Akademie věd České republiky
Ústav teorie informace a automatizace

Academy of Sciences of the Czech Republic
Institute of Information Theory and Automation

RESEARCH REPORT

Josef Andrýsek

Projection Based Estimation of Dynamic Probabilistic Mixtures

No. 2098 April 2004

AV ČR S1075351, GA ČR 102/03/0049

ÚTIA AV ČR, P. O. Box 18, 182 08 Prague, Czech Republic
Telex: 122018 atom c, Fax: (+420)(2)6884 903
E-mail: utia@utia.cas.cz

Contents

Chapter 1

Introduction

The choice of a suitable model is essential for both control and decision making when dealing with complex systems. One way to face complexity is the principle of adaptivity, i.e. using models which evolve during their use. The demand for adaptivity of the model leads to the recursive estimation of its parameters, i.e. permanent updating of its parameter estimates by the new data. In other words, statistics describing estimates are corrected by newly acquired data. The model should be chosen from a sufficiently rich family of models to capture all properties of the modelled system. Naturally, computational cost associated with estimation of parameters of the model grows with complexity of the model. If the modelled system is non-linear, its model should be non-linear too. In this paper, we study finite probabilistic mixture of linear models. The finite mixtures provide a universal approximation of almost any probabilistic density function [1] and thus can be successfully used in modelling of complex systems. Invoking the principle of adaptivity, we seek an efficient recursive estimation of the mixture model parameters.

The resulting model can be then used both for control and decision making tasks. Universal algorithms for mixture-based control [2] were derived, but quality of the resulting control strategy strongly depends on the quality of the estimated model. Practical experience indicates that this is a weak element of adaptive control and that an improvement of the estimation part improves the overall control quality. Hence, we try to develop better estimation algorithms for the mixture model. The control algorithms [2] as well as efficient structure estimation algorithms were derived using the Bayesian theory [3]. The unknown model parameters are treated as random variables and all subsequent task are defined in terms of posterior distributions of the parameters rather then thier point estimates.

The recursive Bayesian estimation evaluates the posterior distribution on parameters at time t as an update of the the posterior distribution at time $t-1$ using the Bayes' rule and the data acquired at time t. The recursion starts at $t = 1$ with update of the prior distribution which must be chosen before the estimation starts. The posterior distribution obtained by the Bayes' rule may not be, however, analytically tractable and thus unsuitable for the next update.

In practice, mostly such prior distribution is used so that the posterior distribution in each estimation step has the same functional form as the prior distribution. Hence, just the sufficient statistics determining the posterior density are updated. Such a prior distribution is then known as conjugate with the observation model. For example, conjugate prior distribution is available for all models from the exponential family. If the conjugate prior does not exists the exact recursive estimation can not be achieved. In such a case, we seek approximate recursive estimation. This is the case of the probabilistic mixture model. Using the exact Bayesian update, the complexity of posterior density grows exponentially with number of the data samples. The quasi-Bayes algorithm [4] [2] or a modification of the EM algorithm [1] are examples of approximate algorithms facing this problem.

This paper introduces a new approximate estimation method, which can be viewed as a generalization of quasi-Bayes algorithm. The basis of both approaches is finding the approximate posterior density in particular (well manipulable) class of densities.

The new algorithm finds the optimal projection of the correct Bayesian density into the selected class of densities. The projection is optimal in the sense of Kulback Leibler distance [5]. It should be mentioned that the Kullback Leibler distance is not symmetric. Algorithm presented in this paper minimizes the Kulback Leibler distance with the argument order, which conforms with Bayesian principles [6]. An algorithm minimizing the Kullback Leibler distance with arguments in different order can be found in [7].

Chapter 2

Notions and notations

x^* *denotes the range of* x, $x \in x^*$.

$\mathring{x}$ *denotes the number of entries in the vector* x.

$\equiv$ *means the equality by definition.*

x_t *is a quantity (vector)* x *at the discrete time labelled by* $t \in t^* \equiv \{1, \ldots, \mathring{t}\}$.

$x_{i;t}$ *is an* i*-th entry of the vector* x_t. The semicolon in the subscript indicates that the symbol following it is the time index.

$x_{k\_l;t}$ *is a subvector of the vector* x_t. $x_{k\_l;t} = (x_{k;t}, \cdots, x_{l;t})$.

$x(k\_l) \equiv x_k, \ldots, x_l$.

$x(t) \equiv x(1\_t)$.

$x(t)$ *is an empty sequence and reflects just the prior information if* $t < 1$.

d *is data array,* d_t *is data record at time* t *(vector with entries* $(d_{1;t}, \cdots, d_{\mathring{d}_t;t}$ *)).*

Θ *unknown parameter, finite-dimensional vector*

f, π *are the letters reserved for probability density functions(pdf).*

$f(d_t|d(t-1), \Theta)$ *means model of the system.*

$f_c(d_t|d(t-1), \Theta_c)$ *is component of the mixture.*

$\pi_0(\Theta)$ *denotes prior density of the unknown parameter* Θ.

$\pi_t(\Theta|d(t)) \equiv \pi_t(\Theta|\mathcal{G}_t)$ *means (approximate) posterior density of the parameter* Θ *determined by the sufficient statistic* $\mathcal{G}_t$.

$\propto$ *is the proportion sign,* $h \propto g$ *means that function* h *equals to the function* g *up to the normalization. I.e.* $\frac{h}{\int h} = \frac{g}{\int g}$.

∂ *is the model order.*

$\mathcal{D}(\ ||\ )$ *means the Kullback-Leibler distance[5]. This "distance" is familiarly used in Bayesian analysis as the measure how good the second pdf approximates the first pdf. For conciseness, the Kullback-Leibler distance is referred to as the KL distance.* $\mathcal{D}\left(f \,\middle|\middle|\, g\right) = \int f \ln\left(\frac{f}{g}\right)$

$\Gamma(x)$ *means gamma function,* $\Gamma(x) = \int_0^{+\infty} t^{x-1} \exp(-t) dt$.

$\psi_0(x)$ *is digamma function,* $\psi_0(x) = \frac{\partial \ln \Gamma(x)}{\partial x}$.

δ *denotes identity matrix. I.e.* $\delta_{ij} = 1$ *iff* $i = j$, *otherwise* $\delta_{ij} = 0$.

$\otimes$ *denotes the Kronecker product of two matrices*

Agreement 1 (Multimatrix, multivector) *Multimatrix of type* m, n

$$M = \begin{pmatrix} M_{11} & \cdots & M_{1n} \\ \vdots & \ddots & \vdots \\ M_{m1} & \cdots & M_{mn} \end{pmatrix}$$

is a mathematical object, where M_{ij} *is either matrix or multimatrix. Hence matrix is a multimatrix. Multimatrix need not be a matrix. Definition of Multivector is analogical.*

Agreement 2 (Multimatrix indexing) *For* M *being a multimatrix of type* m, n *the following notation is used:*

M_{ij} *is* ij*-th entry of* M.

$M_{\bullet j}$ *is multimatrix* $\begin{pmatrix} M_{1j} \\ \vdots \\ M_{mj} \end{pmatrix}$.

$M_{i\bullet}$ *is multimatrix* $(M_{i1}, \cdots, M_{in})$.

$M_{\bullet\bullet}$ *means the same as* M*.We use this notation when we want to stress that* M *is a multimatrix (matrix).*

Agreement 3 (Other Matrix notations) *Let's* M *be a matrix of type* m, n *and* c *some scalar. Let's define the following operations:*

$M \pm c$ *is matrix of type* m, n, $(M \pm c)_{ij} = M_{ij} \pm c$.

$\exp(M)$ *is matrix of type* m, n, $(\exp(M))_{ij} = \exp(M_{ij})$.

$\max M$ *is scalar with maximal value of* M.

$\mathbf{vec}M$ *is column vector of length* $m * n$ *containing all columns of* M.

$|M|$ *is determinant of matrix* M.

Chapter 3

Basic elements and tools

3.1 Recursive parameter estimation

The task of recursive parameter estimation is to determine the posterior density $\pi_t(\Theta|d(t))$ based on the knowledge of

- last posterior density $\pi_{t-1}(\Theta|d(t-1))$
- new data record d_t
- model of the system $f\big(d_t|d(t-1),\Theta\big)$ parameterized by unknown parameter Θ .

The algorithm starts from prior pdf $\pi_0(\Theta) \equiv \pi_0(\Theta|d(0))$. We assume existence of the sufficient statistic $\mathcal{G}_t$ for posterior pdfs, i.e.

$$\pi_t(\Theta|d(t)) \equiv \pi_t(\Theta|\mathcal{G}_t).$$

Next, consider that the actual data record d_t doesn't depend on all historical data $d(t-1)$ but only on a subset $\phi_{t-1} = (d_{t-1}, d_{t-2}, \cdots, d_{t-\partial})$. Hence,

$$f\big(d_t|d(t-1),\Theta\big) \equiv f\big(d_t|\phi_{t-1},\Theta\big).$$

The standard Bayesian approach determines $\pi_t(\Theta|\mathcal{G}_t)$ as

$$\pi_t(\Theta|\mathcal{G}_t) \propto f\big(d_t|\phi_{t-1},\Theta\big)\pi_{t-1}(\Theta|\mathcal{G}_{t-1}). \tag{3.1}$$

3.1.1 Recursive parameter estimation with conjugate pdf

The considered approach (3.1) can be effectively used in the case when $\pi_0(\Theta)$ is conjugate pdf to the system model $f\big(d_t|\phi_{t-1},\Theta\big)$. In such a case, $\pi_t(\Theta|\mathcal{G}_t)$ has the same functional form as $\pi_0(\Theta)$. Hence, we can get

$$\pi_t(\Theta|\mathcal{G}_t) \equiv \pi(\Theta|\mathcal{G}_t), \quad \forall t.$$

When updating from $\pi(\Theta|\mathcal{G}_{t-1})$ to $\pi(\Theta|\mathcal{G}_t)$ it suffices to update the sufficient statistics: $(\mathcal{G}_{t-1}, d_t) \longrightarrow \mathcal{G}_t$.

3.1.2 Recursive parameter estimation without conjugate pdf

If the pdf conjugate to the system model doesn't exist, the dimension of sufficient statistic grows with number of data samples. Then, of course, complexity of π_t grows as well. In such a case we can proceed in the following way:

- we choose prior pdf in an arbitrary well manipulable functional form,
- we seek an approximate posterior pdf's of the same functional form,
- we set, in each step of estimation, the statistic determining the approximate posterior pdf in such a way that it is "closest" to the "correct Bayesian" pdf.

We need to specify what we mean by: "correct Bayesian" and "closest" . Let's have the approximate posterior pdf $\pi(\Theta|\mathcal{G}_{t-1})$, which depends on the statistic $\mathcal{G}_{t-1}$. If we handle the approximate posterior pdf $\pi(\Theta|\mathcal{G}_{t-1})$ as the correct posterior pdf, the "correct Bayesian" posterior pdf in the next step $\hat{\pi}(\Theta|\mathcal{G}_{t-1}, d_t, \phi_{t-1})$ is (according to (3.1)) obtained as

$$\hat{\pi}(\Theta|\mathcal{G}_{t-1}, d_t, \phi_{t-1}) = \frac{f\left(d_t|\phi_{t-1}, \Theta\right)\pi(\Theta|\mathcal{G}_{t-1})}{\int f\left(d_t|\phi_{t-1}, \Theta\right)\pi(\Theta|\mathcal{G}_{t-1})d\Theta}.$$

The term "closest" means closest in sense of the KL distance. It means that we want to find $\mathcal{G}_t$ so that

$$\mathcal{D}\left(\hat{\pi}(\Theta|\mathcal{G}_{t-1}, d_t, \phi_{t-1}) \,\middle|\middle|\, \pi(\Theta|\mathcal{G}_t)\right) \tag{3.2}$$

is minimized.

Remarks 1

1. *Applicability of the presented algorithm strictly depends on the complexity of the KL distance. Except of trivial cases, it is usable only if the KL distance can be evaluated analytically.*

2. *The algorithm uses the approximate posterior pdf obtained in step $t-1$ as the true posterior pdf in step t. This leads to error accumulation.*

3.2 Dynamic probabilistic mixture

In this paper, we consider the parameterized model of the system in the form of a finite probabilistic mixture:

$$\begin{aligned}
f(d_t|\phi_{t-1}, \Theta) &\equiv \sum_{c\in c^*} \alpha_c f_c(d_t|\phi_{c;t-1}, \Theta_c),\ c^* = \{1, \ldots, \mathring{c}\},\ \mathring{c} < \infty, \quad \text{where} &(3.3)\\
f_c(d_t|\phi_{c;t-1}, \Theta_c) &\equiv \text{c-th } \textit{component} \text{ given by component parameters } \Theta_c \text{ and the state}\\
\phi_{c;t-1} &\equiv \text{subset of } \phi_{t-1}\\
\alpha_c &\equiv \text{the probabilistic } \textit{component weight}\\
\Theta &\equiv \text{mixture parameter formed by the component weights and parameters}\\
\Theta \in \Theta^* &\equiv \left\{\{\Theta_c \in \Theta_c^*\}_{c\in c^*},\ \alpha \equiv [\alpha_1, \ldots, \alpha_{\mathring{c}}] \in \alpha^* \equiv \left\{\alpha_c \geq 0,\ \sum_{c\in c^*} \alpha_c = 1\right\}\right\}.
\end{aligned}$$

Before fixing and refining nomenclature related to the mixture, we split the individual components into so called *factors* that provide flexibility of the parametric description.

Using the chain rule, the pdfs $f_c(d_t|\phi_{c;t-1}, \Theta_c)$ can be written as a product of pdfs of individual entries of d_t. Before applying the chain rule, entries of d_t can be permuted and some permutations may lead to parameterizations with less parameters. This motivates inclusion of permutations into the model description

$$d \rightarrow d_c \quad \text{with} \quad d_{ic} = d_{j_{ic}}, \quad \text{where} \tag{3.4}$$

j_{ic} is i-th entry of the permuted indices $[1, \ldots, \mathring{d}]$. The assignment (3.4) is applied component-wise and together with the chain rule give

$$f_c(d_t|\phi_{c;t-1}, \Theta_c) = \prod_{i\in i^*} f_{ic}(d_{ic;t}|d_{(i+1)\_\mathring{d}c;t}, \phi_{c;t-1}, \Theta_{ic}) \equiv \prod_{i\in i^*} f_{ic}(d_{ic;t}|\psi_{ic;t}, \Theta_{ic}). \tag{3.5}$$

The additional subscript i of the parameter Θ_{ic} indicates that only some entries of Θ_c may occur in i-th pdf (*factor*) in (3.5). Similarly, the *regression vector* $\psi_{ic;t}$ is generally a sub-vector of the vector

$$[d_{(i+1)\_\mathring{d}c;t}, \phi'_{c;t-1}, 1]'. \tag{3.6}$$

Agreement 4 (Nomenclature related to mixtures)

Pdfs: *The pdf $f_c(d_t|\phi_{c;t-1}, \Theta_c)$ in (3.3) is called* parameterized component *of a mixture and*

α_c *is the* weight of the c-th parameterized component.

The pdf $f_{ic}(d_{ic;t}|\psi_{ic;t}, \Theta_{ic})$ in (3.5) is called parameterized factor.

Data: *The vector d_t containing data measured at time t is called* data record*.*

The vector $\phi_{c;t-1}$ is the observable state of the parameterized component*.*

The parameterized factor is determined by regression vector $\psi_{ic;t}$ *defined as a sub-selection of the vector* $[d_{i+1\_\mathring{d}c;t}, \phi'_{c;t-1}, 1]'$ *(3.6).*

The coupling $\Psi_{ic;t} \equiv [d_{ic;t}, \psi'_{ic;t}]'$ is called data vector *of the factor.*

Remarks 2

1. *We added the number 1 to the definition of the regression vector, because it helps us to effectively express the constant shifts in mean values of factors.*

2. *The adopted dynamic mixture model is not sufficiently general. The component weights should also depend on the state vector. The choice is driven by our inability to estimate this "natural" and more realistic model. See discussion in [8]*

3.3 Form of the prior and the posterior pdf

According to the general hints in section 3.1.2 we need to choose the prior pdf in a form that is well manipulable, i.e. analytically tractable.

Agreement 5 (Considered forms of pdfs on Θ^*) *The prior $\pi(\Theta) \equiv \pi(\Theta|d(0))$ and the posterior $\pi(\Theta|d(t)) \equiv \pi(\Theta|\mathcal{G}_t)$ are considered to be of the common form:*

$$\begin{aligned} \pi(\Theta|\mathcal{G}_t) &= Di_\alpha(\kappa_t) \prod_{i\in i^*, c\in c^*} \pi_{ic}(\Theta_{ic}|\mathcal{S}_{ic;t}),\ t \in \{0\} \cup t^* \quad, \textit{where} \qquad (3.7)\\ & \mathcal{G}_t \equiv (\kappa_{\bullet;t}, \mathcal{S}_{\bullet\bullet;t}),\\ & Di_\alpha(\kappa_t) \textit{ is Dirichlet distribution, } Di_\alpha(\kappa_\bullet) \equiv \frac{\prod_{c\in c^*} \alpha_c^{\kappa_c - 1}}{\mathcal{B}(\kappa)},\ \mathcal{B}(\kappa) \equiv \frac{\prod_{c\in c^*} \Gamma(\kappa_c)}{\Gamma(\sum_{c\in c^*} \kappa_c)},\\ & \textit{each pdf } \pi_{ic}(\Theta_{ic}|\mathcal{S}_{ic;t}) \textit{ is conjugate to the factor } f_{ic}(d_{ic;t}|\psi_{ic;t}, \Theta_{ic}). \end{aligned}$$

Parameters Θ_{ic}, $i \in i^* \equiv \{1, \ldots, \mathring{d}\}$, $c \in c^*$, of the individual parameterized factors are mutually conditionally independent, and also, independent of the component weights α. The component weights have Dirichlet distribution $Di_\alpha(\kappa)$ with support on the probabilistic simplex *α^*.*

Dirichlet distribution $Di_\alpha(\kappa_t)$ is recalled and analyzed in Chapter C.4 in detail.

