\sigma^2+\lambda^2)}+\frac{\nu}{2}\right)-\psi\left(\frac{\nu+1}{2}\right)\right]}{\Gamma\left(\frac{\nu+1}{2}\right)} \\
&=\frac{\frac{1}{2}\left(\frac{(x-\mu)^2}{2(\sigma^2+\lambda^2)}+\frac{\nu}{2}\right)^{\frac{\nu+1}{2}}\left[\frac{\nu+1}{\frac{(x-\mu)^2}{\sigma^2+\lambda^2}+\nu}+\ln\left(\frac{(x-\mu)^2}{2(\sigma^2+\lambda^2)}+\frac{\nu}{2}\right)-\psi\left(\frac{\nu+1}{2}\right)\right]}{\Gamma\left(\frac{\nu+1}{2}\right)}.
\end{aligned}
$$

Inserting the derivative from above and using Lemma 3.10 for the derivative of the t distribution function, $\frac{d}{d\nu}K(\nu)$ can be represented as

$$
\begin{aligned}
&\frac{\frac{1}{2}\left(\frac{(x-\mu)^2}{2(\sigma^2+\lambda^2)}+\frac{\nu}{2}\right)^{\frac{\nu+1}{2}}\left[\frac{\nu+1}{\frac{(x-\mu)^2}{\sigma^2+\lambda^2}+\nu}+\ln\left(\frac{\nu}{2}+\frac{(x-\mu)^2}{2(\sigma^2+\lambda^2)}\right)-\psi\left(\frac{\nu+1}{2}\right)\right]}{\Gamma\left(\frac{\nu+1}{2}\right)F_T\left(\frac{\lambda(x-\mu)}{\sigma^2+\lambda^2}\sqrt{\frac{\nu+1}{\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}}}\,\middle|\,0,\frac{\sigma^2}{\sigma^2+\lambda^2},\nu+1\right)} \\
&-\frac{\frac{1}{2}\left(\frac{(x-\mu)^2}{2(\sigma^2+\lambda^2)}+\frac{\nu}{2}\right)^{\frac{\nu+1}{2}}\int_{-\infty}^{\frac{\lambda(x-\mu)}{\sigma^2+\lambda^2}}f_T\left(t,0,\frac{\sigma^2\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)}{(\sigma^2+\lambda^2)(\nu+1)},\nu+1\right)c(t)\,dt}{\Gamma\left(\frac{\nu+1}{2}\right)\left(F_T\left(\frac{\lambda(x-\mu)}{\sigma^2+\lambda^2}\sqrt{\frac{\nu+1}{\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}}}\,\middle|\,0,\frac{\sigma^2}{\sigma^2+\lambda^2},\nu+1\right)\right)^2}.
\end{aligned}
$$

All components for $\mathrm{E}(\ln(G) \mid X = x)$ are calculated. Inserting yields that $\mathrm{E}(\ln(G) \mid X = x)$ equals

$$
\begin{aligned}
&\mathrm{E}(G \mid X=x) - \frac{2\frac{d}{d\nu}\left(K(\nu)\right)\Gamma\left(\frac{\nu+1}{2}\right)}{\left(\frac{(x-\mu)^2}{2(\sigma^2+\lambda^2)}+\frac{\nu}{2}\right)^{\frac{\nu+1}{2}}} F_T\left(\frac{\lambda(x-\mu)}{\sigma^2+\lambda^2}\sqrt{\frac{\nu+1}{\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}}}\,\middle|\, \frac{\sigma^2}{\sigma^2+\lambda^2}, \nu+1\right)\\
&= \mathrm{E}(G \mid X=x) - \frac{\nu+1}{\frac{(x-\mu)^2}{\sigma^2+\lambda^2}+\nu} - \ln\left(\frac{\nu}{2}+\frac{(x-\mu)^2}{2(\sigma^2+\lambda^2)}\right) + \psi\left(\frac{\nu+1}{2}\right)\\
&+ \frac{\int_{-\infty}^{\frac{\lambda(x-\mu)}{\sigma^2+\lambda^2}} f_T\left(t, 0, \frac{\sigma^2\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)}{(\sigma^2+\lambda^2)(\nu+1)}, \nu+1\right) c(t)\, dt}{F_T\left(\frac{\lambda(x-\mu)}{\sigma^2+\lambda^2}\sqrt{\frac{\nu+1}{\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}}}\,\middle|\, 0, \frac{\sigma^2}{\sigma^2+\lambda^2}, \nu+1\right)}.
\end{aligned}
$$

Replacing $\mathrm{E}(G \mid X=x)$ with its definition and simplifying yields

$$
\begin{aligned}
&\frac{\nu+1}{\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}}\left[\frac{f_T\left(\frac{\lambda(x-\mu)}{\sigma^2+\lambda^2}\sqrt{\frac{\nu+3}{\nu+\frac{(x-\mu)^2}{(\sigma^2+\lambda^2)}}}\,\middle|\, 0, \frac{\sigma^2}{\sigma^2+\lambda^2}, \nu+3\right)}{F_T\left(\frac{\lambda(x-\mu)}{\sigma^2+\lambda^2}\sqrt{\frac{\nu+1}{\nu+\frac{(x-\mu)^2}{(\sigma^2+\lambda^2)}}}\,\middle|\, 0, \frac{\sigma^2}{\sigma^2+\lambda^2}, \nu+1\right)} - 1\right] - \ln\left(\frac{\nu}{2}+\frac{(x-\mu)^2}{2(\sigma^2+\lambda^2)}\right)\\
&+ \psi\left(\frac{\nu+1}{2}\right) + \frac{\int_{-\infty}^{\frac{\lambda(x-\mu)}{\sigma^2+\lambda^2}} f_T\left(t, 0, \frac{\sigma^2\left(\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}\right)}{(\sigma^2+\lambda^2)(\nu+1)}, \nu+1\right) c(t)\, dt}{F_T\left(\frac{\lambda(x-\mu)}{\sigma^2+\lambda^2}\sqrt{\frac{\nu+1}{\nu+\frac{(x-\mu)^2}{\sigma^2+\lambda^2}}}\,\middle|\, 0, \frac{\sigma^2}{\sigma^2+\lambda^2}, \nu+1\right)}.
\end{aligned}
$$

c) Utilizing the independence of the random variates and inserting the definition of the first moment of the truncated t distribution (see Lemma 2.6 a)).

d) Utilizing the independence of the random variates and inserting the definition of the second moment of the truncated t distribution (see Lemma 2.6 b)).

□

B.12 Proof of Lemma 3.6

We want to maximize $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})$ subject to $\sum_{i=1}^{n} \omega_i = 1$, $\omega_i \geq 0$, $i = 1, \ldots, n$ with respect to $\boldsymbol{\theta}$. Hence we use the Lagrangian to optimize with constraints.

Proof. First we calculate the maximum for the parameter μ.