3.4 Notations related to mixtures

In the sequel, we use the following elements: $i \in i^* \equiv \{1, \ldots, \mathring{d}\}$, $c \in c^*$

$$\begin{aligned} \text{Factor prediction} \quad \mathcal{I}_{ic;t} &= \int f_{ic}(d_{ic;t}|\psi_{ic;t}, \Theta_{ic}) \pi_{ic}(\Theta_{ic}|\mathcal{S}_{ic;t-1}) d\Theta_{ic}\\ \text{Component prediction} \quad \beta_{c;t} &= \prod_{i=1}^{\mathring{d}} \mathcal{I}_{ic;t} && (3.8)\\ \text{Estimate of component weight} \quad \hat{\alpha}_{c;t} &= \frac{\kappa_{c;t}}{\sum_{c\in c^*} \kappa_{c;t}} && (3.9)\\ \text{QB weight of data} \quad w_{c;t} &= \frac{\hat{\alpha}_{c;t-1}\beta_{c;t}}{\sum_{c=1}^{\mathring{c}} \hat{\alpha}_{c;t-1}\beta_{c;t}} && (3.10)\\ \text{"Correct" estimate of factor parameters} \quad \pi_{ic}(\Theta_{ic}|\mathcal{S}^U_{ic;t}) &= \frac{f_{ic}(d_{ic;t}|\psi_{ic;t}, \Theta_{ic}) \pi_{ic}(\Theta_{ic}|\mathcal{S}_{ic;t-1})}{\mathcal{I}_{ic;t}} && (3.11) \end{aligned}$$

Remarks 3

1. *The assumption of conjugacy of $\pi_{ic}(\Theta_{ic}|\mathcal{S}_{ic;t-1})$ to the factor $f_{ic}(d_{ic;t}|\psi_{ic;t}, \Theta_{ic})$ implies that $\frac{f_{ic}(d_{ic;t}|\psi_{ic;t}, \Theta_{ic})\pi_{ic}(\Theta_{ic}|\mathcal{S}_{ic;t-1})}{\mathcal{I}_{ic;t}}$ has the same functional form as $\pi_{ic}(\Theta_{ic}|\mathcal{S}_{ic;t-1})$, and thus we need to evaluate only the statistic $\mathcal{S}^U_{ic;t}$.*

2. *As values of* $\mathcal{I}_{ic;t}$ *can be very close to zero, it is numerically advantageous to evaluate the weights* $w_{\bullet;t}$ *using* $\mathcal{L}_{ic;t} = \ln \mathcal{I}_{ic;t}$. $\mathcal{L}_{ic;t}$ *can be computed directly without evaluating* $\mathcal{I}_{ic;t}$.

Algorithm 1 $w_{\bullet;t} = \text{EVAL\_WEIGHT}(\mathcal{L}_{\bullet\bullet;t}, \kappa_{\bullet;t-1})$

1. *For each component c evaluate* $\mathcal{H}_{c;t} = \ln \kappa_{c;t-1} + \sum_i \mathcal{L}_{ic;t}$
2. $\bar{\mathcal{H}}_{\bullet;t} = H_{\bullet;t} - \max H_{\bullet;t}$
3. $w_{\bullet;t} = \frac{\exp(\bar{\mathcal{H}}_{\bullet;t})}{\sum_c \exp(\bar{\mathcal{H}}_{\bullet;t})}$

Remarks 4 $w_{c;t}$ *evaluated in this algorithm is the same as defined in (3.10):*

$$\begin{aligned} w_{c;t} &= \frac{\exp\left(\mathcal{H}_{c;t} - \max \mathcal{H}_{\bullet;t}\right)}{\sum\left(\exp\left(\mathcal{H}_{c;t} - \max \mathcal{H}_{\bullet;t}\right)\right)} = \frac{\exp\left(\mathcal{H}_{c;t}\right) \exp\left(\max \mathcal{H}_{\bullet;t}\right)}{\exp\left(\max \mathcal{H}_{\bullet;t}\right) \sum \exp\left(\mathcal{H}_{c;t}\right)} = \\ &= \frac{\kappa_{c;t-1}\beta_{c;t}}{\sum \kappa_{c;t-1}\beta_{c;t}} = \frac{\frac{\kappa_{c;t-1}\beta_{c;t}}{\sum \kappa_{c;t-1}}}{\frac{\sum \kappa_{c;t-1}\beta_{c;t}}{\sum \kappa_{c;t-1}}} = \frac{\hat{\alpha}_{c;t-1}\beta_{c;t}}{\sum \hat{\alpha}_{c;t-1}\beta_{c;t}}. \end{aligned}$$

Chapter 4

Problem formulation and general solution

In this Section, we apply the approximation from section 3.1.2 to the introduced mixture model (3.3). We seek the statistic $\mathcal{G}_t$ that minimizes $\mathcal{D}\left(\hat{\pi}(\Theta|\mathcal{G}_{t-1}, \overbrace{d_t, \phi_{t-1}}^{\equiv\Psi_t}) \,\Big|\Big|\, \pi(\Theta|\mathcal{G}_t)\right)$, where

$$
\begin{aligned}
\hat{\pi}(\Theta|\mathcal{G}_{t-1}, \Psi_t) &= \frac{f(d_t|\phi_{t-1}, \Theta)\pi(\Theta|\mathcal{G}_{t-1})}{\int f(d_t|\phi_{t-1}, \Theta)\pi(\Theta|\mathcal{G}_{t-1})d\Theta} \\
\pi(\Theta|\mathcal{G}_{t-1}) &= Di_\alpha(\kappa_{t-1}) \prod_{i=1,c=1}^{\mathring{d},\mathring{c}} \pi_{ic}(\Theta_{ic}|\mathcal{S}_{ic;t-1}) \\
f(d_t|\phi_{t-1}, \Theta) &= \sum_{c=1}^{\mathring{c}} \alpha_c \prod_{i=1}^{\mathring{d}} f_{ic}(d_{ic;t}|\psi_{ic;t}, \Theta_{ic}).
\end{aligned}
$$

In this case, the statistic $\mathcal{G}_t$ consist of vector κ_t (of $\mathring{c}$ elements) and multimatrix $\mathcal{S}_{\bullet\bullet;t}$ of type $(\mathring{d}, \mathring{c})$.

The next proposition summarizes the form of $\hat{\pi}(\Theta|\mathcal{G}_{t-1}, \Psi_t)$.

Proposition 1

$$
\hat{\pi}(\Theta|\mathcal{G}_{t-1}, \Psi_t) = \sum_{c=1}^{\mathring{c}} w_{c;t} Di_\alpha(\kappa_{t-1} + \delta_{\bullet c}) \prod_{\substack{j,r=1 \\ r\neq c}}^{\mathring{d},\mathring{c}} \pi_{jr}(\Theta_{jr}|\mathcal{S}_{jr;t-1}) \prod_{j=1}^{\mathring{d}} \pi_{jc}(\Theta_{jc}|\mathcal{S}^U_{jc;t}). \tag{4.1}
$$

Proof:

$$
\begin{aligned}
f(d_t|\phi_{t-1}, \Theta)\pi(\Theta|\mathcal{G}_{t-1}) &= \sum_{c=1}^{\mathring{c}} \alpha_c \prod_{i=1}^{\mathring{d}} f_{ic}(d_{ic;t}|\psi_{ic;t}, \Theta_{ic}) Di_\alpha(\kappa_{t-1}) \prod_{j=1,r=1}^{\mathring{d},\mathring{c}} \pi_{jr}(\Theta_{jr}|\mathcal{S}_{jr;t-1}) = \\
&= \sum_{c=1}^{\mathring{c}} \alpha_c Di_\alpha(\kappa_{t-1}) \beta_{c;t} \prod_{\substack{j,r=1 \\ r\neq c}}^{\mathring{d},\mathring{c}} \pi_{jr}(\Theta_{jr}|\mathcal{S}_{jr;t-1}) \prod_{j=1}^{\mathring{d}} \pi_{jc}(\Theta_{jc}|\mathcal{S}^U_{jc;t}) \\
&= \sum_{c=1}^{\mathring{c}} \hat{\alpha}_{c;t-1}\beta_{c;t} \underbrace{Di_\alpha(\kappa_{t-1} + \delta_{\bullet c}) \prod_{\substack{j,r=1 \\ r\neq c}}^{\mathring{d},\mathring{c}} \pi_{jr}(\Theta_{jr}|\mathcal{S}_{jr;t-1}) \prod_{j=1}^{\mathring{d}} \pi_{jc}(\Theta_{jc}|\mathcal{S}^U_{jc;t})}_{\text{This part is pdf, hence it integrates to 1}}
\end{aligned}
$$

It's clear that the normalizing integral $\int f(d_t|\phi_{t-1}, \Theta)\pi(\Theta|\mathcal{G}_{t-1})d\Theta = \sum_{c=1}^{\mathring{c}} \hat{\alpha}_{c;t-1}\beta_{c;t}$, hence

$$
\hat{\pi}(\Theta|\mathcal{G}_{t-1}, \Psi_t) = \sum_{c=1}^{\mathring{c}} \underbrace{\frac{\hat{\alpha}_{c;t-1}\beta_{c;t}}{\sum_{\tilde{c}=1}^{\mathring{c}} \hat{\alpha}_{\tilde{c};t-1}\beta_{\tilde{c};t}}}_{=\,w_{c;t}} Di_\alpha(\kappa_{t-1} + \delta_{\bullet c}) \prod_{\substack{j,r=1 \\ r\neq c}}^{\mathring{d},\mathring{c}} \pi_{jr}(\Theta_{jr}|\mathcal{S}_{jr;t-1}) \prod_{j=1}^{\mathring{d}} \pi_{jc}(\Theta_{jc}|\mathcal{S}^U_{jc;t})
$$

□

Proposition 2 (Minimization of KL distance) *For $\mathcal{G}_t \equiv \{\mathcal{S}_{\bullet\bullet;t},\ \kappa_t\}$ minimizing*

$$\mathcal{D}\left(\hat{\pi}(\Theta|\mathcal{G}_{t-1},\Psi_t)\ \middle|\middle|\ \pi(\Theta|\mathcal{G}_t)\right),$$

it holds:

$$\begin{aligned}
\kappa_t &\in \operatorname*{Arg\,min}_{\kappa_t}\left[\sum_{c=1}^{\mathring{c}} w_{c;t}\mathcal{D}\left(Di_\alpha(\kappa_{t-1}+\delta_{\bullet,c})\ \middle|\middle|\ Di_\alpha(\kappa_t)\right)\right] \qquad (4.2)\\
\mathcal{S}_{ic;t} &\in \operatorname*{Arg\,min}_{\mathcal{S}_{ic;t}}\left[(1-w_{c;t})\mathcal{D}\left(\pi_{ic}(\Theta_{ic}|\mathcal{S}_{ic;t-1})\ \middle|\middle|\ \pi_{ic}(\Theta_{ic}|\mathcal{S}_{ic;t})\right)+w_{c;t}\mathcal{D}\left(\pi_{ic}(\Theta_{ic}|\mathcal{S}^U_{ic;t})\ \middle|\middle|\ \pi_{ic}(\Theta_{ic}|\mathcal{S}_{ic;t})\right)\right].
\end{aligned}$$

Proof:

Instead of working with KL distance, we will evaluate the so called Carridge distance $\mathcal{K}\left(\hat{\pi}(\Theta|\mathcal{G}_{t-1},\Psi_t)\ \middle|\middle|\ \pi(\Theta|\mathcal{G}_t)\right)$. Details about this "distance", it's properties and it's relation to KL distance are discussed in section A.2.

$$\begin{aligned}
&\mathcal{K}\left(\sum_{c=1}^{\mathring{c}} w_{c;t}Di_\alpha(\kappa_{t-1}+\delta_{\bullet c})\prod_{\substack{j,r=1\\ r\neq c}}^{\mathring{d},\mathring{c}}\pi_{jr}(\Theta_{jr}|\mathcal{S}_{jr;t-1})\prod_{j=1}^{\mathring{d}}\pi_{jc}(\Theta_{jc}|\mathcal{S}^U_{jc;t})\ \middle|\middle|\ Di_\alpha(\kappa_t)\prod_{i=1,c=1}^{\mathring{d},\mathring{c}}\pi_{ic}(\Theta_{ic}|\mathcal{S}_{ic;t})\right)\overbrace{=}^{prop.\ 8}\\
&=\sum_{c=1}^{\mathring{c}} w_{c;t}\mathcal{K}\left(Di_\alpha(\kappa_{t-1}+\delta_{\bullet c})\prod_{\substack{j,r=1\\ r\neq c}}^{\mathring{d},\mathring{c}}\pi_{jr}(\Theta_{jr}|\mathcal{S}_{jr;t-1})\prod_{j=1}^{\mathring{d}}\pi_{jc}(\Theta_{jc}|\mathcal{S}^U_{jc;t})\ \middle|\middle|\ Di_\alpha(\kappa_t)\prod_{i=1,c=1}^{\mathring{d},\mathring{c}}\pi_{ic}(\Theta_{ic}|\mathcal{S}_{ic;t})\right)\overbrace{=}^{prop.\ 9}\\
&=\sum_{c=1}^{\mathring{c}} w_{c;t}\left[\mathcal{K}\left(Di_\alpha(\kappa_{t-1}+\delta_{\bullet c})\ \middle|\middle|\ Di_\alpha(\kappa_t)\right)+\sum_{\substack{j,r=1\\ j\neq r}}^{\mathring{d},\mathring{c}}\mathcal{K}\left(\pi_{jr}(\Theta_{jr}|\mathcal{S}_{jr;t-1})\ \middle|\middle|\ \pi_{jr}(\Theta_{jr}|\mathcal{S}_{jr;t})\right)+\right.\\
&\qquad\left.+\sum_{j=1}^{\mathring{d}}\mathcal{K}\left(\pi_{jc}(\Theta_{jc}|\mathcal{S}^U_{jc;t-1})\ \middle|\middle|\ \pi_{jc}(\Theta_{jc}|\mathcal{S}_{jc;t})\right)\right]
\end{aligned}$$

Let's now temporarily denote

$$\begin{aligned}
\mathcal{K}_{jc} &= \mathcal{K}\left(\pi_{jc}(\Theta_{jc}|\mathcal{S}_{jc;t-1})\ \middle|\middle|\ \pi_{jc}(\Theta_{jc}|\mathcal{S}_{jc;t})\right)\\
\mathcal{K}^U_{jc} &= \mathcal{K}\left(\pi_{jc}(\Theta_{jc}|\mathcal{S}^U_{jc;t-1})\ \middle|\middle|\ \pi_{jc}(\Theta_{jc}|\mathcal{S}_{jc;t})\right)
\end{aligned}$$

The problem becomes the form

$$\begin{aligned}
&\sum_{c=1}^{\mathring{c}} w_{c;t}\mathcal{K}\left(Di_\alpha(\kappa_{t-1}+\delta_{\bullet c})\ \middle|\middle|\ Di_\alpha(\kappa_t)\right)+\sum_{c=1}^{\mathring{c}} w_{c;t}\sum_{\substack{j,r=1\\ j\neq r}}^{\mathring{d},\mathring{c}}\mathcal{K}_{jr}+\sum_{j,c=1}^{\mathring{d},\mathring{c}} w_{c;t}\mathcal{K}^U_{jc}\overbrace{=}^{prop.\ 12}\\
&=\sum_{c=1}^{\mathring{c}} w_{c;t}\mathcal{K}\left(Di_\alpha(\kappa_{t-1}+\delta_{\bullet c})\ \middle|\middle|\ Di_\alpha(\kappa_t)\right)+\sum_{j,c=1}^{\mathring{d},\mathring{c}}\left[w_{c;t}\mathcal{K}^U_{jc}+(1-w_{c;t})\mathcal{K}_{jc}\right]
\end{aligned}$$

Now it is clear that minimization of this expression can be done in parts.

$$\begin{aligned}
\kappa_t &\in \operatorname*{Arg\,min}_{\kappa_t}\left[\sum_{c=1}^{\mathring{c}} w_{c;t}\mathcal{K}\left(Di_\alpha(\kappa_{t-1}+\delta_{\bullet,c})\ \middle|\middle|\ Di_\alpha(\kappa_t)\right)\right]\\
\mathcal{S}_{ic;t} &\in \operatorname*{Arg\,min}_{\mathcal{S}_{lc;t}}\left[(1-w_{c;t})\mathcal{K}_{cj}+w_{c;t}\mathcal{K}^U_{cj}\right]=\\
&= \operatorname*{Arg\,min}_{\mathcal{S}_{lc;t}}\left[(1-w_{c;t})\mathcal{K}\left(\pi_{lc}(\Theta_{lc}|\mathcal{S}_{lc;t-1})\ \middle|\middle|\ \pi_{lc}(\Theta_{lc}|\mathcal{S}_{lc;t})\right)+w_{c;t}\mathcal{K}\left(\pi_{lc}(\Theta_{lc}|\mathcal{S}^U_{lc;t})\ \middle|\middle|\ \pi_{lc}(\Theta_{lc}|\mathcal{S}_{lc;t})\right)\right]
\end{aligned}$$

The proposition 7 says that the argument minimizing Carridge distance also minimizes KL distance.

□

Remarks 5 *The previous proposition split the overall problem into two subproblems. The subproblem (4.2) can be solved in general, as presented in section 4.2. Solution of the second subproblem depends on the choice of the system model. Solution for the Normal models is presented in chapter 5.*

4.1 General algorithm

Following the proposition 2 we sketch the general algorithm of one mixture estimation step. We naturally suppose that $\Psi_{ic;t}$ can be obtained from $d(t)$.

Algorithm 2

Inputs	-	$\kappa_{\bullet;t-1},\ \mathcal{S}_{\bullet\bullet;t-1},\ \Psi_{\bullet\bullet;t}$
Outputs	-	$\kappa_{\bullet;t},\ \mathcal{S}_{\bullet\bullet;t}$

1. *For each factor ic evaluate* $\mathcal{L}_{ic;t} = \ln \mathcal{I}_{ic;t}$

2. $w_{\bullet;t} = \mathrm{EVAL\_WEIGHT}(\mathcal{L}_{\bullet\bullet;t}, \kappa_{\bullet;t-1})$ *(Algorithm 1)*

3. $\kappa_t \in \operatorname{Arg\,min}_{\kappa_t>0} \left[\sum_{c=1}^{\mathring{c}} w_{c;t}\mathcal{D}\left(Di_\alpha(\kappa_{t-1}+\delta_{\bullet c}) \,\middle|\middle|\, Di_\alpha(\kappa_t)\right)\right]$

4. *For each factor ic evaluate* $\mathcal{S}^U_{ic;t}$ *so that* $\pi_{ic}(\Theta_{ic}|\mathcal{S}^U_{ic;t}) = \frac{\pi_{ic}(\Theta_{ic}|\mathcal{S}_{ic;t-1}) f_{ic}(d_{ic;t}|\psi_{ic;t},\Theta_{ic})}{\mathcal{I}_{ic}}$

5. *For each factor ic evaluate*
$\mathcal{S}_{ic;t} \in \operatorname{Arg\,min}_{\mathcal{S}_{ic;t}} \left[(1-w_{c;t})\mathcal{D}\left(\pi_{ic}(\Theta_{ic}|\mathcal{S}_{ic;t-1}) \,\middle|\middle|\, \pi_{ic}(\Theta_{ic}|\mathcal{S}_{ic;t})\right) + w_{c;t}\mathcal{D}\left(\pi_{ic}(\Theta_{ic}|\mathcal{S}^U_{ic;t}) \,\middle|\middle|\, \pi_{ic}(\Theta_{ic}|\mathcal{S}_{ic;t})\right)\right]$

Steps 1,4,5 depends on the specific choice of the system model, step 2 is solved, and step 3 is discussed in the next section.

4.2 Minimization with respect to κ_t

The following proposition converts the problem of KL distance minimization of κ-part to minimization of an algebraic expression.

Proposition 3 (Minimization with respect to κ_t)
For κ_t minimizing

$$\sum_{c=1}^{\mathring{c}} w_{c;t}\mathcal{D}\left(Di_\alpha(\kappa_{t-1}+\delta_{\bullet c}) \,\middle|\middle|\, Di_\alpha(\kappa_t)\right)$$

it holds

$$\kappa_{\bullet;t} \in \operatorname{Arg\,min}\left\{\sum_{c=1}^{\mathring{c}}\left[\ln\left(\Gamma\left(\kappa_{c;t}\right)\right) - \kappa_{c;t}\xi_{c;t}\right] - \ln\left(\Gamma\left(\sum_{c=1}^{\mathring{c}}\kappa_{c;t}\right)\right)\right\}$$

where

$$\xi_{c;t} = \left(\psi_0\left(\kappa_{c;t-1}\right) + \frac{w_{c,t}}{\kappa_{c;t-1}} - \psi_0\left(\sum_{c=1}^{\mathring{c}}\kappa_{c;t-1}+1\right)\right).$$

Proof: According to proposition 29, which evaluates KL distance of two dirichlet pdfs, it is clear, that we can minimize

$$\sum_{c=1}^{\mathring{c}} w_{c;t}\mathcal{Z}(\kappa_{t-1},\kappa_t,c) \equiv \sum_{c=1}^{\mathring{c}} w_{c;t}\mathcal{Z}_{c;t}\ , \text{ where}$$

$$\mathcal{Z}(\kappa_{t-1},\kappa_t,c) = \sum_{j=1}^{\mathring{c}}\left[\ln\left(\Gamma\left(\kappa_{j;t}\right)\right) - \kappa_{j;t}\psi_0\left(\kappa_{j;t-1}+\delta_{cj}\right)\right] - \left[\ln(\Gamma(\sum_{c=1}^{\mathring{c}}\kappa_{c;t}) - \sum_{j=1}^{\mathring{c}}\kappa_{j;t}\psi_0\left(\sum_{c=1}^{\mathring{c}}\kappa_{c;t-1}+1\right)\right].$$

$$\begin{aligned}\sum_{c=1}^{\mathring{c}} w_{c;t}\mathcal{Z}_{c;t} &= \sum_{j=1}^{\mathring{c}}\sum_{c=1}^{\mathring{c}} w_{c;t}\left[\ln\left(\Gamma\left(\kappa_{j;t}\right)\right) - \kappa_{j;t}\left(\psi_0\left(\kappa_{j;t-1}\right) + \frac{\delta_{c,j}}{\kappa_{j;t-1}}\right)\right] - \\ &\quad - \left[\ln\left(\Gamma\left(\sum_{c=1}^{\mathring{c}}\kappa_{c;t}\right)\right) - \sum_{j}^{\mathring{c}}\kappa_{j;t}\psi_0\left(\sum_{c=1}^{\mathring{c}}\kappa_{c;t-1}+1\right)\right] =\end{aligned}$$

$$
\begin{aligned}
&= \sum_{j=1}^{\mathring{c}} \left[\ln\left(\Gamma\left(\kappa_{j;t}\right)\right) - \kappa_{j;t}\psi_0\left(\kappa_{j;t-1}\right)\right] - \sum_{j}^{\mathring{c}} \kappa_{j;t} \sum_{c}^{\mathring{c}} \delta_{c,j} \frac{w_{c,t}}{\kappa_{j;t-1}} - \\
&- \left[\ln\left(\Gamma\left(\sum_{c=1}^{\mathring{c}} \kappa_{c;t}\right)\right) - \sum_{j}^{\mathring{c}} \kappa_{j;t}\psi_0\left(\sum_{c=1}^{\mathring{c}} \kappa_{c;t-1} + 1\right)\right] = \\
&= \sum_{j=1}^{\mathring{c}} \left[\ln\left(\Gamma\left(\kappa_{j;t}\right)\right) - \kappa_{j;t} \underbrace{\left(\psi_0\left(\kappa_{j;t-1}\right) + \frac{w_{j,t}}{\kappa_{j;t-1}} - \psi_0\left(\sum_{c=1}^{\mathring{c}} \kappa_{c;t-1} + 1\right)\right)}_{\xi_{j;t}}\right] - \ln\left(\Gamma\left(\sum_{c=1}^{\mathring{c}} \kappa_{c;t}\right)\right)
\end{aligned}
$$

□

Proposition 3 yields the following algorithm.

Algorithm 3 $\kappa_{\bullet;t} = \mathrm{NEW\_KAPPA}(w_{\bullet;t}, \kappa_{\bullet;t-1})$

1. *For each component c evaluate* $\xi_{c;t} = \psi_0\left(\kappa_{c;t-1}\right) + \frac{w_{c,t}}{\kappa_{c;t-1}} - \psi_0\left(\sum_{c=1}^{\mathring{c}} \kappa_{c;t-1} + 1\right)$

2. $\kappa_{\bullet;t} \in \operatorname{Arg\,min}\left\{\sum_{j=1}^{\mathring{c}} \left[\ln\left(\Gamma\left(\kappa_{j;t}\right)\right) - \kappa_{j;t}\xi_{j;t}\right] - \ln\left(\Gamma\left(\sum_{c}^{\mathring{c}} \kappa_{c;t}\right)\right)\right\}$

Remarks 6

1. *Minimization of the term (4.2) can be simply approximated by changing* $\mathcal{D}\left(Di_\alpha(\kappa_1) \,\middle|\middle|\, Di_\alpha(\kappa_2)\right)$ *into square of the Euclidean norm* $||\kappa_1 - \kappa_2||^2$. *The problem is then transformed into minimization of* $\min\limits_{x} \sum\limits_{c} w_c ||x - x_c||^2$ *which has explicit solution:* $x = \sum\limits_{c} w_c x_c$. *Applied to our case it yields* $\kappa_t = \kappa_{t-1} + w_t$, *which is identical to the solution obtained using the quasi-Bayes algorithm [4].*

2. *The minimization problem in step 2 must be solved numerically or by suitable approximation. For detailed solution of this problem see [9].*

We have completed all steps which can be done on this general level. In the next parts of the paper, we are dealing with the special case of the factors.

Chapter 5

Application to normal factors

In this chapter, we assume the parameterized factor to be dynamic Gaussian pdf with parameters $\Theta_{ic} \equiv (\theta_{ic}, r_{ic})$, where θ_{ic} is so called vector of regression coefficients and r_{ic} is noise variance of the factor.

$$f_{ic}(d_{ic;t}|\psi_{ic;t}, \Theta_{ic}) = \mathcal{N}_{d_{ic;t}}(\theta'_{ic}\psi_{ic;t}, r_{ic}) = \frac{1}{\sqrt{2\pi r_{ic}}} \exp\left(-\frac{(d_{ic;t} - \theta'_{ic}\psi_{ic;t})^2}{2r_{ic}}\right)$$

We don't need to introduce a shift in the mean value, because the regression vector can contain number 1. See Remarks 2. The shifting constant is then placed to the corresponding place of the vector of regression coefficients.

The prior conjugate to this model is the Gauss inverse Wishart pdf with parameters $\mathcal{S}_{ic;t} = (\nu_{ic;t}, V_{ic;t})$, where $\nu_{ic;t}$ is scalar count of degrees of freedom and $V_{ic;t}$ is so called extended information matrix (symmetric, positive definite, of type $(\mathring{\Psi}_{ic;t}, \mathring{\Psi}_{ic;t})$).

$$\pi_{ic}(\Theta_{ic}|\mathcal{S}_{ic;t}) = GiW_{\theta_{ic},r_{ic}}(V_{ic;t}, \nu_{ic;t}) \propto r_{ic}^{-0.5(\nu_{ic;t}+\mathring{\psi}_{ic;t}+2)} \exp\left\{-\frac{1}{2r_{ic}} \mathrm{tr}\left(V_{ic;t}[-1, \theta'_{ic}]'[-1, \theta'_{ic}]\right)\right\}$$

The details and important properties of this pdf are outlined in the appendix C. Note that the matrix V_{ic} can be equivalently manipulated through its $L'DL$ decomposition (i.e. with lower triangular matrix L_{ic} and diagonal matrix D_{ic} which fulfills the relation $V_{ic} = L'_{ic}D_{ic}L_{ic}$). Next, the matrices L_{ic} and D_{ic} can be equivalently expressed via matrix C_{ic}, vector $\hat{\theta}_{ic}$ and scalar ${}^{\lfloor d}D_{ic}$. The relations between individual representations can be found in section C.2.2.

Because all three representations described above are equivalent, we will not formally distinguish between them. If V_{ic} is a statistic of GiW factor, under the terms $L_{ic}, D_{ic}, \hat{\theta}_{ic}, C_{ic}, {}^{\lfloor d}D_{ic}$ we automatically mean the parts of corresponding representation of the matrix V_{ic}.

Now, we specify the steps 1,4,5 in the general algorithm 2 for Normal factors.

5.1 Evaluating $\mathcal{I}_{ic;t}$

$\mathcal{I}_{ic}$ is defined as

$$\mathcal{I}_{ic;t} = \int f_{ic}(d_{ic;t}|\psi_{ic;t}, \Theta_{ic})\pi_{ic}(\Theta_{ic}|\mathcal{S}_{ic;t-1})d\Theta_{ic} = \int \mathcal{N}_{d_{ic;t}}(\theta'_{ic}\psi_{ic;t}, r_{ic})GiW_{\theta_{ic},r_{ic}}(V_{ic;t-1}, \nu_{ic;t-1})d\theta_{ic}dr_{ic}.$$

According to the proposition 24, $\mathcal{I}_{ic;t}$ is for normal factors evaluated as:

$$\mathcal{I}_{ic;t} = \frac{\Gamma(0.5(\nu_{ic;t-1}+1))\left[{}^{\lfloor d}D_{ic;t-1}(1+\zeta_{ic;t})\right]^{-0.5}}{\sqrt{\pi}\Gamma(0.5\nu_{ic;t-1})\left(1+\frac{\hat{e}^2_{ic;t}}{{}^{\lfloor d}D_{ic;t-1}(1+\zeta_{ic;t})}\right)^{0.5(\nu_{ic;t-1}+1)}} \tag{5.1}$$

where

$$\begin{aligned} \hat{e}_{ic;t} &\equiv d_{ic;t} - \hat{\theta}'_{ic;t-1}\psi_{ic;t} \equiv \textit{prediction error} \\ \zeta_{ic;t} &\equiv \psi'_{ic;t}C_{ic;t-1}\psi_{ic;t} \end{aligned}$$

Remarks 7 *According to remarks 3, we need to evaluate $\mathcal{L}_{ic;t} = \ln \mathcal{I}_{ic;t}$. It can be done efficiently via the product form of (5.1). The following algorithm summarizes this task. Recall that $\Psi_{ic;t} = [d_{ic;t}, \psi_{ic;t}]$.*

Algorithm 4 (evaluation of $\mathcal{L}_{ic;t}$) $\mathcal{L}_{ic;t} = \text{FACNORM}(C_{ic;t-1}, \hat{\theta}_{ic;t-1}, {}^{\lfloor d}D_{ic;t-1}, \nu_{ic;t-1}, \Psi_{ic;t})$

1. *Evaluate* $\zeta_{ic;t} = \psi'_{ic;t} C_{ic;t-1} \psi_{ic;t}$

2. *Evaluate* $\hat{e}_{ic;t} \equiv d_{ic;t} - \hat{\theta}'_{ic;t-1}\psi_{ic;t}$

3. *Evaluate*

$$\begin{aligned}\mathcal{L}_{ic;t} = \ln \mathcal{I}_{ic;t} &= \ln\Gamma\left(0.5(\nu_{ic;t-1}+1)\right) - \ln\Gamma\left(0.5\nu_{ic;t-1}\right) - 0.5\ln\left({}^{\lfloor d}D_{ic;t-1}\right) - 0.5\ln\left(1+\zeta_{ic;t}\right) - \\ &\quad -0.5(\nu_{ic;t-1}+1)\ln\left(1+\frac{\hat{e}^2_{ic;t}}{{}^{\lfloor d}D_{ic;t-1}(1+\zeta_{ic;t})}\right) - 0.5\ln\left(\pi\right)\end{aligned}$$

Remarks 8 *Function* $\ln\Gamma$ *can be evaluated without computing* Γ *first [10].*

5.2 Evaluating $\mathcal{S}^U_{ic;t}$

According to the proposition 23, $\mathcal{S}^U_{ic;t} \equiv [V^U_{ic}, \nu^U_{ic}]$ can be evaluated in the following way:

$$\begin{aligned} V^U_{ic;t} &= V_{ic;t-1} + \Psi_{ic;t}\Psi'_{ic;t} \\ \nu^U_{ic;t} &= \nu_{ic;t-1} + 1 \end{aligned} \tag{5.2}$$

Using the proposition 28, the relation (5.2) can be in the following way rewritten into the $C, \hat{\theta}, {}^{\lfloor d}D$ representation:

$$\begin{aligned} C^U_{ic;t} &= C_{ic;t-1} - \frac{1}{1+\zeta_{ic;t}} z_{ic;t} z'_{ic;t} \qquad , \ \hat{\theta}^U_{ic;t} = \hat{\theta}_{ic;t-1} + \frac{\hat{e}_{ic;t}}{1+\zeta_{ic;t}} z_{ic;t} \\ {}^{\lfloor d}D^U_{ic;t} &= {}^{\lfloor d}D_{ic;t-1} + \frac{\hat{e}^2_{ic;t}}{1+\zeta_{ic;t}} \qquad\qquad z_{ic;t} = C_{ic;t-1}\psi_{ic;t} \end{aligned}$$

5.3 Minimizing the KL distance

According to the proposition 2, we need to minimize

$$(1-w_{c;t})\mathcal{D}\left(\pi_{ic}(\Theta_{ic}|\mathcal{S}_{ic;t-1}) \,\Big|\Big|\, \pi_{ic}(\Theta_{ic}|\mathcal{S}_{ic;t})\right) + w_{c;t}\mathcal{D}\left(\pi_{ic}(\Theta_{ic}|\mathcal{S}^U_{ic;t}) \,\Big|\Big|\, \pi_{ic}(\Theta_{ic}|\mathcal{S}_{ic;t})\right)$$

for each factor i within the component c. The minimization can be done factor-vise (see alg. 2), thus we can simplify the notation by considering one particular factor.

$$\begin{array}{lcll} \mathcal{S}_{ic;t-1} & \equiv & (V_{ic;t-1}, \nu_{ic;t-1}) & \rightarrow (V, \nu) \\ \mathcal{S}_{ic;t} & \equiv & (V_{ic;t}, \nu_{ic;t}) & \rightarrow (V^\spadesuit, \nu^\spadesuit) \\ \mathcal{S}^U_{ic;t} & \equiv & (V^U_{ic;t-1}, \nu^U_{ic;t-1}) & \rightarrow (V^U, \nu^U) \\ w_{c;t} & & & \rightarrow w \\ \psi_{ic;t}, \Psi_{ic;t}, d_t & & & \rightarrow \psi, \Psi, d \end{array}$$

Thus, we minimize

$$\min_{V^\spadesuit, \nu^\spadesuit} \left\{(1-w)\mathcal{D}\left(GiW_{\theta,r}(V,\nu) \,\Big|\Big|\, GiW_{\theta,r}(V^\spadesuit,\nu^\spadesuit)\right) + w\mathcal{D}\left(GiW_{\theta,r}(V^U,\nu^U) \,\Big|\Big|\, GiW_{\theta,r}(V^\spadesuit,\nu^\spadesuit)\right)\right\}. \tag{5.3}$$

Proposition 4 *For* $V^\spadesuit \equiv (C^\spadesuit, \hat{\theta}^\spadesuit, {}^{\lfloor d}D^\spadesuit), \nu^\spadesuit$ *minimizing (5.3) it holds:*

$$(\nu^\spadesuit, {}^{\lfloor d}D^\spadesuit) = \arg\min_{\nu^\spadesuit, {}^{\lfloor d}D^\spadesuit} \left\{ \underbrace{(1-w)M\left(\nu, {}^{\lfloor d}D, \nu^\spadesuit, {}^{\lfloor d}D^\spadesuit\right) + wM\left(\nu^U, {}^{\lfloor d}D^U, \nu^\spadesuit, {}^{\lfloor d}D^\spadesuit\right)}_{MO\left(w,\nu,{}^{\lfloor d}D,\nu^U,{}^{\lfloor d}D^U,\nu^\spadesuit,{}^{\lfloor d}D^\spadesuit\right)} \right\} \tag{5.4}$$

$$(\hat{\theta}^\spadesuit, C^\spadesuit) = \arg\min_{\hat{\theta}^\spadesuit, C^\spadesuit} \left\{ \underbrace{(1-w)G\left(\hat{\theta}, C, \nu, {}^{\lfloor d}D, \hat{\theta}^\spadesuit, C^\spadesuit\right) + wG\left(\hat{\theta}^U, C^U, \nu^U, {}^{\lfloor d}D^U, \hat{\theta}^\spadesuit, C^\spadesuit\right)}_{GO\left(\hat{\theta},C,\nu,{}^{\lfloor d}D,\hat{\theta}^U,C^U,\nu^U,{}^{\lfloor d}D^U,\hat{\theta}^\spadesuit,C^\spadesuit\right)} \right\} \tag{5.5}$$

where

$$
\begin{aligned}
M\left(\nu,\ \lfloor^{d}D,\nu^{\spadesuit},\ \lfloor^{d}D^{\spadesuit}\right) &= \ln\left(\Gamma\left(0.5\nu^{\spadesuit}\right)\right)+0.5\nu^{\spadesuit}\ln\left(\frac{\lfloor^{d}D}{\lfloor^{d}D^{\spadesuit}}\right)-0.5\nu^{\spadesuit}\psi_0\left(0.5\nu\right)+0.5\frac{\nu\,\lfloor^{d}D^{\spadesuit}}{\lfloor^{d}D} \\
G\left(\hat\theta,C,\nu,\ \lfloor^{d}D,\hat\theta^{\spadesuit},C^{\spadesuit}\right) &= -0.5\ln\left|CC^{\spadesuit^{-1}}\right|+0.5\mathrm{tr}\left[CC^{\spadesuit^{-1}}\right]+0.5\frac{\nu}{\lfloor^{d}D}\left(\hat\theta-\hat\theta^{\spadesuit}\right)'C^{\spadesuit^{-1}}\left(\hat\theta-\hat\theta^{\spadesuit}\right)
\end{aligned}
$$

Proof: We use the proposition 22, which evaluates the KL distance of two GiW pdfs. The term $const$ means all elements not depending on optimized variables (with superscript $\spadesuit$).

$$
\begin{aligned}
\mathcal{D}\left(GiW_{\theta,r}(V,\nu)\ \middle|\middle|\ GiW_{\theta,r}(V^{\spadesuit},\nu^{\spadesuit})\right) &= const+\ln\left(\Gamma\left(0.5\nu^{\spadesuit}\right)\right)-0.5\ln\left|CC^{\spadesuit^{-1}}\right|+0.5\nu^{\spadesuit}\ln\left(\frac{\lfloor^{d}D}{\lfloor^{d}D^{\spadesuit}}\right) \\
&\quad -0.5\nu^{\spadesuit}\psi_0\left(0.5\nu\right)+0.5\mathrm{tr}\left[CC^{\spadesuit^{-1}}\right]+ \\
&\quad +0.5\frac{\nu}{\lfloor^{d}D}\left[\left(\hat\theta-\hat\theta^{\spadesuit}\right)'C^{\spadesuit^{-1}}\left(\hat\theta-\hat\theta^{\spadesuit}\right)+\ \lfloor^{d}D^{\spadesuit}\right].
\end{aligned}
$$