This $\dfrac{dQ(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})}{d\mu^{(k)}}$ equals

$$\begin{aligned}
&= -\frac{d}{d\mu^{(k)}} \left(\frac{\sum_{i=1}^{n}(x_i^2 - 2x_i\mu^{(k)} + (\mu^{(k)})^2)\mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)})}{2((\sigma^{(k)})^2 + (\lambda^{(k)})^2)} \right) \\
&- \frac{d}{d\mu^{(k)}} \left(\frac{(\sigma^{(k)})^2 + (\lambda^{(k)})^2}{2(\sigma^{(k)})^2} \sum_{i=1}^{n} \mathrm{E}\left(g_i \left(z_i - \frac{\lambda^{(k)}(x_i - \mu^{(k)})}{(\sigma^{(k)})^2 + (\lambda^{(k)})^2} \right)^2 \,\middle|\, x_i, \boldsymbol{\theta}^{(k)} \right) \right) \\
&= \frac{\sum_{i=1}^{n} x_i \mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)}) - \mu^{(k)} \sum_{i=1}^{n} \mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)})}{(\sigma^{(k)})^2 + (\lambda^{(k)})^2} \\
&- \frac{d}{d\mu^{(k)}} \Bigg(\frac{(\sigma^{(k)})^2 + (\lambda^{(k)})^2}{2(\sigma^{(k)})^2} \Bigg(\sum_{i=1}^{n} \mathrm{E}(g_i z_i^2 \mid x_i, \boldsymbol{\theta}^{(k)}) \\
&-2 \sum_{i=1}^{n} \mathrm{E}\left(g_i z_i \frac{\lambda^{(k)}(x_i - \mu^{(k)})}{(\sigma^{(k)})^2 + (\lambda^{(k)})^2} \,\middle|\, x_i, \boldsymbol{\theta}^{(k)} \right) + \sum_{i=1}^{n} \mathrm{E}\left(g_i \left(\frac{\lambda^{(k)}(x_i - \mu^{(k)})}{(\sigma^{(k)})^2 + (\lambda^{(k)})^2} \right)^2 \,\middle|\, x_i, \boldsymbol{\theta}^{(k)} \right) \Bigg) \Bigg) \\
&= \frac{\sum_{i=1}^{n} x_i \mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)}) - \mu^{(k)} \sum_{i=1}^{n} \mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)})}{(\sigma^{(k)})^2 + (\lambda^{(k)})^2} \\
&- \frac{(\sigma^{(k)})^2 + (\lambda^{(k)})^2}{2(\sigma^{(k)})^2} \Bigg(\frac{-2\lambda^{(k)} \sum_{i=1}^{n} \mathrm{E}(g_i z_i \mid x_i, \boldsymbol{\theta}^{(k)})}{(\sigma^{(k)})^2 + (\lambda^{(k)})^2} + \\
&\frac{2(\lambda^{(k)})^2 \mu^{(k)} \sum_{i=1}^{n} \mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)}) - 2(\lambda^{(k)})^2 \sum_{i=1}^{n} x_i E(g_i \mid x_i, \boldsymbol{\theta}^{(k)})}{((\sigma^{(k)})^2 + (\lambda^{(k)})^2)^2} \Bigg) \\
&= \frac{(\sigma^{(k)})^2 \sum_{i=1}^{n} x_i \mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)}) - \mu^{(k)} (\sigma^{(k)})^2 \sum_{i=1}^{n} \mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)})}{(\sigma^{(k)})^2((\sigma^{(k)})^2 + (\lambda^{(k)})^2)} \\
&+ \frac{-\lambda^{(k)}((\sigma^{(k)})^2 + (\lambda^{(k)})^2) \sum_{i=1}^{n} \mathrm{E}(g_i z_i \mid x_i, \boldsymbol{\theta}^{(k)}) - (\lambda^{(k)})^2 \mu^{(k)} \sum_{i=1}^{n} \mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)})}{(\sigma^{(k)})^2((\sigma^{(k)})^2 + (\lambda^{(k)})^2)} \\
&+ \frac{(\lambda^{(k)})^2 \sum_{i=1}^{n} x_i \mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)})}{(\sigma^{(k)})^2((\sigma^{(k)})^2 + (\lambda^{(k)})^2)} \\
&= \frac{((\sigma^{(k)})^2 + (\lambda^{(k)})^2) \sum_{i=1}^{n} x_i \mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)}) - \mu^{(k)}((\sigma^{(k)})^2 + (\lambda^{(k)})^2) \sum_{i=1}^{n} \mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)})}{(\sigma^{(k)})^2((\sigma^{(k)})^2 + (\lambda^{(k)})^2)} \\
&- \frac{\lambda^{(k)}((\sigma^{(k)})^2 + (\lambda^{(k)})^2) \sum_{i=1}^{n} \mathrm{E}(g_i z_i \mid x_i, \boldsymbol{\theta}^{(k)})}{(\sigma^{(k)})^2((\sigma^{(k)})^2 + (\lambda^{(k)})^2)} \\
&= \frac{\sum_{i=1}^{n} x_i \mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)}) - \mu^{(k)} \sum_{i=1}^{n} \mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)}) - \lambda^{(k)} \sum_{i=1}^{n} \mathrm{E}(g_i z_i \mid x_i, \boldsymbol{\theta}^{(k)})}{(\sigma^{(k)})^2}.
\end{aligned}$$

We set the derivative to zero and solve the equation for μ to achieve the maximum. Hence it follows that

$$\Rightarrow \mu^{(k+1)} = \frac{\sum_{i=1}^{n} x_i \mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)}) - \lambda^{(k)} \sum_{i=1}^{n} \mathrm{E}(g_i z_i \mid x_i, \boldsymbol{\theta}^{(k)})}{\sum_{i=1}^{n} \mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)})}.$$

In the next step we calculate the derivative

$\dfrac{dQ(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})}{d\nu^{(k)}}$, which equals

$$\frac{d}{d\nu^{(k)}}\left(-n\left[\frac{\nu^{(k)}}{2}\ln\left(\frac{\nu^{(k)}}{2}\right)-\ln\left(\sigma\pi\Gamma\left(\frac{\nu^{(k)}}{2}\right)\right)\right]+\frac{\nu^{(k)}}{2}\sum_{i=1}^{n}\mathrm{E}(\ln(g_i) \mid x_i,\boldsymbol{\theta}^{(k)})\right.$$
$$\left.-\frac{\nu^{(k)}}{2}\sum_{i=1}^{n}\mathrm{E}(g_i \mid x_i,\boldsymbol{\theta}^{(k)})\right).$$

Utilizing that $\frac{d}{dx}\ln(\Gamma(x))=\psi(x)$ yields

$$=\frac{n}{2}\left(\ln\left(\frac{\nu^{(k)}}{2}\right)+1-\psi\left(\frac{\nu^{(k)}}{2}\right)\right)+\frac{1}{2}\left(\sum_{i=1}^{n}\mathrm{E}(\ln(g_i) \mid x_i,\boldsymbol{\theta}^{(k)})-\sum_{i=1}^{n}\mathrm{E}(g_i \mid x_i,\boldsymbol{\theta}^{(k)})\right).$$

Setting the equation zero it follows that

$$\Rightarrow \ln\left(\frac{\nu^{(k)}}{2}\right)-\psi\left(\frac{\nu^{(k)}}{2}\right)+1+\frac{\sum_{i=1}^{n}\mathrm{E}(\ln(g_i) \mid x_i,\boldsymbol{\theta}^{(k)})-\sum_{i=1}^{n}\mathrm{E}(g_i \mid x_i,\boldsymbol{\theta}^{(k)})}{n}=0.$$