Now we give together the elements containing $\nu^{\spadesuit}$ a $\lfloor^{d}D^{\spadesuit}$.

$$
\begin{aligned}
\mathcal{D} &= const+\left\{\ln\left(\Gamma\left(0.5\nu^{\spadesuit}\right)\right)+0.5\nu^{\spadesuit}\ln\left(\frac{\lfloor^{d}D}{\lfloor^{d}D^{\spadesuit}}\right)-0.5\nu^{\spadesuit}\psi_0\left(0.5\nu\right)+0.5\frac{\nu\,\lfloor^{d}D^{\spadesuit}}{\lfloor^{d}D}\right\}+ \\
&+\ \left\{-0.5\ln\left|CC^{\spadesuit^{-1}}\right|+0.5\mathrm{tr}\left[CC^{\spadesuit^{-1}}\right]+0.5\frac{\nu}{\lfloor^{d}D}\left(\hat\theta-\hat\theta^{\spadesuit}\right)'C^{\spadesuit^{-1}}\left(\hat\theta-\hat\theta^{\spadesuit}\right)\right\}= \\
&= const+M(\nu,\ \lfloor^{d}D,\nu^{\spadesuit},\ \lfloor^{d}D^{\spadesuit})+G(\hat\theta,C,\nu,\ \lfloor^{d}D,\hat\theta^{\spadesuit},C^{\spadesuit})
\end{aligned}
$$

□

Remarks 9

1. *The proposition 4 changes the problem of minimizing weighted sum of KL distance to two independent algebraic subproblems. First of them is minimization on two dimensional space* ($\nu^{\spadesuit}$, $\lfloor^{d}D^{\spadesuit}$), *the second is more complex. Both subproblems are solved in next sections.*
2. *If we approximate* $\mathcal{D}\left(GiW_{\theta,r}(V,\nu)\ \middle|\middle|\ GiW_{\theta,r}(V^{\spadesuit},\nu^{\spadesuit})\right)$ *with* $||V-V^{\spadesuit}||^2+||\nu-\nu^{\spadesuit}||^2$, *we can quickly achieve the result* $V^{\spadesuit}=V+w\Psi\Psi'$, $\nu^{\spadesuit}=\nu+w$, *which is exactly the same as the quasi-Bayes update [4].*

5.3.1 Searching for $\lfloor^{d}D^{\spadesuit}$ and $\nu^{\spadesuit}$

Proposition 5 *For* $\nu^{\spadesuit}$, $\lfloor^{d}D^{\spadesuit}$ *minimizing (5.3) it holds:*

$$
\begin{aligned}
\frac{\nu^{\spadesuit}}{\lfloor^{d}D^{\spadesuit}} &= (1-w)\frac{\nu}{\lfloor^{d}D}+w\frac{\nu^{U}}{\lfloor^{d}D^{U}} \\
\psi_0\left(0.5\nu^{\spadesuit}\right)-\ln\left(0.5\nu^{\spadesuit}\right) &= \Upsilon,\ \textit{where} \qquad (5.6)\\
\Upsilon &\equiv (1-w)\left(\psi_0\left(0.5\nu\right)-\ln\left(\lfloor^{d}D\right)\right)+w\left(\psi_0\left(0.5\nu^{U}\right)-\ln\left(\lfloor^{d}D^{U}\right)\right)- \\
&\quad -\ln\left(0.5(1-w)\frac{\nu}{\lfloor^{d}D}+0.5w\frac{\nu^{U}}{\lfloor^{d}D^{U}}\right) \qquad (5.7)
\end{aligned}
$$

Proof: We want to find minimum the of the function $MO\left(w,\nu,\ \lfloor^{d}D,\nu^{U},\ \lfloor^{d}D^{U},\nu^{\spadesuit},\ \lfloor^{d}D^{\spadesuit}\right)$ from (5.4). We will use the differential approach. The general proposition on minimizing 2-variable functions [10] says:

If the function $f(x,y)$ *fulfills*

$$
\begin{aligned}
\frac{\partial f}{\partial x}(x_0,y_0) &= 0 \qquad (5.8)\\
\frac{\partial f}{\partial y}(x_0,y_0) &= 0 \qquad (5.9)
\end{aligned}
$$

$$\frac{\partial^2 f}{\partial x^2}(x_0,y_0) > 0 \tag{5.10}$$
$$\frac{\partial^2 f}{\partial y^2}(x_0,y_0) > 0 \tag{5.11}$$
$$\begin{vmatrix} \frac{\partial^2 f}{\partial x\partial y}(x_0,y_0) & \frac{\partial^2 f}{\partial x^2}(x_0,y_0) \\ \frac{\partial^2 f}{\partial y^2}(x_0,y_0) & \frac{\partial^2 f}{\partial x\partial y}(x_0,y_0) \end{vmatrix} < 0 \tag{5.12}$$

then it has the local minimum in the point (x_0, y_0).

First, we will solve the equations (5.8),(5.9).

$$0 = (1-w)\frac{\partial M(\nu, {}^{\lfloor d}D, \nu^{\spadesuit}, {}^{\lfloor d}D^{\spadesuit})}{\partial \nu^{\spadesuit}} + w\frac{\partial M(\nu^U, {}^{\lfloor d}D^U, \nu^{\spadesuit}, {}^{\lfloor d}D^{\spadesuit})}{\partial \nu^{\spadesuit}} \tag{5.13}$$
$$0 = (1-w)\frac{\partial M(\nu, {}^{\lfloor d}D, \nu^{\spadesuit}, {}^{\lfloor d}D^{\spadesuit})}{\partial\, {}^{\lfloor d}D^{\spadesuit}} + w\frac{\partial M(\nu^U, {}^{\lfloor d}D^U, \nu^{\spadesuit}, {}^{\lfloor d}D^{\spadesuit})}{\partial\, {}^{\lfloor d}D^{\spadesuit}} \tag{5.14}$$

Using the relation (B.2), the equation (5.14) becomes the form

$$0 = (1-w)0.5\left(\frac{\nu}{{}^{\lfloor d}D} - \frac{\nu^{\spadesuit}}{{}^{\lfloor d}D^{\spadesuit}}\right) + w0.5\left(\frac{\nu^U}{{}^{\lfloor d}D^U} - \frac{\nu^{\spadesuit}}{{}^{\lfloor d}D^{\spadesuit}}\right) \tag{5.15}$$
$$\frac{\nu^{\spadesuit}}{{}^{\lfloor d}D^{\spadesuit}} = (1-w)\frac{\nu}{{}^{\lfloor d}D} + w\frac{\nu^U}{{}^{\lfloor d}D^U} \overbrace{=}^{denote} X^S \tag{5.16}$$

Using the relation (B.3) the equation (5.13) yields

$$0 = (1-w)0.5\left(\psi_0\left(0.5\nu^{\spadesuit}\right) - \psi_0\left(0.5\nu\right) + \ln\left(\frac{{}^{\lfloor d}D}{{}^{\lfloor d}D^{\spadesuit}}\right)\right) + w0.5\left(\psi_0\left(0.5\nu^{\spadesuit}\right) - \psi_0\left(0.5\nu^U\right) + \ln\left(\frac{{}^{\lfloor d}D^U}{{}^{\lfloor d}D^{\spadesuit}}\right)\right)$$
$$\psi_0\left(0.5\nu^{\spadesuit}\right) - \ln\left({}^{\lfloor d}D^{\spadesuit}\right) = (1-w)\left(\psi_0\left(0.5\nu\right) - \ln\left({}^{\lfloor d}D\right)\right) + w\left(\psi_0\left(0.5\nu^U\right) - \ln\left({}^{\lfloor d}D^U\right)\right)$$

using (5.16) we get:

$$\begin{aligned} {}^{\lfloor d}D^{\spadesuit} &= \frac{0.5\nu^{\spadesuit}}{0.5X^S} \\ \ln\left({}^{\lfloor d}D^{\spadesuit}\right) &= \ln\left(0.5\nu^{\spadesuit}\right) - \ln\left(0.5X^S\right) \end{aligned}$$

So that

$$\psi_0\left(0.5\nu^{\spadesuit}\right) - \ln\left(0.5\nu^{\spadesuit}\right) = \underbrace{(1-w)\left(\psi_0\left(0.5\nu\right) - \ln\left({}^{\lfloor d}D\right)\right) + w\left(\psi_0\left(0.5\nu^U\right) - \ln\left({}^{\lfloor d}D^U\right)\right) - \ln\left(0.5X^S\right)}_{\Upsilon}$$

The proposition 15, proves that $\Upsilon < 0$. It, together with the proposition 10 implies that the equation (5.6) has always unique positive solution.

Now we have to check the conditions (5.10) and (5.11).

$$\begin{aligned} \frac{\partial^2 MO}{\partial\, {}^{\lfloor d}D^{\spadesuit 2}} &\overbrace{=}^{(B.4)} (1-w)\frac{0.5\nu^{\spadesuit}}{{}^{\lfloor d}D^{\spadesuit 2}} + w\frac{0.5\nu^{\spadesuit}}{{}^{\lfloor d}D^{\spadesuit 2}} = \frac{0.5\nu^{\spadesuit}}{{}^{\lfloor d}D^{\spadesuit 2}} > 0 \\ \frac{\partial^2 MO}{\partial \nu^{\spadesuit 2}} &\overbrace{=}^{(B.5)} (1-w)0.25\psi_1\left(0.5\nu^{\spadesuit}\right) + w0.25\psi_1\left(0.5\nu^{\spadesuit}\right) = 0.25\psi_1\left(0.5\nu^{\spadesuit}\right) > 0 \end{aligned}$$

The second condition holds, because the trigamma function is for positive arguments positive (see proposition 11).
Now we need to evaluate the determinant (5.12)

$$\begin{aligned} \frac{\partial^2 MO}{\partial \nu^{\spadesuit}\partial\, {}^{\lfloor d}D^{\spadesuit}} &\overbrace{=}^{(B.6)} -(1-w)\frac{0.5}{{}^{\lfloor d}D^{\spadesuit}} - w\frac{0.5}{{}^{\lfloor d}D^{\spadesuit}} = -\frac{0.5}{{}^{\lfloor d}D^{\spadesuit}} \\ \begin{vmatrix} \frac{\partial^2 f}{\partial x\partial y}(x_0,y_0) & \frac{\partial^2 f}{\partial x^2}(x_0,y_0) \\ \frac{\partial^2 f}{\partial y^2}(x_0,y_0) & \frac{\partial^2 f}{\partial x\partial y}(x_0,y_0) \end{vmatrix} &= \frac{0.25}{{}^{\lfloor d}D^{\spadesuit 2}} - \frac{0.5\nu^{\spadesuit}}{{}^{\lfloor d}D^{\spadesuit 2}}0.25\psi_1\left(0.5\nu^{\spadesuit}\right) = \\ &= \frac{0.25}{{}^{\lfloor d}D^{\spadesuit 2}}\left[1 - 0.5 * \nu^{\spadesuit}\psi_1\left(0.5\nu^{\spadesuit}\right)\right] < 0 \end{aligned}$$

The last inequality holds, because the function $x\psi_1(x) > 1, \forall x > 0$ (See proposition 11).

□

Straightforward application of Proposition 5 yields the following algorithm. Recall that $\Psi = [d, \psi]$.

Algorithm 5 (Updating $\lfloor^d D$ and ν) $\left(\lfloor^d D^{\spadesuit}, \nu^{\spadesuit}\right) = \text{UPDATE\_DFM}(w, C, \nu, \hat{\theta}, \lfloor^d D, \Psi)$

1. $\hat{e} = d - \hat{\theta}'\psi, \quad \zeta = \psi' C \psi$
2. $\nu^U = \nu + 1, \quad \lfloor^d D^U = \lfloor^d D + \frac{\hat{e}^2}{1+\zeta}$
3. $X^S = (1-w)\frac{\nu}{\lfloor^d D} + w\frac{\nu^U}{\lfloor^d D^U}$
4. $\Upsilon = (1-w)\left[\psi_0(0.5\nu) - \ln\left(\lfloor^d D\right)\right] + w\left[\psi_0\left(0.5\nu^U\right) - \ln\left(\lfloor^d D^U\right)\right] - \ln\left(0.5X^S\right)$
5. *Solve the equation for $\nu^{\spadesuit}$:* $\ln\left(0.5\nu^{\spadesuit}\right) - \psi_0\left(0.5\nu^{\spadesuit}\right) = \Upsilon$
6. $\lfloor^d D^{\spadesuit} = \frac{\nu^{\spadesuit}}{X^S}$

Remarks 10

1. *Step 5 must be solved numerically or using some suitable approximation.*
   For detail description of the numerical solution and for proof of unicity of the solution see [9].

5.3.2 Searching for $\hat{\theta}^{\spadesuit}$ and $C^{\spadesuit}$

Proposition 6 *For $\hat{\theta}^{\spadesuit}$ and $C^{\spadesuit}$ minimizing (5.3) it holds:*

$$C^{\spadesuit} = C + w_c z z' \tag{5.17}$$

$$\hat{\theta}^{\spadesuit} = \hat{\theta} + w_\theta z \tag{5.18}$$

where

$$\begin{aligned} z &= C\psi, \quad \hat{e} = d - \hat{\theta}'\psi, \quad \zeta = \psi' C \psi \\ w_c &= \left[\frac{\hat{e}^2}{(1+\zeta)^2}\frac{XX^U}{X+X^U} - \frac{w}{1+\zeta}\right], \quad w_\theta = \left[\frac{\hat{e}}{1+\zeta}\frac{X^U}{X+X^U}\right] \\ X &= (1-w)\frac{\nu}{\lfloor^d D}, \quad X^U = w\frac{\nu^U}{\lfloor^d D^U} \end{aligned}$$

Proof:

We again use the differential calculus to find the minimizer of function

$$GO\left(\hat{\theta}, C, \nu, \lfloor^d D, \hat{\theta}^U, C^U, \nu^U, \lfloor^d D^U, \hat{\theta}^{\spadesuit}, C^{\spadesuit}\right) = (1-w)G\left(\hat{\theta}, C, \nu, \lfloor^d D, \hat{\theta}^{\spadesuit}, C^{\spadesuit}\right) + wG\left(\hat{\theta}^U, C^U, \nu^U, \lfloor^d D^U, \hat{\theta}^{\spadesuit}, C^{\spadesuit}\right).$$

According to the form of function G, it is better to find ${C^{\spadesuit}}^{-1}$ rather than $C^{\spadesuit}$.

$$0 = (1-w)\frac{\partial G\left(\hat{\theta}, C, \nu, \lfloor^d D, \hat{\theta}^{\spadesuit}, C^{\spadesuit}\right)}{\partial {C^{\spadesuit}}^{-1}} + w\frac{\partial G\left(\hat{\theta}^U, C^U, \nu^U, \lfloor^d D^U, \hat{\theta}^{\spadesuit}, C^{\spadesuit}\right)}{\partial {C^{\spadesuit}}^{-1}} \tag{5.19}$$

$$0 = (1-w)\frac{\partial G\left(\hat{\theta}, C, \nu, \lfloor^d D, \hat{\theta}^{\spadesuit}, C^{\spadesuit}\right)}{\partial \hat{\theta}^{\spadesuit}} + w\frac{\partial G\left(\hat{\theta}^U, C^U, \nu^U, \lfloor^d D^U, \hat{\theta}^{\spadesuit}, C^{\spadesuit}\right)}{\partial \hat{\theta}^{\spadesuit}} \tag{5.20}$$

Using the relation (B.8) , the equation (5.20) yields:

$$\begin{aligned} 0 &= (1-w)\frac{\nu}{\lfloor^d D}{C^{\spadesuit}}^{-1'}\left(\hat{\theta} - \hat{\theta}^{\spadesuit}\right) + w\frac{\nu^U}{\lfloor^d D^U}{C^{\spadesuit}}^{-1'}\left(\hat{\theta}^U - \hat{\theta}^{\spadesuit}\right) \\ X^S\hat{\theta}^{\spadesuit} &= \underbrace{(1-w)\frac{\nu}{\lfloor^d D}}_{X}\hat{\theta} + \underbrace{w\frac{\nu^U}{\lfloor^d D^U}}_{X^U}\hat{\theta}^U \\ \hat{\theta}^{\spadesuit} &= \left(\frac{X}{X^S}\hat{\theta} + \frac{X^U}{X^S}(\hat{\theta} + h_U z)\right) = \hat{\theta} + h_U\frac{X^U}{X^S}z \end{aligned} \tag{5.21}$$

We used the relation $\hat{\theta}^U = \hat{\theta} + h_U z$ which was pronounced in section 5.2.
Using the relation (B.7) , the equation (5.19) yields:

$$\begin{aligned}
0 &= (1-w)\left[0.5(C - C^{\spadesuit}) + 0.5\frac{\nu}{\lfloor^d D}\left(\hat{\theta}-\hat{\theta}^{\spadesuit}\right)'\left(\hat{\theta}-\hat{\theta}^{\spadesuit}\right)\right] + \\
&\quad + w\left[0.5(C^U - C^{\spadesuit}) + 0.5\frac{\nu^U}{\lfloor^d D^U}\left(\hat{\theta}^U-\hat{\theta}^{\spadesuit}\right)'\left(\hat{\theta}^U-\hat{\theta}^{\spadesuit}\right)\right] \\
C^{\spadesuit} &= (1-w)C + wC^U + X\left(\hat{\theta}-\hat{\theta}^{\spadesuit}\right)'\left(\hat{\theta}-\hat{\theta}^{\spadesuit}\right) + X^U\left(\hat{\theta}^U-\hat{\theta}^{\spadesuit}\right)'\left(\hat{\theta}^U-\hat{\theta}^{\spadesuit}\right)
\end{aligned}$$

According to section 5.2 it holds:

$$\hat{\theta}^U = \hat{\theta} + h_U z \quad , \; h_U = \frac{\hat{e}}{1+\zeta} \tag{5.22}$$

$$C^U = C + g_U z z' \quad , \; g_U = -\frac{1}{1+\zeta} \tag{5.23}$$

$$z = C\psi \tag{5.24}$$

It holds:

$$\left(\hat{\theta}-\hat{\theta}^{\spadesuit}\right) = \hat{\theta} - \left(\frac{X}{X^S}\hat{\theta} + \frac{X^U}{X^S}(\hat{\theta}+h_U z)\right) = -h_U\frac{X^U}{X^S}z \tag{5.25}$$

$$\left(\hat{\theta}^U-\hat{\theta}^{\spadesuit}\right) = \hat{\theta} + h_U z - \left(\frac{X}{X^S}\hat{\theta} + \frac{X^U}{X^S}(\hat{\theta}+h_U z)\right) = -h_U\left(\frac{X^U}{X^S}-1\right)z \tag{5.26}$$

Hence

$$C^{\spadesuit} = C + w g_U z z' + X\left(h_U\frac{X^U}{X^S}\right)^2 z z' + X^U\left(h_U\left(\frac{X^U}{X^S}-1\right)\right)^2 z z'$$

$$C^{\spadesuit} = C + \left[w g_U + h_U^2\frac{X X^U}{X^S}\right] z z' \tag{5.27}$$

Simplifying the relations (5.27) and (5.21) we get:

$$\begin{aligned}
C^{\spadesuit} &= C + \left[\frac{\hat{e}^2}{(1+\zeta)^2}\frac{X X^U}{X+X^U} - \frac{w}{1+\zeta}\right] z z' \\
\hat{\theta}^{\spadesuit} &= \hat{\theta} + \left[\frac{\hat{e}}{1+\zeta}\frac{X^U}{X^S}\right] z
\end{aligned}$$

We must now check if the obtained matrix $C^{\spadesuit}$ is positive definite. The proposition 18 proves that.
We also have to prove that the result is a minimum. I.e we need to show the negative definitness of the Hessian matrix.
Because the function GO is a matrix function, we must obtain its hessian by formal differentiation with respect to some vector. Let's define vector $x = [\mathbf{vec} {C^{\spadesuit}}^{-1}; \hat{\theta}^{\spadesuit}]$
Hence we want to investigate the hessian of function

$$GO\left(\hat{\theta}, C, \nu, \lfloor^d D, \hat{\theta}^U, C^U, \nu^U, \lfloor^d D^U, x\right) \equiv GO\left(\hat{\theta}, C, \nu, \lfloor^d D, \hat{\theta}^U, C^U, \nu^U, \lfloor^d D^U, \hat{\theta}^{\spadesuit}, C^{\spadesuit}\right).$$

According to the relation (B.10) it holds:

$$\frac{\partial^2 G(\cdots,x)}{\partial x \partial x'} = \begin{pmatrix} 0.5C^{\spadesuit}\otimes C^{\spadesuit} & -\frac{\nu}{\lfloor^d D} I \otimes \left(\hat{\theta}-\hat{\theta}^{\spadesuit}\right)' \\ -\frac{\nu}{\lfloor^d D} I \otimes \left(\hat{\theta}-\hat{\theta}^{\spadesuit}\right)' & \frac{\nu}{\lfloor^d D} {C^{\spadesuit}}^{-1} \end{pmatrix}$$

$$\begin{aligned}
\frac{\partial^2 GO(\cdots,x)}{\partial x \partial x'} &= (1-w)\begin{pmatrix} 0.5C^{\spadesuit}\otimes C^{\spadesuit} & -\frac{\nu}{\lfloor^d D} I \otimes \left(\hat{\theta}-\hat{\theta}^{\spadesuit}\right)' \\ -\frac{\nu}{\lfloor^d D} I \otimes \left(\hat{\theta}-\hat{\theta}^{\spadesuit}\right)' & \frac{\nu}{\lfloor^d D} {C^{\spadesuit}}^{-1} \end{pmatrix} + \\
&+ \quad w\begin{pmatrix} 0.5C^{\spadesuit}\otimes C^{\spadesuit} & -\frac{\nu^U}{\lfloor^d D^U} I \otimes \left(\hat{\theta}^U-\hat{\theta}^{\spadesuit}\right)' \\ -\frac{\nu^U}{\lfloor^d D^U} I \otimes \left(\hat{\theta}^U-\hat{\theta}^{\spadesuit}\right)' & \frac{\nu^U}{\lfloor^d D^U} {C^{\spadesuit}}^{-1} \end{pmatrix}
\end{aligned}$$

The relations (5.25) and (5.26)gives:

$$\begin{aligned}\left(\hat{\theta} - \hat{\theta}^{\spadesuit}\right) &= -h_U \frac{X^U}{X^S} z \\ \left(\hat{\theta}^U - \hat{\theta}^{\spadesuit}\right) &= -h_U \left(\frac{X^U}{X^S} - 1\right) z\end{aligned}$$

using it, we get:

$$\begin{aligned}\frac{\partial^2 GO(\cdots,x)}{\partial x \partial x'} &= \left(\begin{array}{cc} 0.5C^{\spadesuit} \otimes C^{\spadesuit} & \overbrace{\left[X h_U \frac{X^U}{X^S} + X^U h_U \left(\frac{X^U}{X^S} - 1\right)\right]}^{=0} I \otimes z' \\ \left[X h_U \frac{X^U}{X^S} + X^U h_U \left(\frac{X^U}{X^S} - 1\right)\right] I \otimes z & X^S {C^{\spadesuit}}^{-1} \end{array}\right) = \\ &= \left(\begin{array}{cc} 0.5C^{\spadesuit} \otimes C^{\spadesuit} & 0 \\ 0 & X^S {C^{\spadesuit}}^{-1} \end{array}\right)\end{aligned}$$

$(X^S){C^{\spadesuit}}^{-1}$ is positive definite, because $C^{\spadesuit}$ is positive definite and X^S is a positive scalar.
The sentences about Kronecker product proves that $C^{\spadesuit} \otimes C^{\spadesuit}$ is also positive definite.
Now is the proposition proven. □

Algorithm 6 (Updating $\hat{\theta}$ and C) $(C^{\spadesuit}, \hat{\theta}^{\spadesuit}) = \text{UPDATE\_C}(w, C, \nu, \hat{\theta}, {}^{\lfloor d}D, \Psi)$

1. $\hat{e} = d - \hat{\theta}'\psi, \quad \zeta = \psi' C \psi$

2. $\nu^U = \nu + 1, \quad {}^{\lfloor d}D^U = {}^{\lfloor d}D + \frac{\hat{e}}{1+\zeta}$

3. $X = (1 - w)\frac{\nu}{{}^{\lfloor d}D}, \quad X^U = w \frac{\nu^U}{{}^{\lfloor d}D^U}$

4. $z = C\psi$

5. $C^{\spadesuit} = C + \left[\frac{\hat{e}^2}{(1+\zeta)^2} \frac{X X^U}{X + X^U} - \frac{w}{1+\zeta}\right] z z'$

6. $\hat{\theta}^{\spadesuit} = \hat{\theta} + \left[\frac{\hat{e}}{1+\zeta} \frac{X^U}{X^S}\right] z$

5.3.3 Algorithmic Aspects

We become some relations for updating the statistics of GiW distributions in its $C, \hat{\theta}, {}^{\lfloor d}D$ representation. Because the current solution uses the $L'DL$ representation of the statistics, we need to formulate the relations 5,6 in algorithm 6 in the same representation. We can use the proposition 26. It's clear that the relations from algorithm 6 can be formulated in the following form : ${}^{\lfloor\psi}V^{\spadesuit} = {}^{\lfloor\psi}V + w_1 \psi\psi'$, ${}^{\lfloor d\psi}V^{\spadesuit} = {}^{\lfloor d\psi}V + w_2 d\psi$, where the scalars w_1, w_2 are obtained by solving the equations:

$$\begin{aligned}w_C &= -\frac{w_1}{1 + w_1\zeta} \\ w_{\hat{\theta}} &= \frac{w_2 d + w_1(\hat{e} - d)}{1 + w_1\zeta}\end{aligned}$$

This equations have simple solution

$$\begin{aligned}w_1 &= -\frac{w_C}{1 + w_C\zeta} \\ w_2 &= \frac{w_{\hat{\theta}}(1 + w_1\zeta) - w_1(\hat{e} - d)}{d}\end{aligned}$$

5.4 Resulting PB algorithm

In this Section, we summarize all the elaborated parts into one consistent algorithm.

Algorithm 7 (PB)

Inputs - $\kappa_{\bullet;t-1},\ C_{\bullet\bullet;t-1},\ \hat{\theta}_{\bullet\bullet;t-1},\ {}^{\lfloor d}D_{\bullet\bullet;t-1},\ \nu_{\bullet\bullet;t-1},\ \Psi_{\bullet\bullet;t}$
Outputs - $\kappa_{\bullet;t},\ C_{\bullet\bullet;t},\ \hat{\theta}_{\bullet\bullet;t},\ {}^{\lfloor d}D_{\bullet\bullet;t},\ \nu_{\bullet\bullet;t}$

1. *For each factor ic:* $\mathcal{L}_{ic;t} = \mathrm{FACNORM}(C_{ic;t-1}, \hat{\theta}_{ic;t-1},\ {}^{\lfloor d}D_{ic;t-1}, \nu_{ic;t-1}, \Psi_{ic;t})$. *(algorithm 4)*

2. *Evaluate* $w_{\bullet;t} = \mathrm{EVAL\_WEIGHT}(\mathcal{L}_{\bullet\bullet;t}, \kappa_{\bullet;t-1})$. *(algorithm 1)*

3. *Evaluate* $\kappa_{\bullet;t} = \mathrm{NEW\_KAPPA}(w_{\bullet;t}, \kappa_{\bullet;t-1})$. *(algorithm 3)*

4. *For each factor ic:* $\left({}^{\lfloor d}D_{ic;t}, \nu_{ic;t}\right) = \mathrm{UPDATE\_DFM}(w_{c;t}, C_{ic;t-1}, \nu_{ic;t-1}, \hat{\theta}_{ic;t-1},\ {}^{\lfloor d}D_{ic;t-1}, \Psi_{ic;t})$. *(algorithm 5)*

5. *For each factor ic:* $\left(C_{ic;t}, \hat{\theta}_{ic;t}\right) = \mathrm{UPDATE\_C}(w_{c;t}, C_{ic;t-1}, \nu_{ic;t-1}, \hat{\theta}_{ic;t-1},\ {}^{\lfloor d}D_{ic;t-1}, \Psi_{ic;t})$. *(algorithm 6)*

Chapter 6

Comparison of PB and QB algorithms

In this Section, we compare the performance of the PB algorithm with the performance of the standard QB algorithm. The QB algorithm has been used extensively in real-life applications [11], and it is proven to be reliable and computationally efficient. Therefore, we study differences of the PB algorithm from the QB in terms of numerical properties and quality of estimation. The algorithms are based on different objective criteria for which they are optimal. Therefore, comparison of their behaviour is presented in a subjective way: arguing what seem to be more "rational".

In order to compare the analytical properties, we review the QB algorithm. Then, we investigate the differences between the two algorithms from analytical and computational point of view. Those finding are supported by experimental results.

6.1 The Quasi-Bayes algorithm

The general QB algorithm uses the following rule(see [4]):

$$\begin{aligned} \kappa_t &= \kappa_{t-1} + w_t \\ \pi_{ic}(\Theta_{ic}|\mathcal{S}_{ic;t}) &\propto [f_{ic}(d_{ic;t}|\psi_{ic;t}, \Theta_{ic})]^{w_{c;t}} \pi_{ic}(\Theta_{ic}|\mathcal{S}_{ic;t-1}) \end{aligned}$$

Let's mark the statistics corresponding to the QB algorithm by the subscript Q. Application of the general algorithm to the case with Normal factors yields:

$$V_Q = V + w\Psi\Psi',\ \nu_Q = \nu + w,\ \kappa_{Q_{\bullet;t}} = \kappa_{\bullet;t-1} + w_{\bullet;t} \tag{6.1}$$

We would receive exactly this result, if we approximate the KL distances in the PB algorithm with squares of euclidian norms of the parameter difference (see remarks 6 and 9).

For better comparison of the QB algorithm with the PB algorithm, we rewrite the relations (6.1) in terms of $C, \hat{\theta}$, ${}^{\lfloor d}D$:

$$C_Q = C + w_{QC}zz' \tag{6.2}$$

$$\hat{\theta}_Q = \hat{\theta} + w_{Q\theta}z\ , \quad {}^{\lfloor d}D_Q = {}^{\lfloor d}D + \frac{w\hat{e}^2}{1+w\zeta} \tag{6.3}$$

where

$$\begin{aligned} z &= C\psi, \quad \hat{e} = d - \hat{\theta}'\psi, \quad \zeta = \psi' C\psi \\ w_{QC} &= \frac{-w}{1+w\zeta}\ , \quad w_{Q\theta} = \frac{w\hat{e}}{1+w\zeta} \end{aligned}$$

6.2 Analytical comparison

Nature of both algorithms allows us to divide the analytical investigation into two parts. In the first part, we investigate the update of factors. This part is discussed next. In the second part, computing of the new component weights can be studied, see [9].

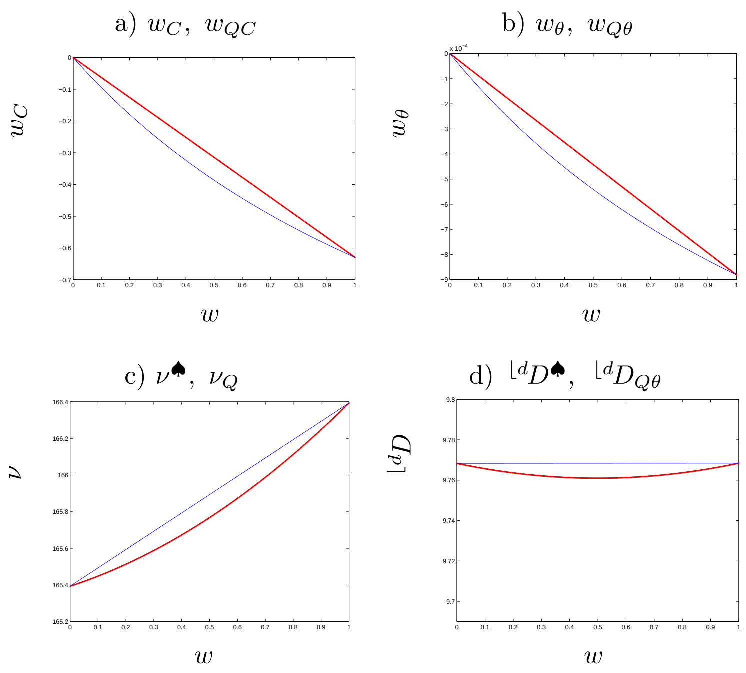

Figure 6.1: Similar behavior of the QB and PB algorithms for case a)

The figure shows the parameters $(\nu^\spadesuit, \nu_Q), (\lfloor^d D^\spadesuit,\ \lfloor^d D_Q), (w_C, w_{Qc}), (w_\theta, w_{Q\theta})$ *as the functions of* $w \in< 0, 1 >$ *for the case a)* $\nu = 165.39,\ \lfloor^d D = 9.77, \hat{e} = -0.0140, \zeta = 0.59$. *The parameters related to PB algorithm are plotted with the thick line. In this case the difference between the QB and PB algorithms is rather small.*

6.2.1 Differences of the algorithms

Note that the expressions for the QB update (6.2), (6.3) are very similar to the expressions for the PB update (5.17),(5.18). Hence, it suffice to investigate differences between the pairs $(\nu^\spadesuit, \nu_Q), (\lfloor^d D^\spadesuit,\ \lfloor^d D_Q)$ $(w_C, w_{QC}), (w_\theta, w_{Q\theta})$. This involves observation of 4 scalar variables, no matter what is the full dimension of the parameters.

We illustrates differences in behavior on the following examples. Consider the following situations:

a) $\nu = 165.39,\ \lfloor^d D = 9.77, \hat{e} = -0.0140, \zeta = 0.59$

b) $\nu = 102.82,\ \lfloor^d D = 1.14, \hat{e} = -0.7386, \zeta = 1.20$

The figures 6.1 and 6.2 shows the parameters $(\nu^\spadesuit, \nu_Q), (\lfloor^d D^\spadesuit,\ \lfloor^d D_Q), (w_C, w_{Qc}), (w_\theta, w_{Q\theta})$ as functions of $w \in< 0, 1 >$. The parameters related to the PB algorithm are plotted with the thick line. It is clear, that values obtained using PB equals to those of QB for $w = 0, w = 1$.

6.2.2 Bahaviour of the PB algorithm

In this Section, we study two particular factors and evaluate marginal distributions of their updates provided by both algorithms. For better comparison, we will also show the marginal pdf of the correct Bayesian update (4.1) which is a mixture of two GiW factors.

Consider the GiW factor $\pi(\Theta|\mathcal{S}) = GiW_{\theta,r}(V, \nu)$ and denote the associated densities as follows:

trial update $\pi(\Theta|\mathcal{S}^U) = GiW_{\theta,r}(V^U, \nu^U)$	$V^U = V + \Psi\Psi',\ \nu^U = \nu + 1$
QB update $\pi(\Theta|\mathcal{S}_Q) = GiW_{\theta,r}(V_Q, \nu_Q)$	$V_Q = V + w\Psi\Psi',\ \nu_Q = \nu + w$
PB update $\pi(\Theta|\mathcal{S}^\spadesuit) = GiW_{\theta,r}(V^\spadesuit, \nu^\spadesuit)$	result of the algorithms 5 and 6
correct update $\hat{\pi}(\Theta) = (1 - w)\pi(\Theta|\mathcal{S}) + w\pi(\Theta|\mathcal{S}^U)$	

Consider the statistics V, ν of the GiW factor, updating weights w and actual data vectors of the factor Ψ, to be:

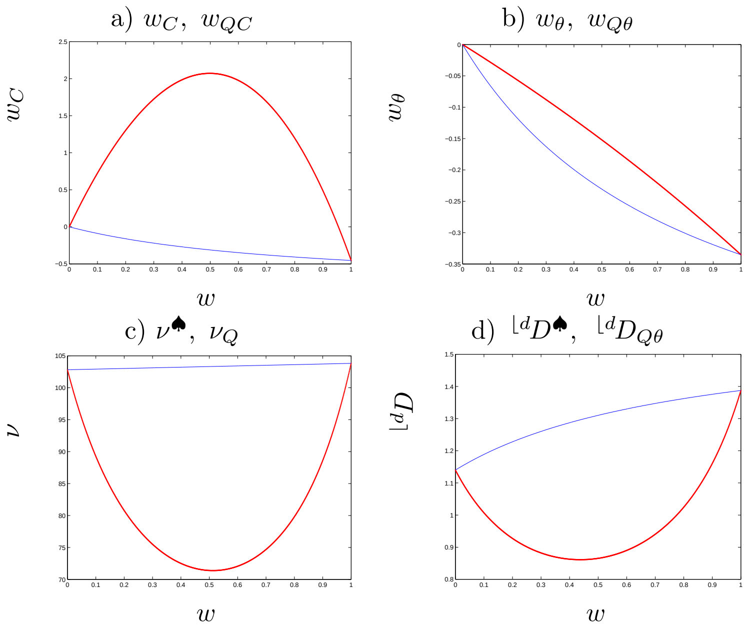

Figure 6.2: Different behavior of the QB and PB algorithms for case b)

The figure shows the parameters $(\nu^{\spadesuit}, \nu_Q), (\lfloor^{d}D^{\spadesuit}, \lfloor^{d}D_Q), (w_C, w_{Qc}), (w_\theta, w_{Q\theta})$ *as the functions of* $w \in < 0, 1 >$ *for the case b)* $\nu = 102.82$, $\lfloor^{d}D = 1.14, \hat{e} = -0.7386, \zeta = 1.20$. *The parameters related to the PB algorithm are plotted with the thick line. In this case, the difference between the QB and PB algorithms is significant.*

a)	*b)*
$V = \begin{pmatrix} 1.16 & 0.12 \\ 0.12 & 0.83 \end{pmatrix}$	$V = \begin{pmatrix} 1.96 & -1.47 \\ -1.47 & 6.07 \end{pmatrix}$
$\nu = 102.82$	$\nu = 108.06$
$\Psi = (-0.59 \quad 1)'$	$\Psi = (-0.79 \quad 1)'$
$w = 0.43$	$w = 0.39$

The figures 6.3 and 6.4 shows marginal pdfs of all discussed densities for both cases. From visual inspection of these figures, we can conclude that the PB algorithm can provide results significantly different from those of the QB algorithm. We also consider behavior of the PB algorithms as reasonable.