This equation cannot be solved explicitly for ν. In the EM algorithm we calculate ν using a root search algorithm. Calculating the derivative of $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})$ with respect to λ yields

$$\begin{aligned}
&\frac{dQ(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})}{d\lambda^{(k)}}=\\
&=\frac{d}{d\lambda^{(k)}}\left(-\frac{\sum_{i=1}^{n}(x_i-\mu^{(k)})^2\mathrm{E}(g_i \mid x_i,\boldsymbol{\theta}^{(k)})}{2((\sigma^{(k)})^2+(\lambda^{(k)})^2)}\right.\\
&\left.-\frac{(\sigma^{(k)})^2+(\lambda^{(k)})^2}{2(\sigma^{(k)})^2}\sum_{i=1}^{n}\mathrm{E}\left(g_i\left(z_i-\frac{\lambda^{(k)}(x_i-\mu^{(k)})}{(\sigma^{(k)})^2+(\lambda^{(k)})^2}\right)^2 \,\middle|\, x_i,\boldsymbol{\theta}^{(k)}\right)\right)\\
&=\frac{\lambda^{(k)}\sum_{i=1}^{n}(x_i-\mu^{(k)})^2\mathrm{E}(g_i \mid x_i,\boldsymbol{\theta}^{(k)})}{((\sigma^{(k)})^2+(\lambda^{(k)})^2)^2}\\
&-\frac{d}{d\lambda^{(k)}}\left(\frac{((\sigma^{(k)})^2+(\lambda^{(k)})^2)\sum_{i=1}^{n}\mathrm{E}(g_iz_i^2 \mid x_i,\boldsymbol{\theta}^{(k)})}{2(\sigma^{(k)})^2}\right)\\
&+\frac{d}{d\lambda^{(k)}}\left(\frac{\lambda^{(k)}\sum_{i=1}^{n}(x_i-\mu^{(k)})\mathrm{E}(g_iz_i \mid x_i,\boldsymbol{\theta}^{(k)})}{(\sigma^{(k)})^2}\right)\\
&-\frac{d}{d\lambda^{(k)}}\left(\frac{(\lambda^{(k)})^2\sum_{i=1}^{n}(x_i-\mu^{(k)})^2\mathrm{E}(g_i \mid x_i,\boldsymbol{\theta}^{(k)})}{2(\sigma^{(k)})^2((\sigma^{(k)})^2+(\lambda^{(k)})^2)}\right)\\
&=\frac{\lambda^{(k)}\sum_{i=1}^{n}(x_i-\mu^{(k)})^2\mathrm{E}(g_i \mid x_i,\boldsymbol{\theta}^{(k)})}{((\sigma^{(k)})^2+(\lambda^{(k)})^2)^2}\\
&-\frac{\lambda^{(k)}\sum_{i=1}^{n}\mathrm{E}(g_iz_i^2 \mid x_i,\boldsymbol{\theta}^{(k)})}{(\sigma^{(k)})^2}+\frac{\sum_{i=1}^{n}(x_i-\mu^{(k)})\mathrm{E}(g_iz_i \mid x_i,\boldsymbol{\theta}^{(k)})}{(\sigma^{(k)})^2}\\
&-\frac{\lambda^{(k)}\sum_{i=1}^{n}(x_i-\mu^{(k)})^2\mathrm{E}(g_i \mid x_i,\boldsymbol{\theta}^{(k)})}{((\sigma^{(k)})^2+(\lambda^{(k)})^2)^2}
\end{aligned}$$

$$= \frac{\sum_{i=1}^{n}(x_i - \mu^{(k)})\mathrm{E}(g_i z_i \mid x_i, \boldsymbol{\theta}^{(k)})}{(\sigma^{(k)})^2} - \frac{\lambda^{(k)} \sum_{i=1}^{n} \mathrm{E}(g_i z_i^2 \mid x_i, \boldsymbol{\theta}^{(k)})}{(\sigma^{(k)})^2}.$$

Setting the derivative zero and solving the equation for λ yields

$$\Rightarrow \lambda^{(k+1)} = \frac{\sum_{i=1}^{n}(x_i - \mu^{(k+1)})\mathrm{E}(g_i z_i \mid x_i, \boldsymbol{\theta}^{(k)})}{\sum_{i=1}^{n} \mathrm{E}(g_i z_i^2 \mid x_i, \boldsymbol{\theta}^{(k)})}.$$

For the EM algorithm we first need to calculated $\mu^{(k+1)}$ and then insert it in the equation for $\lambda^{(k)}$. The derivative of $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})$ with respect to σ^2 equals

$$\begin{aligned}
&\frac{dQ(\theta \mid \theta^{(k)})}{d\sigma^{(k)}} = 0 \\
&= -\frac{d}{d\sigma^{(k)}}\left(n \ln \sigma^{(k)}\right) - \frac{d}{d\sigma^{(k)}}\left(\frac{\sum_{i=1}^{n}(x_i - \mu^{(k)})^2 \mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)})}{2((\sigma^{(k)})^2 + (\lambda^{(k)})^2)}\right) \\
&- \frac{d}{d\sigma^{(k)}}\left(\frac{(\sigma^{(k)})^2 + (\lambda^{(k)})^2}{2(\sigma^{(k)})^2} \sum_{i=1}^{n} \mathrm{E}\left(g_i \left(z_i - \frac{\lambda^{(k)}(x_i - \mu^{(k)})}{(\sigma^{(k)})^2 + (\lambda^{(k)})^2}\right)^2 \middle| \, x, \theta^{(k)}\right)\right) \\
&= -\frac{n}{\sigma^{(k)}} + \frac{\sigma^{(k)} \sum_{i=1}^{n}(x_i - \mu^{(k)})^2 \mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)})}{((\sigma^{(k)})^2 + (\lambda^{(k)})^2)^2} \\
&- \frac{d}{d\sigma^{(k)}}\left(\frac{((\sigma^{(k)})^2 + (\lambda^{(k)})^2) \sum_{i=1}^{n} \mathrm{E}(g_i z_i^2 \mid x_i, \boldsymbol{\theta}^{(k)})}{2(\sigma^{(k)})^2}\right) \\
&+ \frac{d}{d\sigma^{(k)}}\left(\frac{\lambda^{(k)} \sum_{i=1}^{n}(x_i - \mu^{(k)})\mathrm{E}(g_i z_i \mid x_i, \boldsymbol{\theta}^{(k)})}{(\sigma^{(k)})^2}\right) \\
&- \frac{d}{d\sigma^{(k)}}\left(\frac{(\lambda^{(k)})^2 \sum_{i=1}^{n}(x_i - \mu^{(k)})^2 \mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)})}{2(\sigma^{(k)})^2((\sigma^{(k)})^2 + (\lambda^{(k)})^2)}\right) \\
&= -\frac{n}{\sigma^{(k)}} + \frac{\sigma^{(k)} \sum_{i=1}^{n}(x_i - \mu^{(k)})^2 \mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)})}{((\sigma^{(k)})^2 + (\lambda^{(k)})^2)^2} \\
&+ \frac{(\lambda^{(k)})^2 \sum_{i=1}^{n} \mathrm{E}(g_i z_i^2 \mid x_i, \boldsymbol{\theta}^{(k)})}{(\sigma^{(k)})^3} \\
&- \frac{2\lambda^{(k)} \sum_{i=1}^{n}(x_i - \mu^{(k)})\mathrm{E}(g_i z_i \mid x_i, \boldsymbol{\theta}^{(k)})}{(\sigma^{(k)})^3}) \\
&- \frac{(\lambda^{(k)})^2((2\sigma^{(k)})^2 + (\lambda^{(k)})^2) \sum_{i=1}^{n}(x_i - \mu^{(k)})^2 \mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)})}{(\sigma^{(k)})^3((\sigma^{(k)})^2 + (\lambda^{(k)})^2)^2} \\
&= -\frac{n}{\sigma^{(k)}} + \frac{\left[(\sigma^{(k)})^4 - (\lambda^{(k)})^2((2\sigma^{(k)})^2 + (\lambda^{(k)})^2)\right] \sum_{i=1}^{n}(x_i - \mu^{(k)})^2 \mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)})}{(\sigma^{(k)})^3((\sigma^{(k)})^2 + (\lambda^{(k)})^2)^2} \\
&+ \frac{(\lambda^{(k)})^2 \sum_{i=1}^{n} \mathrm{E}(g_i z_i^2 \mid x_i, \boldsymbol{\theta}^{(k)})}{(\sigma^{(k)})^3} - \frac{2\lambda^{(k)} \sum_{i=1}^{n}(x_i - \mu^{(k)})\mathrm{E}(g_i z_i \mid x_i, \boldsymbol{\theta}^{(k)})}{(\sigma^{(k)})^3}
\end{aligned}$$