6.3 Experimental comparison

Intensive tests consisting of 1396 data sets were done. Data used for this test represent various types of systems (static, dynamic, multidimensional) and are part of standard testing procedure of new algorithms. As a quality measure, we used the likelihood [2] of the estimated model. For each set, we evaluated a criterion h which is the difference between the likelihood obtained by the PB algorithm and the QB algorithm. (i.e $h > 0$ if the PB algorithm was better.) The table 6.1 shows the results. Mean value of h over all sets is 6.18.

6.4 Comparing of computational complexity

We compare all 5 steps of the PB algorithm (algorithm 7).

1. This step is needed in both algorithms.
2. This step is needed in both algorithms.
3. We have to find minimizer of a convex function with $\mathring{c}$ variables. There exist a good approximation of the starting point for iterative numerical algorithm, which warrants quick solution of this task [9].
4. Solution of one-dimensional nonlinear equation must be found. However, a good approximation which always leads to solving the equation in a few steps was found [9].
5. This step has the same complexity in both algorithms.

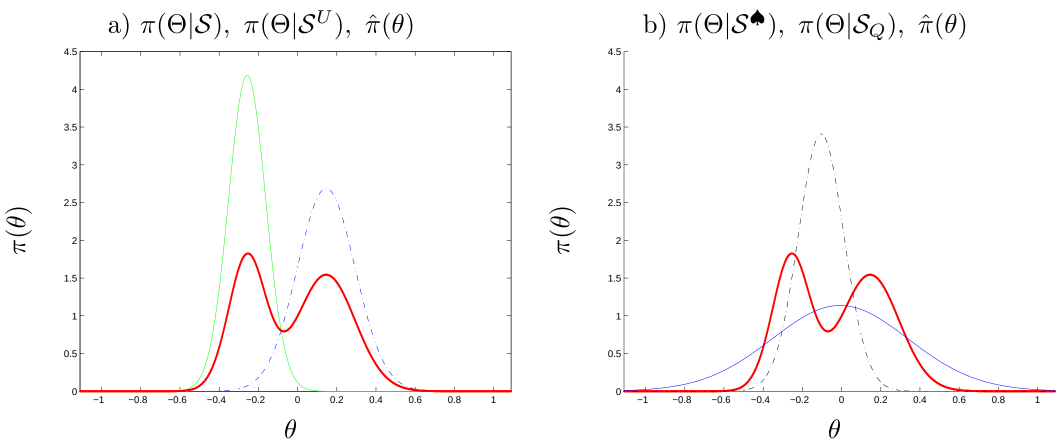

Figure 6.3: Marginal pdfs of the QB and PB updates for the case a)

The left part shows original factor (dashdot), its trial update (dotted) and the correct Bayesian update (thick), i.e. the mixture of the two mentioned factors. The right part shows how the QB update (dashdot) and the PB update (solid) approximates the correct Bayesian update (thick). It can be seen that the PB update is in this case flatter then the QB update which concentrates on smaller interval.

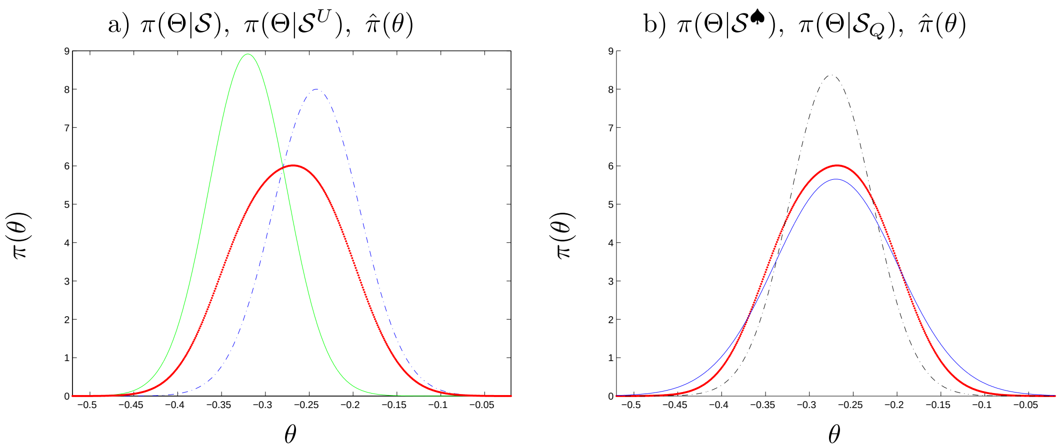

Figure 6.4: Marginal pdfs of the QB and PB updates for the case b)

The left part shows original factor (dashdot), its trial update (dotted) and the correct Bayesian update (thick), i.e. the mixture of the two mentioned factors. The right part shows how the QB update (dashdot) and the PB update (solid) approximates the correct Bayesian update (thick). It can be seen that the PB update in this case better approximates the correct pdf.

condition	number of sets	percentage
$h > 0$	1125	80.6%
$h < 0$	271	19.4%
abs(h) < 2	1126	80.6%
$h > 2$	251	18.0%
$h < -2$	19	1.4%

Table 6.1: Results of experimental comparison

The table shows some conditions for h and number of sets fulfilling each condition.

Addressing the previous considerations, we conclude that computational cost of numerical evaluation of the PB algorithm is comparable to the computational cost associated with the QB algorithms. Detailed case study of the computational costs of both algorithms can be found in [12].

Conclusions

This work describes a novel and efficient algorithm for recursive estimation of finite probabilistic mixture. The algorithm has the potential of providing more accurate results than the well-established quasi-Bayes estimator. This improvement is important as mixtures represent a universal approximating tool for modelling of non-linear stochastic systems. Therefore, mixture models can be used to address complex control and decision-making problems in changing environments, such as multiple-participants decision making. Each participant (or group of participants) can be modelled by a component of the overall mixture model. All subsequent decision-making task can be easily formalized within the consistent formal framework of probabilistic mixture models. We believe, that the algorithms presented in this paper will be an important part of this framework.

Acknowledgments

This work was supported by AV ČR S1075351, GA ČR 102/03/0049

Appendix A

Proximity meassures

A.1 Kullback-Leibler distance

Kullback-Leibler distance measures well proximity of a pair of pdfs. Let f, g be a pair of pdfs acting on a common set x^*. Then, the Kullback-Leibler distance $\mathcal{D}(f||g)$ is defined by the formula

$$\mathcal{D}(f||g) \equiv \int_{x^*} f(x) \ln \left(\frac{f(x)}{g(x)} \right) dx. \tag{A.1}$$

For conciseness, the Kullback-Leibler distance is referred to as the *KL distance*.

A.2 Kerridge distance

We can rearrange the expression of KL distance:

$$\int_{x^*} f(x) \ln \left(\frac{f(x)}{g(x)} \right) dx = \int_{x^*} f(x) \ln \left(f(x) \right) dx - \int_{x^*} f(x) \ln \left(g(x) \right) dx \tag{A.2}$$

It's clear that the first element doesn't influence the result when minimizing the KL distance with respect to the function $g(x)$. We define so called *Kerridge distance*:

Let f, g be a pair of pdfs acting on a common set x^*. Then, the Kerridge distance $\mathcal{K}(f||g)$ is defined by the formula

$$\mathcal{K}(f||g) \equiv -\int_{x^*} f(x) \ln \left(g(x) \right) dx. \tag{A.3}$$

For conciseness, the Kerridge distance distance is referred to as the *K distance*.

Proposition 7

$$\operatorname{Arg} \min_g \mathcal{D}(f||g) = \operatorname{Arg} \min_g \mathcal{K}(f||g) \tag{A.4}$$

Proof:

$$\min_g \mathcal{D}(f||g) = \min_g \int f \ln \frac{f}{g} = \min_g \left\{ \int f \ln f - \int f \ln g \right\} = \int f \ln f + \min_g \left\{ - \int f \ln g \right\}$$

$\square$

Proposition 8

$$\mathcal{K}\left(\sum_{c=1}^{\mathring{c}} \alpha_c f_c(x) \;\Big|\Big|\; g(x) \right) = \sum_{c=1}^{\mathring{c}} \alpha_c \mathcal{K}\left(f_c(x) \;\Big|\Big|\; g(x) \right) \tag{A.5}$$

Proof:

$$\mathcal{K}\left(\sum_{c=1}^{\mathring{c}} \alpha_c f_c(x) \;\Big|\Big|\; g(x) \right) = -\int \sum_{c=1}^{\mathring{c}} \alpha_c f_c(x) \ln(g(x)) = \sum_{c=1}^{\mathring{c}} \alpha_c \left\{ -\int f_c(x) \ln(g(x)) \right\}$$

$\square$

Proposition 9

$$\mathcal{K}\Big(f(x)h(y) \;\Big|\Big|\; g(x)v(y)\Big) = \mathcal{K}\Big(f(x) \;\Big|\Big|\; g(x)\Big) + \mathcal{K}\Big(h(y) \;\Big|\Big|\; v(y)\Big) \tag{A.6}$$

Proof:

$$\begin{aligned}
\mathcal{K}\Big(f(x)h(y) \;\Big|\Big|\; g(x)v(y)\Big) &= -\int f(x)h(y)\ln\Big(g(x)v(y)\Big)dxdy = -\int f(x)h(y)\Big(\ln(g(x)) + \ln(v(y))\Big)dxdy = \\
&= -\int f(x)h(y)\ln(g(x))dxdy - \int f(x)h(y)\ln(v(y))dxdy = \\
&= -\int f(x)\ln(g(x))dx - \int h(y)\ln(v(y)dy
\end{aligned}$$

□

Appendix B

Auxiliary propositions

B.1 Properties of the digamma and trigamma functions

$$\text{digamma} \qquad \psi_0(x) = \frac{d \ln(\Gamma(x))}{dx}$$

$$\text{trigamma} \qquad \psi_1(x) = \frac{d\psi_0(x)}{dx}$$

Proposition 10 (Properties of the function $h(x) = \psi_0(x) - \ln(x)$)

- *$h(x)$ is for positive arguments increasing and negative.*
- $h((0, +\infty)) = (-\infty, 0)$

Proposition 11 (Properties of the function $\psi_1(x)$)

- *$\psi_1(x)$ is for positive arguments decreasing and positive.*
- $x\psi_1(x) > 1, \forall x > 0$

Remarks 11 *Proofs of the presented propositions can be found for example in [9].*

B.2 Other relations

Proposition 12 *Let $\sum_{c=1}^{\mathring{c}} w_{c;t} = 1$. It holds:*

$$\sum_{j,c=1}^{\mathring{d},\mathring{c}} w_{c;t}\mathcal{K}_{jc}^{U} + \sum_{c=1}^{\mathring{c}} w_{c;t} \sum_{\substack{j,r=1 \\ r\neq c}}^{\mathring{d},\mathring{c}} \mathcal{K}_{jr} = \sum_{j,c=1}^{\mathring{d},\mathring{c}} \left[w_{c;t}\mathcal{K}_{jc}^{U} + (1 - w_{c;t})\mathcal{K}_{jc}\right] \tag{B.1}$$

Proof:

$$\sum_{j,c=1}^{\mathring{d},\mathring{c}} w_{c;t}\mathcal{K}_{jc}^{U} + \sum_{c=1}^{\mathring{c}} w_{c;t} \sum_{\substack{j,r=1 \\ r\neq c}}^{\mathring{d},\mathring{c}} \mathcal{K}_{jr} = \sum_{j,c=1}^{\mathring{d},\mathring{c}} w_{c;t}\mathcal{K}_{jc}^{U} + \sum_{c=1}^{\mathring{c}} w_{c;t} \left[\sum_{j,r=1}^{\mathring{d},\mathring{c}} \mathcal{K}_{jr} - \sum_{j=1}^{\mathring{d}} \mathcal{K}_{jc}\right] =$$

$$= \sum_{j,c=1}^{\mathring{d},\mathring{c}} w_{c;t}\mathcal{K}_{jc}^{U} + \sum_{j,r=1}^{\mathring{d},\mathring{c}} \mathcal{K}_{jr} - \sum_{j,c=1}^{\mathring{d},\mathring{c}} w_{c;t}\mathcal{K}_{jc} = \sum_{j,c=1}^{\mathring{d},\mathring{c}} \left[w_{c;t}\mathcal{K}_{jc}^{U} + (1 - w_{c;t})\mathcal{K}_{jc}\right]$$

$\square$

Proposition 13 *Let function M be defined as follows:*

$$M\left(\nu, {}^{\lfloor d}D, \nu^{\spadesuit}, {}^{\lfloor d}D^{\spadesuit}\right) = \ln\left(\Gamma\left(0.5\nu^{\spadesuit}\right)\right) + 0.5\nu^{\spadesuit} \ln\left(\frac{{}^{\lfloor d}D}{{}^{\lfloor d}D^{\spadesuit}}\right) - 0.5\nu^{\spadesuit}\psi_0(0.5\nu) + 0.5\frac{\nu\, {}^{\lfloor d}D^{\spadesuit}}{{}^{\lfloor d}D}$$

The partial derivatives can be evaluated in the following way:

$$\frac{\partial M(\nu, {}^{\lfloor d}D, \nu^{\spadesuit}, {}^{\lfloor d}D^{\spadesuit})}{\partial\, {}^{\lfloor d}D^{\spadesuit}} = 0.5\left(\frac{\nu}{{}^{\lfloor d}D} - \frac{\nu^{\spadesuit}}{{}^{\lfloor d}D^{\spadesuit}}\right) \tag{B.2}$$

$$\frac{\partial M(\nu, {}^{\lfloor d}D, \nu^{\spadesuit}, \nu^{\spadesuit})}{\partial\, {}^{\lfloor D}\nu^{\spadesuit}} = 0.5\left(\psi_0\left(0.5\nu^{\spadesuit}\right) - \psi_0\left(0.5\nu\right) + \ln\left(\frac{{}^{\lfloor d}D}{{}^{\lfloor d}D^{\spadesuit}}\right)\right) \tag{B.3}$$

$$\frac{\partial^2 M(\nu, {}^{\lfloor d}D, \nu^{\spadesuit}, {}^{\lfloor d}D^{\spadesuit})}{\partial^2\, {}^{\lfloor d}D^{\spadesuit}} = 0.5\frac{\nu^{\spadesuit}}{{}^{\lfloor d}D^{\spadesuit 2}} \tag{B.4}$$

$$\frac{\partial^2 M(\nu, {}^{\lfloor d}D, \nu^{\spadesuit}, \nu^{\spadesuit})}{\partial^2\, {}^{\lfloor D}\nu^{\spadesuit}} = 0.25\psi_1\left(0.5\nu^{\spadesuit}\right) \tag{B.5}$$

$$\frac{\partial^2 M(\nu, {}^{\lfloor d}D, \nu^{\spadesuit}, \nu^{\spadesuit})}{\partial\, {}^{\lfloor D}\nu^{\spadesuit}\partial\, {}^{\lfloor d}D^{\spadesuit}} = -\frac{0.5}{{}^{\lfloor d}D^{\spadesuit}} \tag{B.6}$$

Proof: Straightforward evaluation □

Proposition 14 *Let function G be defined as follows:*

$$G\left(\hat\theta, C, \nu, {}^{\lfloor d}D, \hat\theta^{\spadesuit}, C^{\spadesuit}\right) = -0.5\ln\left|CC^{\spadesuit^{-1}}\right| + 0.5\mathrm{tr}\left[CC^{\spadesuit^{-1}}\right] + 0.5\frac{\nu}{{}^{\lfloor d}D}\left(\hat\theta - \hat\theta^{\spadesuit}\right)' C^{\spadesuit^{-1}}\left(\hat\theta - \hat\theta^{\spadesuit}\right)$$

Then the partial derivatives can be evaluated as follows:

$$\frac{\partial G(\hat\theta, C, \hat\theta^{\spadesuit}, C^{\spadesuit})}{\partial C^{\spadesuit^{-1}}} = -0.5C^{\spadesuit}C^{-1}C + 0.5C + 0.5\frac{\nu}{{}^{\lfloor d}D}\left(\hat\theta - \hat\theta^{\spadesuit}\right)'\left(\hat\theta - \hat\theta^{\spadesuit}\right) \tag{B.7}$$

$$\frac{\partial G(\hat\theta, C, \hat\theta^{\spadesuit}, C^{\spadesuit})}{\partial \hat\theta^{\spadesuit}} = -\frac{\nu}{{}^{\lfloor d}D}C^{\spadesuit^{-1\prime}}\left(\hat\theta - \hat\theta^{\spadesuit}\right) \tag{B.8}$$

$$\tag{B.9}$$

In order to evaluate the hessian of the function G, we have to formally differentiate it with respect to some vector. Let's define vector $x = [\mathbf{vec}C^{\spadesuit^{-1}}, \hat\theta^{\spadesuit}]$.

$$G\left(\hat\theta, C, \nu, {}^{\lfloor d}D, x\right) \equiv G\left(\hat\theta, C, \nu, {}^{\lfloor d}D, \hat\theta^{\spadesuit}, C^{\spadesuit}\right)$$

$$\frac{\partial^2 G(x)}{\partial x \partial x'} = \begin{pmatrix} 0.5C^{\spadesuit}\otimes C^{\spadesuit} & -\frac{\nu}{{}^{\lfloor d}D}I\otimes\left(\hat\theta - \hat\theta^{\spadesuit}\right)' \\ -\frac{\nu}{{}^{\lfloor d}D}I\otimes\left(\hat\theta - \hat\theta^{\spadesuit}\right)' & \frac{\nu}{{}^{\lfloor d}D}C^{\spadesuit^{-1}} \end{pmatrix} \tag{B.10}$$

Proof: Almost everything is straightforward a evaluation. Just a little formalism will be recalled. Let's denote $y = \mathbf{vec}C^{\spadesuit^{-1}}$

$$\frac{\partial^2 G(\cdots, x)}{\partial x \partial x'} = \begin{pmatrix} \frac{\partial^2 G(\cdots,x)}{\partial y \partial y'} & \frac{\partial^2 G(\cdots,x)}{\partial \hat\theta^{\spadesuit}\partial y'} \\ \frac{\partial^2 G(\cdots,x)}{\partial y \partial(\hat\theta^{\spadesuit})'} & \frac{\partial^2 G(\cdots,x)}{\partial \hat\theta^{\spadesuit}\partial(\hat\theta^{\spadesuit})'} \end{pmatrix}$$

$$\frac{\partial^2 G(\cdots, x)}{\partial y \partial y'} = 0.5C^{\spadesuit}\otimes C^{\spadesuit} \tag{B.11}$$

$$\frac{\partial G(\cdots, x)}{\partial(\hat\theta^{\spadesuit})'} = -\frac{\nu}{{}^{\lfloor d}D}C^{\spadesuit^{-1}}\left(\hat\theta - \hat\theta^{\spadesuit}\right) \tag{B.12}$$

$$\frac{\partial^2 G(\cdots, x)}{\partial y \partial(\hat\theta^{\spadesuit})'} = -\frac{\nu}{{}^{\lfloor d}D}I\otimes\left(\hat\theta - \hat\theta^{\spadesuit}\right)' \tag{B.13}$$

$$\frac{\partial G(\cdots, x)}{\partial(\hat\theta^{\spadesuit})\partial(\hat\theta^{\spadesuit})'} = \frac{\nu}{{}^{\lfloor d}D}C^{\spadesuit^{-1}} \tag{B.14}$$