$$
\begin{aligned}
&= -\frac{n}{\sigma^{(k)}} + \frac{\sum_{i=1}^{n}(x_i - \mu^{(k)})^2 \mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)}) + (\lambda^{(k)})^2 \sum_{i=1}^{n} \mathrm{E}(g_i z_i^2 \mid x_i, \boldsymbol{\theta}^{(k)})}{(\sigma^{(k)})^3} \\
&- \frac{2\lambda^{(k)} \sum_{i=1}^{n}(x_i - \mu^{(k)}) \mathrm{E}(g_i z_i \mid x_i, \boldsymbol{\theta}^{(k)})}{(\sigma^{(k)})^3}
\end{aligned}
$$

Setting the derivative zero and solving it for σ^2 yields

$$
\begin{aligned}
\Rightarrow (\sigma^{(k+1)})^2 &= \frac{\sum_{i=1}^{n}(x_i - \mu^{(k+1)})^2 \mathrm{E}(g_i \mid x_i, \boldsymbol{\theta}^{(k)}) + (\lambda^{(k+1)})^2 \sum_{i=1}^{n} \mathrm{E}(g_i z_i^2 \mid x_i, \boldsymbol{\theta}^{(k)})}{n} \\
&- \frac{2\lambda^{(k+1)} \sum_{i=1}^{n}(x_i - \mu^{(k+1)}) \mathrm{E}(g_i z_i \mid x_i, \boldsymbol{\theta}^{(k)})}{n}
\end{aligned}
$$

□

Bibliography

Agresti, A. (2007). *An Introduction to Categorical Data Analysis*. Wiley Series in Probability and Statistics. Wiley.

Aigner, D., C. K. Lovell, and P. Schmidt (1977). "Formulation and estimation of stochastic frontier production function models." *Journal of econometrics* 6.1, pp. 21–37.

Aitkin, M. and D. B. Rubin (1985). "Estimation and hypothesis testing in finite mixture models." *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 67–75.

Aldrich, J. (1997). "RA Fisher and the making of maximum likelihood 1912-1922." *Statistical science*, pp. 162–176.

Anderson, E. and E. Thompson (2002). "A model-based method for identifying species hybrids using multilocus genetic data." *Genetics* 160.3, pp. 1217–1229.

Andrews, D. F. and C. L. Mallows (1974). "Scale mixtures of normal distributions." *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 99–102.

Ankerst, D. P., J. Höfler, S. Bock, P. J. Goodman, A. Vickers, J. Hernandez, L. J. Sokoll, M. G. Sanda, J. T. Wei, R. J. Leach, et al. (2014). "Prostate Cancer Prevention Trial risk calculator 2.0 for the prediction of low-vs high-grade prostate cancer." *Urology* 83.6, pp. 1362–1368.

Ankerst, D. P., T. Koniarski, Y. Liang, R. J. Leach, Z. Feng, M. G. Sanda, A. W. Partin, D. W. Chan, J. Kagan, L. Sokoll, et al. (2012). "Updating risk prediction tools: a case study in prostate cancer." *Biometrical Journal* 54.1, pp. 127–142.

Ankerst, D. P., J. Groskopf, J. R. Day, A. Blase, H. Rittenhouse, B. H. Pollock, C. Tangen, D. Parekh, R. J. Leach, and I. Thompson (2008). "Predicting prostate cancer risk through incorporation of prostate cancer gene 3." *The Journal of urology* 180.4, pp. 1303–1308.

Antoniak, C. E. (1974). "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems." *The annals of statistics*, pp. 1152–1174.

Arellano-Valle, R. and A. Azzalini (2006). "On the Unification of Families of Skew-normal Distributions." *Scandinavian Journal of Statistics* 33.3, pp. 561–574.

Arellano-Valle, R., H. Bolfarine, and V. Lachos (2007). "Bayesian inference for skew-normal linear mixed models." *Journal of Applied Statistics* 34.6, pp. 663–682.

El-Arini, K. (2008). "Dirichlet Processes: a Gentle Tutorial." In: *Select Lab Meeting*. Vol. 10.

Arnold, B. C., R. J. Beaver, A. Azzalini, N. Balakrishnan, A. Bhaumik, D. Dey, C. Cuadras, and J. M. Sarabia (2002). "Skewed multivariate models related to hidden truncation and/or selective reporting." *Test* 11.1, pp. 7–54.

Azzalini, A. (1986). "Further results on a class of distributions which includes the normal ones." *Statistica 46*, pp. 199–208.

Azzalini, A. (2019). *The R package sn: The Skew-Normal and Related Distributions such as the Skew-t (version 1.5-4)*. Università di Padova, Italia. URL: http://azzalini.stat.unipd.it/SN.

Azzalini, A. (2005). "The skew-normal distribution and related multivariate families." *Scandinavian Journal of Statistics* 32.2, pp. 159–188.

Azzalini, A. and A. Capitanio (1999). “Statistical applications of the multivariate skew normal distribution.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61.3, pp. 579–602.

Azzalini, A. and A. D. Valle (1996). “The multivariate skew-normal distribution.” *Biometrika* 83.4, pp. 715–726.

Beard, T. R., S. B. Caudill, and D. M. Gropper (1991). “Finite mixture estimation of multiproduct cost functions.” *The review of economics and statistics*, pp. 654–664.

Beven, K. and A. Binley (1992). “The future of distributed models: model calibration and uncertainty prediction.” *Hydrological processes* 6.3, pp. 279–298.

Birnbaum, Z. W. (1950). “Effect of linear truncation on a multinormal population.” *The Annals of Mathematical Statistics* 21.2, pp. 272–279.

Blackwell, D. and J. B. MacQueen (1973). “Ferguson distributions via Pólya urn schemes.” *The annals of statistics*, pp. 353–355.

Böck, A. (2014). “Statistical modeling of risk and trends in the life sciences with applications to forestry, plant breeding, phenology, and cancer.” Dissertation. München: Technische Universität München.