□

Proposition 15 *Let's suppose all notations from the section 5.3.1. It holds:*

$$\Upsilon < 0$$

Proof:

$$\begin{aligned}
\Upsilon &= (1-w)\psi_0\left(0.5\nu\right) + w\psi_0\left(0.5\nu^U\right) - \ln\left(0.5\,{}^{\lfloor d}D^{1-w}\,{}^{\lfloor d}D^{U^w}X^S\right) \\
&= (1-w)\psi_0\left(0.5\nu\right) + w\psi_0\left(0.5\nu^U\right) - \ln\left(0.5\,{}^{\lfloor d}D^{1-w}\,{}^{\lfloor d}D^{U^w}\left((1-w)\frac{\nu}{{}^{\lfloor d}D} + w\frac{\nu^U}{{}^{\lfloor d}D^U}\right)\right) \\
&= (1-w)\psi_0\left(0.5\nu\right) + w\psi_0\left(0.5\nu^U\right) - \ln\left((1-w)0.5\nu\left(\frac{{}^{\lfloor d}D^U}{{}^{\lfloor d}D}\right)^w + w0.5\nu^U\left(\frac{{}^{\lfloor d}D^U}{{}^{\lfloor d}D}\right)^{w-1}\right)
\end{aligned}$$

Let's denote $x = \frac{{}^{\lfloor d}D^U}{{}^{\lfloor d}D}$. We will now show that the most pessimistic estimate of $\Upsilon \equiv \Upsilon(x, w, \nu, \nu^U)$ is negative. It is clear that $\Upsilon <= \max_x \Upsilon(x, w, \nu, \nu^U)$

$$\frac{\partial\Upsilon}{\partial x} = -\frac{(1-w)w0.5\nu x^{w-1} + w(w-1)0.5\nu^U x^{w-2}}{((1-w)0.5\nu x^w + w0.5\nu^U x^{w-1})}$$

Now we have to solve the equation $\frac{\partial\Upsilon}{\partial x} = 0$

$$(1-w)w0.5x^{w-2}\left(\nu x - \nu^U\right) = 0$$

$$\nu x - \nu^U = 0, x = \frac{\nu^U}{\nu}$$

Because the denominator of the derivativ is always positive, is easy to see, that $x = \frac{\nu^U}{\nu}$ is global maximizer of Υ. We proved that $\Upsilon(x, w, \nu, \nu^U) <= \Upsilon_1(w, \nu, \nu^U) \equiv \Upsilon(\frac{\nu^U}{\nu}, w, \nu, \nu^U)$
Next, we analyze the function Υ_1

$$\begin{aligned}
\Upsilon_1(w, \nu, \nu^U) &= -\ln\left(0.5\nu\left(\frac{\nu^U}{\nu}\right)^w\right) + 2 * \left((1-w)0.5\psi_0\left(0.5\nu\right) + w0.5\psi_0\left(0.5\nu^U\right)\right) = \\
&= -\ln(0.5\nu) - w\ln\left(\frac{\nu^U}{\nu}\right) + 2 * \left((1-w)0.5\psi_0\left(0.5\nu\right) + w0.5\psi_0\left(0.5\nu^U\right)\right)
\end{aligned}$$

Now we want to find $\max_w \Upsilon_1(w, \nu, \nu^U)$

$$\frac{\partial\Upsilon_1}{\partial w} = -\ln\left(\frac{\nu^U}{\nu}\right) - \psi_0\left(0.5\nu\right) + \psi_0\left(0.5\nu^U\right) = \left(\ln(0.5\nu) - \psi_0\left(0.5\nu\right)\right) - \left(\ln(0.5\nu^U) - \psi_0\left(0.5\nu^U\right)\right)$$

Because the function $\ln x - \psi_0\left(x\right)$ is for positive arguments positive and descending, the derivative $\frac{\partial\Upsilon_1}{\partial w}$ is positive and thus Υ_1 is maximized by $w = 1$. (note that $\nu^U = \nu + 1$)
We proved $\Upsilon <= \Upsilon_2(\nu, \nu^U) \equiv \Upsilon_1(1, \nu, \nu^U)$

$$\Upsilon_2(\nu, \nu^U) = \Upsilon_3(\nu^U) = \psi_0\left(0.5\nu^U\right) - \ln(0.5\nu^U)$$

We know that this function is for positive arguments negative(see proposition 10). We proved $\Upsilon <= \Upsilon_3(\nu^U) < 0$
□

Proposition 16 *Let matrix C be regular and matrix A be symmetric and positive definite. Then the matrix $C'AC$ is symmetric positive definite.*

Proof:
The matrix A is positive definite, i.e for each $y \neq 0$ it holds: $y'Ay > 0$. We want to show that for each $x \neq 0$, $x'C'ACx > 0$.
C is regular, hence $Cx \neq 0$ for $x \neq 0$, hence $x'C'ACx = \underbrace{(Cx)'}_{z'}A\underbrace{(Cx)}_{z} = z'Az > 0$ □

Proposition 17 (Determinant of the matrix I+xx') *Let x be a column vector of the length n. Then*

$$|I + xx'| = 1 + x'x$$

Proof: First, we will prove that x is eigenvector of the matrix $(I + xx')$ with eigenvalue $1 + x'x$.

$$(I + xx')x = x + xx'x = x(1 + x'x) = (1 + x'x)x$$

Let's now take such linear independent vectors $y_1, \cdots, y_{\mathring{x}-1}$, so that $x'y_i = 0,\ \forall i$. We will prove, that this vectors are eigenvectors of the matrix $(I + xx')$ with eigenvalues 1.

$$(I + xx')y_i = y_i + xx'y_i = y_i + x(x'y_i) = y_i$$

$\square$

Proposition 18 *Let's suppose all notations from the section 5.3.2. $C^{\spadesuit} = C + w_C zz'$ is positive definite*

Proof: C is positive definite, hence there exists the square root $C^{\frac{1}{2}}$: $C = C^{\frac{1}{2}}C^{\frac{1}{2}}$, which is symmetric and regular.

$$C^{\spadesuit} = C^{\frac{1}{2}}\left(I + w_C C^{-\frac{1}{2}}zz'C^{-\frac{1}{2}}\right)C^{\frac{1}{2}}$$

Thanks to the proposition 16 it suffice only to prove the posit. definitnes of the matrix:

$$\left(I + w_C C^{-\frac{1}{2}}zz'C^{-\frac{1}{2}}\right).$$

According to the proposition 17 , the sufficient condition for the previous matrix to be positive definite is

$$0 < 1 + w_C z'C^{-\frac{1}{2}}C^{-\frac{1}{2}}z = 1 + w_C z'C^{-1}z.$$

Because $z = C\psi$ and $\psi'z = \zeta$, it suffice to check

$$0 < 1 + w_C\psi'CC^{-1}z = 1 + w_C\psi'z = 1 + w_C\zeta$$

$$1 + w_C\zeta = 1 + \left[wg_U + h_U^2\frac{XX^U}{X^S}\right]\zeta = 1 + wg_U\zeta + \underbrace{h_U^2\frac{XX^U}{X^S}}_{>0}$$

$$1 + wg_U\zeta = 1 - \frac{w\zeta}{1+\zeta} = \frac{1 + \zeta(1-w)}{1+\zeta} > 0$$

The last inequality holds because $w \leq 1$ and $\zeta = \psi'z = \psi'C\psi > 0$ (C is positive definite) $\square$

Appendix C

Normal factors and its properties

Because this chapter deals with only one factor in one specific time moment, we can omit the indexes $_{ic;t}$. i.e.

$$f_{ic}(d_{ic;t}|\psi_{ic;t}, \Theta_{ic}) \equiv f(d|\psi, \Theta)$$

C.1 Gauss dynamic pdf

The normal parameterized factor, we deal with, predicts a real-valued variable d by the pdf

$$f(d|\psi, \Theta) = \mathcal{N}_d(\theta'\psi, r), \text{ where} \tag{C.1}$$

$\Theta \equiv [\theta, r] \equiv$ [regression coefficients, noise variance] $\in \Theta^* \subset$ ($\mathring{\psi}$-dimensional, non-negative) real variables,

$$\mathcal{N}_d(\theta'\psi, r) \equiv (2\pi r)^{-0.5} \exp\left\{-\frac{(d-\theta'\psi)^2}{2r}\right\}. \tag{C.2}$$

C.2 Gauss-invers wishart dynamic pdf

C.2.1 Definition

Normal factors belong to the exponential family, so that they possess conjugate (self-reproducing) prior. This pdf is known as Gauss-inverse-Wishart pdf (GiW).

$$GiW_\Theta(V, \nu) \quad \equiv \quad GiW_{\theta,r}(V,\nu) \equiv \frac{r^{-0.5(\nu+\mathring{\psi}+2)}}{\mathcal{I}(V,\nu)} \exp\left\{-\frac{1}{2r}\mathrm{tr}\left(V[-1,\theta']'[-1,\theta']\right)\right\}. \tag{C.3}$$

The value of the normalization integral $\mathcal{I}(V, \nu)$ is described below, together with other properties of this important pdf.

C.2.2 Parameters

The parameter ν is positive scalar. The parameter V is $(\mathring{\Psi}, \mathring{\Psi})$-dimensional symmetric *positive definite extended information matrix*.

We often manipulate the matrix V through it's $L'DL$ decomposition. (i.e. with lower triangular matrix L and diagonal matrix D which fulfills the relation $V = L'DL$)

Let us split the information matrix V and its $L'DL$ decomposition as follows:

$$\begin{aligned} V &= \begin{bmatrix} {}^{\lfloor d}V & {}^{\lfloor d\psi}V' \\ {}^{\lfloor d\psi}V & {}^{\lfloor \psi}V \end{bmatrix}, \ {}^{\lfloor d}V \text{ is scalar,} \\ L &= \begin{bmatrix} 1 & 0 \\ {}^{\lfloor d\psi}L & {}^{\lfloor \psi}L \end{bmatrix}, \ D = \begin{bmatrix} {}^{\lfloor d}D & 0 \\ 0 & {}^{\lfloor \psi}D \end{bmatrix}, \ {}^{\lfloor d}D \text{ is scalar.} \end{aligned} \tag{C.4}$$

Next, the matrices L and D can be equivalently expressed with help of matrix C, vector $\hat{\theta}$ and scalar ${}^{\lfloor d}D$ defined by:

$$\hat{\theta} \equiv {}^{\lfloor \psi}L^{-1}\, {}^{\lfloor d\psi}L \equiv \text{ least-squares estimate of } \theta \tag{C.5}$$

$$C \equiv {}^{\lfloor \psi}L^{-1}\, {}^{\lfloor \psi}D^{-1} \left({}^{\lfloor \psi}L'\right)^{-1} \equiv \text{ covariance factor of least-squares estimate} \tag{C.6}$$

Proposition 19 *It holds:*

$$C = {}^{\lfloor\psi}V^{-1} \tag{C.7}$$
$$\hat{\theta} = {}^{\lfloor\psi}V^{-1}\,{}^{\lfloor d\psi}V \tag{C.8}$$

Proof: The relation (C.7) is clear from the following form of the matrix V.

$$V = L'DL = \begin{bmatrix} 1 & {}^{\lfloor d\psi}L' \\ 0 & {}^{\lfloor\psi}L' \end{bmatrix}\begin{bmatrix} {}^{\lfloor d}D & 0 \\ 0 & {}^{\lfloor\psi}D \end{bmatrix}\begin{bmatrix} 1 & 0 \\ {}^{\lfloor d\psi}L & {}^{\lfloor\psi}L \end{bmatrix} \tag{C.9}$$
$$= \begin{bmatrix} {}^{\lfloor d}D & {}^{\lfloor d\psi}L'\,{}^{\lfloor\psi}D \\ 0 & {}^{\lfloor\psi}L'\,{}^{\lfloor\psi}D \end{bmatrix}\begin{bmatrix} 1 & 0 \\ {}^{\lfloor d\psi}L & {}^{\lfloor\psi}L \end{bmatrix} = \tag{C.10}$$
$$= \begin{bmatrix} {}^{\lfloor d}D + {}^{\lfloor d\psi}L'\,{}^{\lfloor\psi}D\,{}^{\lfloor d\psi}L & {}^{\lfloor d\psi}L'\,{}^{\lfloor\psi}D\,{}^{\lfloor\psi}L \\ {}^{\lfloor\psi}L'\,{}^{\lfloor\psi}D\,{}^{\lfloor d\psi}L & {}^{\lfloor\psi}L'\,{}^{\lfloor\psi}D\,{}^{\lfloor\psi}L \end{bmatrix} \tag{C.11}$$

The proof od the second relation is simple.

$${}^{\lfloor\psi}V^{-1}\,{}^{\lfloor d\psi}V = {}^{\lfloor\psi}L^{-1}\,{}^{\lfloor\psi}D^{-1}\,{}^{\lfloor\psi}L'^{-1}\;\;{}^{\lfloor\psi}L'\,{}^{\lfloor\psi}D\,{}^{\lfloor d\psi}L = {}^{\lfloor\psi}L^{-1}\,{}^{\lfloor d\psi}L = \hat{\theta}$$

□

Proposition 20 (Alternative expressions of the GiW pdf) *$GiW_\Theta(V,\nu)$ has the following alternative expressions*

$$GiW_\Theta(V,\nu) \equiv \frac{r^{-0.5(\nu+\mathring{\psi}+2)}}{\mathcal{I}(L,D,\nu)}\exp\left\{-\frac{1}{2r}\left[\left({}^{\lfloor\psi}L\theta - {}^{\lfloor d\psi}L\right)'\,{}^{\lfloor\psi}D\left({}^{\lfloor\psi}L\theta - {}^{\lfloor d\psi}L\right) + {}^{\lfloor d}D\right]\right\} \equiv$$
$$\equiv \frac{r^{-0.5(\nu+\mathring{\psi}+2)}}{\mathcal{I}(C,\hat{\theta},\,{}^{\lfloor d}D,\nu)}\exp\left\{-\frac{1}{2r}\left[(\theta-\hat{\theta})'C^{-1}(\theta-\hat{\theta}) + {}^{\lfloor d}D\right]\right\} \tag{C.12}$$

Proposition 21 (Normalization integral)
The normalization integral is

$$\mathcal{I}(L,D,\nu) = \Gamma(0.5\nu)\,{}^{\lfloor d}D^{-0.5\nu}\left|{}^{\lfloor\psi}D\right|^{-0.5} 2^{0.5\nu}(2\pi)^{0.5\mathring{\psi}} \tag{C.13}$$

Repeatedly, we need to evaluate the KL distance of a pair of GiW pdfs.

Proposition 22 (KL distance of GiW pdfs) *Let $f(\Theta) = GiW_\Theta(L,D,\nu)$, $\tilde{f}(\Theta) = GiW_\Theta\left(\tilde{L},\tilde{D},\tilde{\nu}\right)$ be a pair of GiW pdfs of parameters $\Theta \equiv (\theta,r) =$ (regression coefficients, noise variance). Let D_{ii} stand for the diagonal element of the matrix D. Then, the KL distance of f and $\tilde{f}$ is given by the formula*

$$\begin{aligned}\mathcal{D}\left(f||\tilde{f}\right) &= \ln\left(\frac{\Gamma(0.5\tilde{\nu})}{\Gamma(0.5\nu)}\right) - 0.5\ln\left|C\tilde{C}^{-1}\right| + 0.5\tilde{\nu}\ln\left(\frac{{}^{\lfloor d}D}{{}^{\lfloor d}\tilde{D}}\right) + \\ &+ 0.5(\nu-\tilde{\nu})\psi_0(0.5\nu) - 0.5\mathring{\psi} - 0.5\nu + 0.5\mathrm{tr}\left[C\tilde{C}^{-1}\right] + \\ &+ 0.5\frac{\nu}{{}^{\lfloor d}D}\left[\left(\hat{\theta}-\hat{\tilde{\theta}}\right)'\tilde{C}^{-1}\left(\hat{\theta}-\hat{\tilde{\theta}}\right) + {}^{\lfloor d}\tilde{D}\right].\end{aligned} \tag{C.14}$$

Proposition 23 (Estimation of the normal factor)

$$\frac{GiW_\Theta(V_{t-1},\nu_{t-1})\mathcal{N}_{d_t}(\theta'\psi_t,r)}{\int dtto\;\; d\Theta} = GiW_\Theta(V_{t-1}+\Psi_t\Psi_t',\nu_{t-1}+1) \tag{C.15}$$

Proposition 24 (Prediction of the normal factor) *The predictive pdf is known as* Student pdf. *For any data vector $\Psi = [d,\psi']'$, its values can be found numerically as the ratio*

$$f(d|\psi,L,D,\nu) = \frac{\Gamma(0.5(\nu+1))\left[{}^{\lfloor d}D(1+\zeta)\right]^{-0.5}}{\sqrt{\pi}\Gamma(0.5\nu)\left(1+\frac{\hat{e}^2}{{}^{\lfloor d}D(1+\zeta)}\right)^{0.5(\nu+1)}}, \quad \textit{where} \tag{C.16}$$
$$\hat{e} \equiv d - \hat{\theta}'\psi \equiv \textit{prediction error}$$
$$\zeta \equiv \psi'C\psi,$$

The relations for evaluating ζ and $\hat{e}$ are based on $(C, \hat{\theta}, {}^{\lfloor d}D)$ representation of matrix V. We will now show how to compute these characteristic from $L'DL$ representation without converting between representations.