Böck, A., J. Dieler, P. Biber, H. Pretzsch, and D. P. Ankerst (2014). “Predicting tree mortality for European beech in southern Germany using spatially explicit competition indices.” *Forest Science* 60.4, pp. 613–622.

Branco, M. D. and D. K. Dey (2001). “A general class of multivariate skew-elliptical distributions.” *Journal of Multivariate Analysis* 79.1, pp. 99–113.

Branco, M. and D. Dey (2002). “Regression model under skew elliptical error distribution.” *Journal of Mathematical Sciences* 1, pp. 151–169.

Bravo, F., M. Fabrika, C. Ammer, S. Barreiro, K. Bielak, L. Coll, T. Fonseca, A. Kangur, M. Löf, K. Merganičová, et al. (2019). “Modelling approaches for mixed forests dynamics prognosis. Research gaps and opportunities.” *Forest Systems* 28.1, eR002.

Buchinsky, M. (1994). “Changes in the US wage structure 1963-1987: Application of quantile regression.” *Econometrica: Journal of the Econometric Society*, pp. 405–458.

Cabral, C. R. B., V. H. Lachos, and M. O. Prates (2012). “Multivariate mixture modeling using skew-normal independent distributions.” *Computational Statistics & Data Analysis* 56.1, pp. 126–142.

Caillette, F., A. Galata, and T. Howard (2005). “Real-Time 3-D Human Body Tracking using Variable Length Markov Models.” In: *BMVC*. Vol. 1.

Cambanis, S., S. Huang, and G. Simons (1981). “On the theory of elliptically contoured distributions.” *Journal of Multivariate Analysis* 11.3, pp. 368–385.

Cameron, A. C. and P. K. Trivedi (2013). *Regression analysis of count data*. Vol. 53. Cambridge university press.

Carter, H. B., P. C. Albertsen, M. J. Barry, R. Etzioni, S. J. Freedland, K. L. Greene, L. Holmberg, P. Kantoff, B. R. Konety, M. H. Murad, et al. (2013). “Early detection of prostate cancer: AUA Guideline.” *The Journal of urology* 190.2, pp. 419–426.

Chang, M. and P. Sangrey (2019). *Bypassing the Curse of Dimensionality: Feasible Multivariate Density Estimation*.

Chang, W., J. Cheng, J. Allaire, Y. Xie, and J. McPherson (2019). *shiny: Web Application Framework for R*. R package version 1.3.2. URL: https://CRAN.R-project.org/package=shiny.

Cheng, B. and D. M. Titterington (1994). “Neural networks: A review from a statistical perspective.” *Statistical science*, pp. 2–30.

Dagum, L. and R. Menon (1998). “OpenMP: an industry standard API for shared-memory programming.” *IEEE computational science and engineering* 5.1, pp. 46–55.

Dalrymple, M. L., I. L. Hudson, and R. P. K. Ford (2003). “Finite mixture, zero-inflated Poisson and hurdle models with application to SIDS.” *Computational Statistics & Data Analysis* 41.3-4, pp. 491–504.

Deb, P. and P. K. Trivedi (1997). “Demand for medical care by the elderly: a finite mixture approach.” *Journal of applied Econometrics*, pp. 313–336.

Dechter, R. (1986). *Learning while searching in constraint-satisfaction problems*. University of California, Computer Science Department, Cognitive Systems Laboratory.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). “Maximum likelihood from incomplete data via the EM algorithm.” *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38.

Dunning, M. J., S. L. Vowler, E. Lalonde, H. Ross-Adams, P. Boutros, I. G. Mills, A. G. Lynch, and A. D. Lamb (2017). “Mining human prostate cancer datasets: The “camcAPP” Shiny App.” *EBioMedicine* 17, pp. 5–6.

Duong, T. and M. L. Hazelton (2005). “Convergence rates for unconstrained bandwidth matrix selectors in multivariate kernel density estimation.” *Journal of Multivariate Analysis* 93.2, pp. 417–433.

Eberlein, E., U. Keller, et al. (1995). “Hyperbolic distributions in finance.” *Bernoulli* 1.3, pp. 281–299.

Escobar, M. D. and M. West (1995). “Bayesian density estimation and inference using mixtures.” *Journal of the american statistical association* 90.430, pp. 577–588.

Fabrika, M., H. Pretzsch, and F. Bravo (2018). “Models for mixed forests.” In: *Dynamics, Silviculture and Management of Mixed Forests*. Springer, pp. 343–380.

Fang, K. W. (2018). *Symmetric multivariate and related distributions*. CRC Press.

Ferguson, T. S. (1973). “A Bayesian analysis of some nonparametric problems.” *The annals of statistics*, pp. 209–230.

Finkenstadt, B. and H. Rootzén (2003). *Extreme values in finance, telecommunications, and the environment*. CRC Press.

Fortin, M., S. Bédard, J. DeBlois, and S. Meunier (2008). “Predicting individual tree mortality in northern hardwood stands under uneven-aged management in southern Québec, Canada.” *Annals of Forest Science* 65.2, pp. 205–205.

Francis, R. M. (2017). “POPHELPER: an R package and web app to analyse and visualize population structure.” *Molecular Ecology Resources* 17.1, pp. 27–32.

Freedman, A. N., D. Seminara, M. H. Gail, P. Hartge, G. A. Colditz, R. Ballard-Barbash, and R. M. Pfeiffer (2005). “Cancer risk prediction models: a workshop on development, evaluation, and application.” *Journal of the National Cancer Institute* 97.10, pp. 715–723.

Freedman, A. N., M. L. Slattery, R. Ballard-Barbash, G. Willis, B. J. Cann, D. Pee, M. H. Gail, and R. M. Pfeiffer (2009). "Colorectal cancer risk prediction tool for white men and women without known susceptibility." *Journal of clinical oncology* 27.5, p. 686.

Frühwirth-Schnatter, S. and S. Pyne (2010). "Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-t distributions." *Biostatistics* 11.2, pp. 317–336.

Gail, M. H. (2015). "Twenty-five years of breast cancer risk models and their applications." *JNCI: Journal of the National Cancer Institute* 107.5.

Gail, M. H., L. A. Brinton, D. P. Byar, D. K. Corle, S. B. Green, C. Schairer, and J. J. Mulvihill (1989). "Projecting individualized probabilities of developing breast cancer for white females who are being examined annually." *Journal of the National Cancer Institute* 81.24, pp. 1879–1886.

Gail, M. H. and J. P. Costantino (2001). *Validating and improving models for projecting the absolute risk of breast cancer*.

Gail, M. H. and P. L. Mai (2010). *Comparing breast cancer risk assessment models*.

Galassi, M., J. Davies, J. Theiler, B. Gough, G. Jungman, M. Booth, and F. Rossi (2009). "GNU Scientific Library Reference Manual (Network Theory Ltd., 2009)." *URL http://www. gnu. org/s/gsl*.

Genton, M. G. (2004). *Skew-elliptical distributions and their applications: a journey beyond normality*. CRC Press.

Georgii, H.-O. (2013). *Stochastics: introduction to probability and statistics*. Walter de Gruyter.

Goes, P. B. (2014). "Design science research in top information systems journals." *MIS Quarterly: Management Information Systems* 38.1, pp. iii–viii.

Green, P. J. (1995). "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination." *Biometrika*, pp. 711–732.

Green, P. J. and D. I. Hastie (2009). "Reversible jump MCMC." *Genetics* 155.3, pp. 1391–1403.

Grill, S., D. P. Ankerst, M. H. Gail, N. Chatterjee, and R. M. Pfeiffer (2017). "Comparison of approaches for incorporating new information into existing risk prediction models." *Statistics in medicine* 36.7, pp. 1134–1156.

Grill, S., M. Fallah, R. J. Leach, I. M. Thompson, S. Freedland, K. Hemminki, and D. P. Ankerst (2015a). "Incorporation of detailed family history from the Swedish Family Cancer Database into the PCPT risk calculator." *The Journal of urology* 193.2, pp. 460–465.

Grill, S., M. Fallah, R. J. Leach, I. M. Thompson, K. Hemminki, and D. P. Ankerst (2015b). "A simple-to-use method incorporating genomic markers into prostate cancer risk prediction tools facilitated future validation." *Journal of clinical epidemiology* 68.5, pp. 563–573.

Gupta, A. (2003). "Multivariate skew t-distribution." *Statistics: A Journal of Theoretical and Applied Statistics* 37.4, pp. 359–363.

Hao, L., D. Naiman, D. Naiman, and i. Sage Publications (2007). *Quantile Regression*. Quantile Regression no. 149. SAGE Publications. ISBN: 9781412926287.

Hartmann, H., C. F. Moura, W. R. Anderegg, N. K. Ruehr, Y. Salmon, C. D. Allen, S. K. Arndt, D. D. Breshears, H. Davi, D. Galbraith, et al. (2018). "Research frontiers for improving our understanding of drought-induced tree and forest mortality." *New Phytologist* 218.1, pp. 15–28.

Ho, H. J. and T.-I. Lin (2010). "Robust linear mixed models using the skew t distribution with application to schizophrenia data." *Biometrical Journal* 52.4, pp. 449–469.

Höfler, J. (2019). *fitmixst4: Fitting mixture of skew t distribution*. R package version 0.281. URL: https://R-Forge.R-project.org/projects/fitmixst/.

Holmes, C., D. Denison, and B. Mallick (1999). "Bayesian partitioning for classification and regression." *Manuscript, Imperial College*.

Hornik, K., F. Leisch, and A. Zeileis (2003). "JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling." In: *Proceedings of DSC*. Vol. 2, pp. 1–1.

Jahanshiri, E. and A. R. M. Shariff (2014). "Developing web-based data analysis tools for precision farming using R and Shiny." In: *IOP Conference Series: Earth and Environmental Science*. Vol. 20. 1. IOP Publishing, p. 012014.

Jie, M., G. S. Collins, E. W. Steyerberg, J. Y. Verbakel, B. van Calster, et al. (2019). "A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models." *Journal of clinical epidemiology*.

Johnson, T. D., R. M. Elashoff, and S. J. Harkema (2003). "A bayesian change-point analysis of electromyographic data: Detecting muscle activation patterns and associated applications." *Biostatistics* 4.1, pp. 143–164.

Jones, M. and M. Faddy (2003). "A skew extension of the t-distribution, with applications." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65.1, pp. 159–174.

Kelker, D. (1970). "Distribution theory of spherical distributions and a location-scale parameter generalization." *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 419–430.

Kiernan, D., E. Bevilacqua, R. Nyland, and L. Zhang (2009). "Modeling tree mortality in low-to medium-density uneven-aged hardwood stands under a selection system using generalized estimating equations." *Forest science* 55.4, pp. 343–351.

King, G. and L. Zeng (2001). "Logistic regression in rare events data." *Political analysis* 9.2, pp. 137–163.

Klotz, L., L. Zhang, A. Lam, R. Nam, A. Mamedov, and A. Loblaw (2009). "Clinical results of long-term follow-up of a large, active surveillance cohort with localized prostate cancer." *Journal of Clinical Oncology* 28.1, pp. 126–131.

Koenker, R., V. Chernozhukov, X. He, and L. Peng (2017). *Handbook of Quantile Regression*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press.

Koenker, R. (2019). *quantreg: Quantile Regression*. R package version 5.51. URL: https://CRAN.R-project.org/package=quantreg.

Koenker, R. and G. Bassett Jr (1978). "Regression quantiles." *Econometrica: journal of the Econometric Society*, pp. 33–50.

Koenker, R. and K. F. Hallock (2001). "Quantile regression." *Journal of economic perspectives* 15.4, pp. 143–156.

Kotz, S. and S. Nadarajah (2004). *Multivariate t-distributions and their applications*. Cambridge University Press.

Lachos, V. H., P. Ghosh, and R. Arellano-Valle (2010). "Likelihood based inference for skew-normal independent linear mixed models." *Statistica Sinica*, pp. 303–322.

Lachos, V. H., H. Bolfarine, R. Arellano-Valle, and L. C. Montenegro (2007). "Likelihood-based inference for multivariate skew-normal regression models." *Communications in Statistics—Theory and Methods* 36.9, pp. 1769–1786.

Lambert, D. (1992). "Zero-inflated Poisson regression, with an application to defects in manufacturing." *Technometrics* 34.1, pp. 1–14.

Laney, D. (2001). "3D data management: Controlling data volume, velocity and variety." *META Group Research Note* 6.70.

Lange, K. L., R. J. Little, and J. M. Taylor (1989). "Robust statistical modeling using the t distribution." *Journal of the American Statistical Association* 84.408, pp. 881–896.

Lee, S. and G. McLachlan (2011). "On the fitting of mixtures of multivariate skew t-distributions via the EM algorithm." *arXiv preprint arXiv:1109.4706*.

Lee, S. and G. McLachlan (2012). "EMMIX-uskew: an R package for fitting mixtures of multivariate skew t-distributions via the EM algorithm." *arXiv preprint arXiv:1211.5290*.

Lee, S. and G. McLachlan (2014). "Finite mixtures of multivariate skew t-distributions: some recent and new results." *Statistics and Computing* 24.2, pp. 181–202.

Lee, S. and G. Mclachlan (2013). "On mixtures of skew normal and skew t -distributions." *Advances in Data Analysis and Classification* 7.3, pp. 241–266.

Lee, S., G. McLachlan, M. S. X. Lee, and M. Depends (2014). "Package EMMIXuskewt."

Li, H.-D., Q.-S. Xu, and Y.-Z. Liang (2012). "Random frog: an efficient reversible jump Markov Chain Monte Carlo-like approach for variable selection with applications to gene selection and disease classification." *Analytica Chimica Acta* 740, pp. 20–26.

Lin, T. I., J. C. Lee, and W. J. Hsieh (2007). "Robust mixture modeling using the skew t distribution." *Statistics and computing* 17.2, pp. 81–92.

Lin, T.-I. (2010). "Robust mixture modeling using multivariate skew t distributions." *Statistics and Computing* 20.3, pp. 343–356.

Lin, T.-I., H. J. Ho, and C.-R. Lee (2014). "Flexible mixture modelling using the multivariate skew-t-normal distribution." *Statistics and Computing* 24.4, pp. 531–546.

Lindsay, B. G. (1995). *Mixture models: theory, geometry and applications*.

Liseo, B. and N. Loperfido (2003). "A Bayesian interpretation of the multivariate skew-normal distribution." *Statistics & probability letters* 61.4, pp. 395–401.