Proposition 25 (Evaluating $\zeta, \hat{e}$ in $L'DL$ representation)

$$\begin{aligned} \hat{e} &= d - {}^{\lfloor d\psi}L'x \\ \zeta &= x'\,{}^{\lfloor\psi}D^{-1}x, \quad \textit{where } {}^{\lfloor\psi}L'x = \psi \end{aligned}$$

Proof:

$$\begin{aligned} \hat{e} &= d - \hat{\theta}'\psi = d - {}^{\lfloor d\psi}L'\left({}^{\lfloor\psi}L'\right)^{-1}\psi \\ \zeta &= \psi'\,{}^{\lfloor\psi}L^{-1}\,{}^{\lfloor\psi}D^{-1}\left({}^{\lfloor\psi}L'\right)^{-1}\psi \\ &\text{Let's denote} \\ x &= \left({}^{\lfloor\psi}L'\right)^{-1}\psi \\ &\text{then the characteristics become the form} \\ \hat{e} &= d - {}^{\lfloor d\psi}L'x \\ \zeta &= x'\,{}^{\lfloor\psi}D^{-1}x \end{aligned}$$

□

Remarks 12

1. *We can see that with known vector x the evaluation of $\hat{e}$ consists of computing scalar product of two vectors. Similarly, because matrix ${}^{\lfloor\psi}D$ is diagonal, evaluation of ζ consists of "scalar product" of three vectors.*

2. *The vector x can be simply computed from the equation ${}^{\lfloor\psi}L'x = \psi$. Because ${}^{\lfloor\psi}L'$ is triangular matrix with unite diagonal, the solving consists of backward run.*

Algorithm 8 (Evaluating $\zeta, \hat{e}$ in $L'DL$ representation) $(\zeta, \hat{e}) = \mathrm{GETCHARS}({}^{\lfloor\psi}L, {}^{\lfloor\psi}D, {}^{\lfloor d\psi}L, \Psi)$

1. *Solve* $\left({}^{\lfloor\psi}L'\right)^{-1}x = \psi$

2. $\hat{e} = d - {}^{\lfloor d\psi}L'x$

3. $\zeta = x'\,{}^{\lfloor\psi}D^{-1}x$

C.3 Properties of the operation $V + w\Psi\Psi'$

This section deals with important operation with GiW parameters. First of all, try to rewrite this operation into block form. Recall that $\Psi = \begin{bmatrix} d \\ \psi \end{bmatrix}$.

$$V = L'DL = \begin{bmatrix} {}^{\lfloor d}D + {}^{\lfloor d\psi}L'\,{}^{\lfloor\psi}D\,{}^{\lfloor d\psi}L & {}^{\lfloor d\psi}L'\,{}^{\lfloor\psi}D\,{}^{\lfloor\psi}L \\ {}^{\lfloor\psi}L'\,{}^{\lfloor\psi}D\,{}^{\lfloor d\psi}L & {}^{\lfloor\psi}L'\,{}^{\lfloor\psi}D\,{}^{\lfloor\psi}L \end{bmatrix} \tag{C.17}$$

$$\Psi\Psi' = \begin{bmatrix} d \\ \psi \end{bmatrix}\begin{bmatrix} d & \psi' \end{bmatrix} = \begin{bmatrix} d^2 & d\psi' \\ \psi d & \psi\psi' \end{bmatrix} \tag{C.18}$$

Proposition 26 *Let's the matrices $C, \hat{\theta}, L, D, V$ are defined according to (C.4),(C.5), (C.6). Then the operation*

$${}^{\lfloor\psi}V^{\spadesuit} = {}^{\lfloor\psi}V + w_1\psi\psi', \; {}^{\lfloor d\psi}V^{\spadesuit} = {}^{\lfloor d\psi}V + w_2 d\psi$$

can be rewritten to

$$\begin{aligned} C^{\spadesuit} &= C - \frac{w_1}{1 + w_1\zeta}zz' \\ \hat{\theta}^{\spadesuit} &= \hat{\theta} + \frac{w_2 d + w_1(\hat{e} - d)}{1 + w_1\zeta}z \\ z &= C\psi \end{aligned}$$

Proof: First let's prove the part dealing with matrix C. According to the relation (C.7), we need to investigate the element ${}^{\lfloor\psi}V^{\spadesuit^{-1}} = C^{\spadesuit}$. We know that ${}^{\lfloor\psi}V^{\spadesuit} = {}^{\lfloor\psi}V + w_1\psi\psi'$. Using the well known proposition about inversion it holds:

$$
{}^{\lfloor\psi}V^{\spadesuit^{-1}} = {}^{\lfloor\psi}V^{-1} - \frac{w_1\,{}^{\lfloor\psi}V^{-1}\psi\psi'\,{}^{\lfloor\psi}V^{-1}}{1 + w_1\psi'\,{}^{\lfloor\psi}V^{-1}\psi}
$$

thus:

$$
C^{\spadesuit} = C - \frac{w_1}{1 + w_1\zeta}zz'
$$

Now we will prove the part dealing with $\hat{\theta}$. Using the relation (C.8), the element $\hat{\theta}^{\spadesuit}$ can be expressed:

$$
\hat{\theta}^{\spadesuit} = {}^{\lfloor\psi}V^{\spadesuit^{-1}}\,{}^{\lfloor d\psi}V^{\spadesuit}
$$

$$
\begin{aligned}
\hat{\theta}^{\spadesuit} &= \left({}^{\lfloor\psi}V^{-1} - \frac{w_1\,{}^{\lfloor\psi}V^{-1}\psi\psi'\,{}^{\lfloor\psi}V^{-1}}{1 + w_1\psi'\,{}^{\lfloor\psi}V^{-1}\psi}\right)\left({}^{\lfloor d\psi}V + w_2 d\psi\right) = \\
&= \hat{\theta} + \left[{}^{\lfloor\psi}V^{-1}\psi - \frac{w_1\,{}^{\lfloor\psi}V^{-1}\psi\psi'\,{}^{\lfloor\psi}V^{-1}\psi}{1 + w_1\psi'\,{}^{\lfloor\psi}V^{-1}\psi}\right]w_2 d - \frac{w_1\,{}^{\lfloor\psi}V^{-1}\psi\psi'\,{}^{\lfloor\psi}V^{-1}}{1 + w_1\psi'\,{}^{\lfloor\psi}V^{-1}\psi}\,{}^{\lfloor d\psi}V = \\
&= \hat{\theta} + \left[{}^{\lfloor\psi}V^{-1}\psi\left(1 - \frac{w_1\psi'\,{}^{\lfloor\psi}V^{-1}\psi}{1 + w_1\psi'\,{}^{\lfloor\psi}V^{-1}\psi}\right)\right]w_2 d - \frac{w_1\,{}^{\lfloor\psi}V^{-1}\psi\psi'\hat{\theta}}{1 + w_1\psi'\,{}^{\lfloor\psi}V^{-1}\psi} = \\
&= \hat{\theta} + \frac{{}^{\lfloor\psi}V^{-1}\psi}{1 + w_1\psi'\,{}^{\lfloor\psi}V^{-1}\psi}w_2 d - \frac{w_1\,{}^{\lfloor\psi}V^{-1}\psi\psi'\hat{\theta}}{1 + w_1\psi'\,{}^{\lfloor\psi}V^{-1}\psi} = \\
&= \hat{\theta} + \frac{{}^{\lfloor\psi}V^{-1}\psi}{1 + w_1\psi'\,{}^{\lfloor\psi}V^{-1}\psi}\left(w_2 d - w_1\psi'\hat{\theta}\right) = \hat{\theta} + \frac{w_2 d + w_1(\hat{e} - d)}{1 + w_1\zeta}z
\end{aligned}
$$

□

Proposition 27 (Determinant of matrix $L'DL + \Psi\Psi'$) *Let the matrices V, ${}^{\lfloor\psi}V$, vectors Ψ, ψ and scalars $\hat{e}, \zeta$ be defined as common. $w \in <0, 1>$. It holds:*

$$
\begin{aligned}
|V + w\Psi\Psi'| &= |V|\left((1 + w\zeta + w\frac{\hat{e}^2}{{}^{\lfloor d}D}\right) \\
\left|{}^{\lfloor\psi}V + w\psi\psi'\right| &= \left|{}^{\lfloor\psi}V\right|(1 + w\zeta)
\end{aligned}
$$

Proof:

$$
\begin{aligned}
|L'DL + w\Psi\Psi'| &= \left|L'\sqrt{D}(\sqrt{D}L + wD^{-\frac{1}{2}}L'^{-1}\Psi\Psi')\right| = \left|L'\sqrt{D}(I + wD^{-\frac{1}{2}}L'^{-1}\Psi\Psi'L^{-1}D^{-\frac{1}{2}})\sqrt{D}L\right| = \\
&= \left|L'\sqrt{D}\right|\left|I + wD^{-\frac{1}{2}}L'^{-1}\Psi\Psi'L^{-1}D^{-\frac{1}{2}}\right|\left|\sqrt{D}L\right| = \text{(proposition 17)} \\
&= |L'DL|\,(1 + w\Psi'L^{-1}D^{-\frac{1}{2}}D^{-\frac{1}{2}}L'^{-1}\Psi) = \\
&= |L'DL|\,(1 + w\Psi'L^{-1}D^{-1}L'^{-1}\Psi) \\
&\ \text{analogically} \\
\left|{}^{\lfloor\psi}L'\,{}^{\lfloor\psi}D\,{}^{\lfloor\psi}L + w\Psi\Psi'\right| &= \left|{}^{\lfloor\psi}L'\,{}^{\lfloor\psi}D\,{}^{\lfloor\psi}L\right|(1 + w\psi'\underbrace{{}^{\lfloor\psi}L^{-1}\,{}^{\lfloor\psi}D^{-1}\,{}^{\lfloor\psi}L'^{-1}}_{C}\psi) = \left|{}^{\lfloor\psi}L'\,{}^{\lfloor\psi}D\,{}^{\lfloor\psi}L\right|(1 + w\zeta)
\end{aligned}
$$

Lets' denote $x = L'^{-1}\Psi$. x is obtained by solving the equation $L'x = \Psi$.

$$
\Psi'L^{-1}D^{-1}L'^{-1}\Psi = x'D^{-1}x = \sum_{i=1}^{n}\frac{x_i^2}{D_{i,i}}.
$$

Analogically denote $y = {}^{\lfloor\psi}L'^{-1}\psi$.

$$
\zeta = \psi'\,{}^{\lfloor\psi}L^{-1}\,{}^{\lfloor\psi}D^{-1}\,{}^{\lfloor\psi}L'^{-1}\psi = y'\,{}^{\lfloor\psi}D^{-1}y = \sum_{i=1}^{n-1}\frac{y_i^2}{{}^{\lfloor\psi}D_{i,i}}.
$$

Because ${}^{\lfloor d\psi}L$ resp. ${}^{\lfloor d\psi}D$ resp. ψ are parts of L resp. D, resp. Ψ, the vector y is part of x. Exactly: $y = (x_2, \cdots, x_n)$ Hence:

$$
\sum_{i=1}^{n-1}\frac{y_i^2}{{}^{\lfloor\psi}D_{i,i}} = \sum_{i=2}^{n}\frac{x_i^2}{D_{i,i}}
$$

Using the previous relations we can obtain the following expression:

$$\Psi' L^{-1} D^{-1} L'^{-1} \Psi = \zeta + \frac{x_1^2}{{}^{\lfloor d}D}$$

Now it is important to evaluate the first element of the vector x.

$$x_1 = d - {}^{\lfloor d\psi}L' y = d - {}^{\lfloor d\psi}L' \,{}^{\lfloor \psi}L'^{-1} \psi = d - \theta' \psi = \hat{e}$$

$$|L'DL| \left(1 + w \Psi' L^{-1} D^{-1} L'^{-1} \Psi\right) = |L'DL| \left(1 + w \left(\zeta + \frac{\hat{e}^2}{{}^{\lfloor d}D}\right)\right)$$

□

Proposition 28 *Let the matrices $C, \hat{\theta}, L, D, V$ be defined according to (C.4),(C.5), (C.6). Then the operation*

$$V^{\spadesuit} = V + w\Psi\Psi'$$

can be rewritten to

$$\begin{aligned} C^{\spadesuit} &= C - \frac{w}{1+w\zeta} zz' \\ \hat{\theta}_2 &= \hat{\theta} + \frac{w\hat{e}}{1+w\zeta} z \\ {}^{\lfloor d}D^{\spadesuit} &= {}^{\lfloor d}D + \frac{w\hat{e}^2}{1+w\zeta} \\ z &= C\psi \end{aligned}$$

Proof: The first two parts can be simple proven using the proposition 26 and considering $w_1 = w_2 = w$. The third part can be simply proved using the proposition 27.

$${}^{\lfloor d}D^{\spadesuit} = \frac{|D^{\spadesuit}|}{|{}^{\lfloor \psi}D^{\spadesuit}|} = \frac{|V^{\spadesuit}|}{|{}^{\lfloor \psi}V^{\spadesuit}|} = \frac{|V|\left(1 + w\zeta + w\frac{\hat{e}^2}{{}^{\lfloor d}D}\right)}{|{}^{\lfloor \psi}V|\left(1+w\zeta\right)} = {}^{\lfloor d}D \left(1 + \frac{w\frac{\hat{e}^2}{{}^{\lfloor d}D})}{1+w\zeta}\right) = {}^{\lfloor d}D + \frac{w\hat{e}^2}{1+w\zeta}$$

□

C.4 Dirichlet multivariete pdf

C.4.1 Definition

$Di_\alpha(\kappa)$ denotes Dirichlet pdf of $\alpha \in \alpha^* \equiv \left\{\alpha_c \geq 0 : \sum_{c\in c^*} \alpha_c = 1\right\}$ of the form :

$$Di_\alpha(\kappa) \equiv \frac{\prod_{c\in c^*} \alpha_c^{\kappa_c - 1}}{\mathcal{B}(\kappa)}, \;\mathcal{B}(\kappa) \equiv \frac{\prod_{c\in c^*} \Gamma(\kappa_c)}{\Gamma(\sum_{c\in c^*} \kappa_c)}.$$

C.4.2 Properties

$$\begin{aligned} \hat{\alpha}_c &= \frac{\kappa_c}{\sum_{c=1}^{\mathring{c}} \kappa_c} && \text{(C.19)} \\ \alpha_c Di_\alpha(\kappa) &= \hat{\alpha}_c Di_\alpha(\kappa + \delta_{\bullet,c}) && \text{(C.20)} \\ \varepsilon\left[\alpha_c | \kappa\right] &= \hat{\alpha}_c && \text{(C.21)} \end{aligned}$$

Proof:

$$\begin{aligned} \mathcal{B}\left(\kappa + \delta_{\bullet,c}\right) &= \frac{\Gamma(\kappa_c + 1) \prod\limits_{k=1,k\neq c} \Gamma(\kappa_k)}{\Gamma\left(\sum \kappa_k + 1\right)} = \frac{\kappa_c \prod\limits_{k=1} \Gamma(\kappa_k)}{\Gamma\left(\sum \kappa_k\right) \sum \kappa_k} = \mathcal{B}\left(\kappa\right) \hat{\alpha}_c \\ \alpha_c Di_\alpha(\kappa) &= \alpha_c \frac{\prod\limits_{k=1}^{\mathring{c}} \alpha_k^{\kappa_k - 1}}{\mathcal{B}\left(\kappa\right)} = \hat{\alpha}_c \frac{\prod\limits_{k=1}^{\mathring{c}} \alpha_k^{\kappa_k - 1 + \delta_{k,c}}}{\mathcal{B}\left(\kappa + \delta_{\bullet,c}\right)} = \hat{\alpha}_c Di_\alpha(\kappa + \delta_{\bullet,c}) \\ \varepsilon\left[\alpha_c | \kappa\right] &= \int \alpha_c Di_\alpha(\kappa) d\alpha = \hat{\alpha}_c \int Di_\alpha(\kappa + \delta_{\bullet c}) d\alpha = \hat{\alpha}_c \end{aligned}$$

□

Proposition 29 (KL distance of Di pdfs) *Let $f(\alpha) = Di_{\alpha}(\kappa)$, $\tilde{f}(\alpha) = Di_{\alpha}(\tilde{\kappa})$ be a pair of Dirichlet pdfs of parameters $\alpha \equiv (\alpha_1, \ldots, \alpha_{\mathring{c}}) \in \alpha^* = \left\{\alpha_c > 0,\ \sum_{c \in c^*} \alpha_c = 1\right\}$, $c^* \equiv \{1, \ldots, \mathring{c}\}$.*

Their KL distance is given by the formula

$$\begin{aligned}
\mathcal{D}(f||\tilde{f}) &= \sum_{c=1}^{\mathring{c}} \left[(\kappa_c - \tilde{\kappa}_c)\, \psi_0\left(\kappa_c\right) + \ln\left(\frac{\Gamma\left(\tilde{\kappa}_c\right)}{\Gamma\left(\kappa_c\right)}\right)\right] - (\nu - \tilde{\nu})\psi_0\left(\nu\right) + \ln\left(\frac{\Gamma(\nu)}{\Gamma(\tilde{\nu})}\right) \\
\nu &\equiv \sum_{c=1}^{\mathring{c}} \kappa_c, \quad \tilde{\nu} \equiv \sum_{c=1}^{\mathring{c}} \tilde{\kappa}_c.
\end{aligned} \tag{C.22}$$

Moreover it holds:

$$\operatorname{Arg}\min_{\tilde{\kappa}} \mathcal{D}(f||\tilde{f}) = \operatorname{Arg}\min_{\tilde{\kappa}} \sum_{j=1}^{\mathring{c}} \left[\ln\left(\Gamma\left(\tilde{\kappa}_j\right)\right) - \tilde{\kappa}_j \psi_0\left(\kappa_j\right)\right] - \left[\ln(\Gamma(\tilde{\nu}) - \tilde{\nu}\psi_0\left(\nu\right)\right] \tag{C.23}$$

Bibliography

[1] D.M. Titterington, A.F.M. Smith, and U.E. Makov, *Statistical Analysis of Finite Mixtures*, John Wiley & Sons, Chichester, New York, Brisbane, Toronto, Singapore, 1985, ISBN 0 471 90763 4.

[2] M. Kárný, J. Böhm, T. V. Guy, and P. Nedoma, "Mixture-based adaptive probabilistic control", *International Journal of Adaptive Control and Signal Processing*, vol. 17, no. 2, pp. 119–132, 2003.

[3] V. Peterka, "Bayesian system identification", in *Trends and Progress in System Identification*, P. Eykhoff, Ed., pp. 239–304. Pergamon Press, Oxford, 1981.

[4] M. Kárný, J. Kadlec, and E. L. Sutanto, "Quasi-Bayes estimation applied to normal mixture", in *Preprints of the 3rd European IEEE Workshop on Computer-Intensive Methods in Control and Data Processing*, J. Rojíček, M. Valečková, M. Kárný, and K. Warwick, Eds., Praha, September 1998, pp. 77–82, ÚTIA AV ČR.

[5] S. Kullback and R. Leibler, "On information and sufficiency", *Annals of Mathematical Statistics*, vol. 22, pp. 79–87, 1951.

[6] L. Berec and M. Kárný, "Identification of reality in Bayesian context", in *Computer-Intensive Methods in Control and Signal Processing: Curse of Dimensionality*, K. Warwick and M. Kárný, Eds., pp. 181–193. Birkhäuser, 1997.

[7] S. J. Roberts and W. D. Penny, "Variational Bayes for generalized autoregressive models", *IEEE Transactions on Signal Processing*, vol. 50, no. 9, pp. 2245–2257, 2002.

[8] L. He and M. Kárný, "Bayesian Modelling and Estimation of ARMAX Model and Its Finite Mixtures", Tech. Rep. 2080, ÚTIA AV ČR, Praha, 2003.

[9] P. Nenutil, "Přibližné průběžné odhadování dynamických směsových modelů", Tech. Rep., ÚTIA AV ČR, Praha, 2004.

[10] M. Abramowitz and I.A. Stegun, *Handbook of mathematical functions*, Dover Publications, Inc., New York, 1972.

[11] M. Kárný, J. Böhm, T.V. Guy, L. Jirsa, I. Nagy, P. Nedoma, L. Tesař, and M. Tichý, "Prodactool background – theory, algorithms and software", Tech. Rep., UTIA AV CR, 2003, Draft of the report, 401 pp.

[12] P. Nedoma and J. Andrýsek, "Benchmarks in mixture processing", in *Proceedings of the International Workshop on Multiple Participant Decision Making 2004*, M. Kárný, J. Kracík, and J. Andrýsek, Eds., Praha, May 2004, ÚTIA AV ČR.