Liu, C. and D. B. Rubin (1994). "The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence." *Biometrika* 81.4, pp. 633–648.

Lloyd, S. (1957). "Least square quantization in PCM. Bell Telephone Laboratories Paper. Published in journal much later: Lloyd, SP: Least squares quantization in PCM." *IEEE Trans. Inform. Theor.(1957/1982)*.

Lu, L., H. Jiang, and W. H. Wong (2013). "Multivariate density estimation by bayesian sequential partitioning." *Journal of the American Statistical Association* 108.504, pp. 1402–1410.

Lunn, D. J., A. Thomas, N. Best, and D. Spiegelhalter (2000). "WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility." *Statistics and computing* 10.4, pp. 325–337.

Maalouf, M., D. Homouz, and T. B. Trafalis (2018). "Logistic regression in large rare events and imbalanced data: A performance comparison of prior correction and weighting methods." *Computational Intelligence* 34.1, pp. 161–174.

MacQueen, J. et al. (1967). "Some methods for classification and analysis of multivariate observations." In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Vol. 1. 14. Oakland, CA, USA., pp. 281–297.

Marden, J. I. (1998). "Bivariate QQ-plots and spider web plots." *Statistica Sinica*, pp. 813–826.

Marr, B. (2014). "Big data: The 5 Vs everyone must know." *LinkedIn Pulse* 6.

Matiu, M., D. P. Ankerst, and A. Menzel (2016). "Asymmetric trends in seasonal temperature variability in instrumental records from ten stations in Switzerland, Germany and the UK from 1864 to 2012." *International Journal of Climatology* 36.1, pp. 13–27.

McLachlan, G. and K. E. Basford (1988). "Mixture models: Inference and applications to clustering." 84.

McLachlan, G. and T. Krishnan (2007). *The EM algorithm and extensions*. Vol. 382. John Wiley & Sons.

McLachlan, G. and D. Peel (2004). *Finite mixture models*. John Wiley & Sons.

McMurdie, P. J. and S. Holmes (2014). "Shiny-phyloseq: web application for interactive microbiome analysis with provenance tracking." *Bioinformatics* 31.2, pp. 282–283.

Meng, X.-L. and D. B. Rubin (1993). "Maximum likelihood estimation via the ECM algorithm: A general framework." *Biometrika* 80.2, pp. 267–278.

Metsalu, T. and J. Vilo (2015). "ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap." *Nucleic acids research* 43.W1, W566–W570.

Monserud, R. A. and H. Sterba (1999). "Modeling individual tree mortality for Austrian forest species." *Forest Ecology and Management* 113.2, pp. 109–123.

Moons, K. G., A. P. Kengne, D. E. Grobbee, P. Royston, Y. Vergouwe, D. G. Altman, and M. Woodward (2012). "Risk prediction models: II. External validation, model updating, and impact assessment." *Heart*, heartjnl–2011.

Moyer, V. A. (2012). "Screening for prostate cancer: US Preventive Services Task Force recommendation statement." *Annals of internal medicine* 157.2, pp. 120–134.

Murphy, A. H. (1973). "A new vector partition of the probability score." *Journal of Applied Meteorology* 12.4, pp. 595–600.

Nelson, J. W., J. Sklenar, A. P. Barnes, and J. Minnier (2017). "The START App: a web-based RNAseq analysis and visualization resource." *Bioinformatics* 33.3, pp. 447–449.

Nelson, L. S. (1964). "The sum of values from a normal and a truncated normal distribution." *Technometrics* 6, pp. 469–471.

Nisa, K. K., H. A. Andrianto, and R. Mardhiyyah (2014). "Hotspot clustering using DBSCAN algorithm and shiny web framework." In: *Advanced Computer Science and Information Systems (ICACSIS), 2014 International Conference on*. IEEE, pp. 129–132.

O'hagan, A. (1976). *Moments of the truncated multivariate-t distribution*. URL: http://www.tonyohagan.co.uk/academic/pdf/trunc_multi_t.PDF.

O'hagan, A. and T. Leonard (1976). "Bayes estimation subject to uncertainty about parameter constraints." *Biometrika* 63.1, pp. 201–203.

OpenMP, A. (2013). "OpenMP application program interface version 4.0."

Owen, D. B. (1980). "A table of normal integrals: A table." *Communications in Statistics-Simulation and Computation* 9.4, pp. 389–419.

Owen, J. and R. Rabinovitch (1983). "On the class of elliptical distributions and their applications to the theory of portfolio choice." *The Journal of Finance* 38.3, pp. 745–752.

Papastamoulis, P., T. Furukawa, N. Van Rhijn, M. Bromley, E. Bignell, and M. Rattray (2017). "Bayesian detection of piecewise linear trends in replicated time-series with application to growth data modelling." *arXiv preprint arXiv:1709.06111*.

Pearson, K. (1894). "Contributions to the mathematical theory of evolution." *Philosophical Transactions of the Royal Society of London. A* 185, pp. 71–110.

Peel, D. and G. McLachlan (2000). "Robust mixture modelling using the t distribution." *Statistics and computing* 10.4, pp. 339–348.

Pernkopf, F. and D. Bouchaffra (2005). "Genetic-based EM algorithm for learning Gaussian mixture models." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27.8, pp. 1344–1348.

Pham, D. L., C. Xu, and J. L. Prince (2000). "Current methods in medical image segmentation." *Annual review of biomedical engineering* 2.1, pp. 315–337.

Plummer, M. (2012). "JAGS: Just another Gibbs sampler." *Astrophysics Source Code Library*.

Prates, M. O., V. H. Lachos, and C. Cabral (2013). "mixsmsn: Fitting finite mixture of scale mixture of skew-normal distributions." *Journal of Statistical Software* 54.12, pp. 1–20.

Pretzsch, H. (1992). "Modellierung der kronenkonkurrenz von fichte und buche in rein-und mischbeständen." *Allgemeine Forst-und Jagdzeitung* 163.11/12, pp. 203–213.

Pretzsch, H. (2001). *Modellierung des Waldwachstums*. Blackwell Wissenschafts-Verlag.

Pretzsch, H., P. Biber, and J. Ďurskỳ (2002). "The single tree-based stand simulator SILVA: construction, application and evaluation." *Forest ecology and management* 162.1, pp. 3–21.

Puhr, R., G. Heinze, M. Nold, L. Lusa, and A. Geroldinger (2017). "Firth's logistic regression with rare events: accurate effect estimates and predictions?" *Statistics in medicine* 36.14, pp. 2302–2317.

Rasmussen, C. E. (2000). "The infinite Gaussian mixture model." In: *Advances in neural information processing systems*, pp. 554–560.

Redner, R. A. and H. F. Walker (1984). "Mixture densities, maximum likelihood and the EM algorithm." *SIAM review* 26.2, pp. 195–239.

Richardson, S. and P. J. Green (1997). "On Bayesian analysis of mixtures with an unknown number of components (with discussion)." *Journal of the Royal Statistical Society: series B (statistical methodology)* 59.4, pp. 731–792.

Ripley, B. D. (1993). "Statistical aspects of neural networks." *Networks and chaos—statistical and probabilistic aspects* 50, pp. 40–123.

Rizzo, M. L. (2007). *Statistical computing with R*. CRC Press.

Roberts, C. (1966). "A correlation model useful in the study of twins." *Journal of the American Statistical Association* 61.316, pp. 1184–1190.

RStudio, Inc (2019). *shiny: Easy web applications in R*. URL: http://shiny.rstudio.com.

Sahu, S. K., D. K. Dey, and M. D. Branco (2003). "A new class of multivariate skew distributions with applications to Bayesian regression models." *Canadian Journal of Statistics* 31.2, pp. 129–150.

Salas-Eljatib, C., A. Fuentes-Ramirez, T. G. Gregoire, A. Altamirano, and V. Yaitul (2018). "A study on the effects of unbalanced data when fitting logistic regression models in ecology." *Ecological Indicators* 85, pp. 502–508.

Sanderson, C. (2010). "Armadillo: An open source C++ linear algebra library for fast prototyping and computationally intensive experiments."

Santago, P. and H. D. Gage (1993). "Quantification of MR brain images by mixture density and partial volume modeling." *IEEE Transactions on Medical Imaging* 12.3, pp. 566–574.

Schlattmann, P. (2009). *Medical applications of finite mixture models*. Springer.

Schmidhuber, J. (2016). "Deep learning." *Encyclopedia of Machine Learning and Data Mining*, pp. 1–11.

Scott, D. W. (2015). *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons.

Self, S. G. and K.-Y. Liang (1987). "Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions." *Journal of the American Statistical Association* 82.398, pp. 605–610.

Sexton, J. and A. R. Swensen (2000). "ECM algorithms that converge at the rate of EM." *Biometrika* 87.3, pp. 651–662.

Silverman, B. W. (2018). *Density estimation for statistics and data analysis*. Routledge.

Sisson, S. A. (2005). "Transdimensional Markov chains: A decade of progress and future perspectives." *Journal of the American Statistical Association* 100.471, pp. 1077–1089.

Sokoll, L. J., M. G. Sanda, Z. Feng, J. Kagan, I. A. Mizrahi, D. L. Broyles, A. W. Partin, S. Srivastava, I. M. Thompson, J. T. Wei, et al. (2010). "A prospective, multicenter, National Cancer Institute Early Detection Research Network study of [- 2] proPSA: improving prostate cancer detection and correlating with cancer aggressiveness." *Cancer Epidemiology and Prevention Biomarkers* 19.5, pp. 1193–1200.

Spitzer, M., J. Wildenhain, J. Rappsilber, and M. Tyers (2014). "BoxPlotR: a web tool for generation of box plots." *Nature methods* 11.2, p. 121.

Steyerberg, E. W. (2008). *Clinical prediction models: a practical approach to development, validation, and updating*. Springer Science & Business Media.

Stigler, S. M. (1984). "Studies in the history of probability and statistics XL Boscovich, Simpson and a 1760 manuscript note on fitting a linear relation." *Biometrika* 71.3, pp. 615–620.

Tanner, M. A. and W. H. Wong (1987). "The calculation of posterior distributions by data augmentation." *Journal of the American statistical Association* 82.398, pp. 528–540.

Thompson, I. M., D. P. Ankerst, C. Chi, P. J. Goodman, C. M. Tangen, M. S. Lucia, Z. Feng, H. L. Parnes, and C. A. Coltman Jr (2006). "Assessing prostate cancer risk: results from the Prostate Cancer Prevention Trial." *Journal of the National Cancer Institute* 98.8, pp. 529–534.

Thompson, T. J., P. J. Smith, and J. P. Boyle (1998). "Finite mixture models with concomitant information: assessing diagnostic criteria for diabetes." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 47.3, pp. 393–404.

Titterington, D., P. Titterington, A. Smith, and U. Makov (1985). *Statistical Analysis of Finite Mixture Distributions*. Applied section. Wiley.

Tohka, J., E. Krestyannikov, I. D. Dinov, A. M. Graham, D. W. Shattuck, U. Ruotsalainen, and A. W. Toga (2007). "Genetic algorithms for finite mixture model based voxel classification in neuroimaging." *IEEE transactions on medical imaging* 26.5, pp. 696–711.

Twitter, Inc (2019). *Bootstrap · Build responsive, mobile-first projects on the web with the world's most popular front-end component library.* URL: http://getbootstrap.com/.

Van Praag, B. M. and B. M. Wesselman (1989). "Elliptical multivariate analysis." *Journal of Econometrics* 41.2, pp. 189–203.

Venables, W. N. and B. D. Ripley (2013). *Modern applied statistics with S-PLUS*. Springer Science & Business Media.

Vickers, A. J., A. M. Cronin, M. J. Roobol, J. Hugosson, J. S. Jones, M. W. Kattan, E. Klein, F. Hamdy, D. Neal, J. Donovan, et al. (2010). "The relationship between prostate-specific antigen and prostate cancer risk: the Prostate Biopsy Collaborative Group." *Clinical Cancer Research* 16.17, pp. 4374–4381.

Vickers, A. J. and E. B. Elkin (2006). "Decision curve analysis: a novel method for evaluating prediction models." *Medical Decision Making* 26.6, pp. 565–574.

Vilca-Labra, F. and V. Leiva-Sánchez (2006). "A new fatigue life model based on the family of skew-elliptical distributions." *Communications in Statistics—Theory and Methods* 35.2, pp. 229–244.

Walker, J. S. (2014). *Big data: A revolution that will transform how we live, work, and think*.

Wang, Y. and T. Lei (1994). "Statistical analysis of MR imaging and its applications in image modeling." In: *Image Processing, 1994. Proceedings. ICIP-94., IEEE International Conference*. Vol. 1. IEEE, pp. 866–870.

Wedel, M. and W. S. DeSarbo (2002). "Market segment derivation and profiling via a finite mixture model framework." *Marketing Letters* 13.1, pp. 17–25.

West, M. and M. D. Escobar (1993). *Hierarchical priors and mixture models, with application in regression and density estimation*. Institute of Statistics and Decision Sciences, Duke University.

Wickham, H., P. Danenberg, and M. Eugster (2018). *roxygen2: In-Line Documentation for R*. R package version 6.1.1. URL: https://CRAN.R-project.org/package=roxygen2.

Wilt, T. J., M. K. Brawer, K. M. Jones, M. J. Barry, W. J. Aronson, S. Fox, J. R. Gingrich, J. T. Wei, P. Gilhooly, B. M. Grob, et al. (2012). "Radical prostatectomy versus observation for localized prostate cancer." *New England Journal of Medicine* 367.3, pp. 203–213.

Yee, T. W. (2015). *Vector Generalized Linear and Additive Models: With an Implementation in R*. New York, USA: Springer.

Yee, T. W. (2019). *VGAM: Vector Generalized Linear and Additive Models*. R package version 1.1-1. URL: https://CRAN.R-project.org/package=VGAM.

Yee, T. W. et al. (2010). "The VGAM package for categorical data analysis." *Journal of Statistical Software* 32.10, pp. 1–34.

Yee, T. W. and C. Wild (1996). "Vector generalized additive models." *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 481–493.

Yu, K., Z. Lu, and J. Stander (2003). "Quantile regression: applications and current research areas." *Journal of the Royal Statistical Society: Series D (The Statistician)* 52.3, pp. 331–350.

Zhai, C. (2007). "A note on the expectation-maximization (em) algorithm." *Course note of CS410*.