Advanced Bayesian Estimation in Hierarchical Gaussian Models: Dirichlet Process Mixtures and Clustering Gain

MASTER'S THESIS

for obtaining the academic degree

Diplom-Ingenieur

as part of the study

Telecommunications

carried out by

Erik Šauša

student number: 1525267

February 2024

Supervision:

Ao.Univ.Prof. Dipl.-Ing. Dr.techn. Franz Hlawatsch
Dipl.-Ing. Thomas John Bucco

Institute of Telecommunications
TU Wien

Statement on Academic Integrity

Hiermit erkläre ich, dass die vorliegende Arbeit gemäß dem Code of Conduct – Regeln zur Sicherung guter wissenschaftlicher Praxis (in der aktuellen Fassung des jeweiligen Mitteilungsblattes der TU Wien), insbesondere ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel, angefertigt wurde. Die aus anderen Quellen direkt oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet. Die Arbeit wurde bisher weder im In– noch im Ausland in gleicher oder in ähnlicher Form in anderen Prüfungsverfahren vorgelegt.

Vienna, February 2024

Author's signature

Supervisor's signature

Acknowledgements

I extend my gratitude to my supervisor, Ao.Univ.Prof. Franz Hlawatsch, whose mentorship has been invaluable throughout this journey. His dedication to academic excellence, unwavering support, and insightful suggestions have significantly shaped this thesis. I am indebted to Dr.techn. Güther Koliander for his expertise in mathematical derivations and proofs, which made this work possible. Additionally, I am grateful to Dipl.-Ing. Bernd Kreidl and Dipl.-Ing. Thomas John Bucco for their valuable insights shared during our collaborative meetings. A special acknowledgment goes to my family for their steadfast encouragement and support. Their belief in my abilities has been a constant source of motivation. Lastly, I express my deepest appreciation to my fiancée, Karen, whose support, motivation, and last but not least, assistance with graphics have been indispensable throughout this endeavor.

Abstract

This Master's thesis explores the use of a Dirichlet process (DP) prior to enhance Bayesian estimation of the parameters of multiple objects. Specifically, it focuses on a hierarchical Gaussian model where each object is linked to one parameter of interest, one noisy measurement, and one hyperparameter, and the hyperparameters are shared among objects within the same cluster. The model permits the derivation of closed-form performance bounds, enabling a quantification of performance improvements relative to the theoretically achievable performance. Our primary objective is to estimate the parameter of interest for each object based on its associated noisy measurement while leveraging the cluster structure of the hyperparameters. Because a closed-form calculation of the posterior distribution is not possible, we employ a Markov chain Monte Carlo sampling method to approximate the minimum mean square error (MMSE) estimator. This methodology yields an estimator that exploits the inherent cluster structure and, as we show through simulations, consistently achieves a mean squared error (MSE) that is lower than the MSE of the MMSE estimator for a scenario without a cluster structure. Additionally, we derive a closed-form MMSE estimator assuming known object-cluster associations and demonstrate its performance through simulations. Our approach of combining estimation and clustering demonstrates superior performance compared to the widely used method of first clustering and then performing estimation within each cluster; however, this performance advantage comes at the cost of a higher computational complexity.

Contents

1 Introduction

1.1 Motivation and Contributions

This Master's thesis investigates the use of a Dirichlet process (DP) prior in the context of Bayesian estimation. We focus on multiple-object estimation within a hierarchical Gaussian model where each object is associated with one parameter of interest, one hyperparameter, and one noisy measurement. The objects are assigned to clusters that are defined by shared hyperparameters. The primary aim is to estimate the parameter of interest for each object from the associated noisy measurement while leveraging the cluster structure of the hyperparameters.

To solve this estimation problem, we present a technique where the measurements are clustered simultaneously with estimating the parameters of interest. This integration of clustering into estimation results in more precise estimates of the parameter of interest, compared to traditional methods that perform clustering and estimation as separate steps. We adopt a Bayesian approach, which makes it possible to seamlessly integrate prior knowledge with observed measurements [1]. The resulting posterior distribution provides not only estimates of the parameters of interest but also a detailed characterization of the uncertainty associated with these estimates.

We consider four different scenarios of the Gaussian estimation problem that differ in the prior distribution and the available prior knowledge; in particular, two of the four scenarios involve a DP prior. For each scenario, we calculate the minimum mean square error (MMSE) estimator and the corresponding MMSE. More specifically, we derive closed-form expressions of the MMSE estimator and the MMSE for three scenarios and a Markov chain Monte Carlo (MCMC) [2] approximation for the remaining scenario. These results allow us to quantify the clustering gain (i.e., the reduction in MMSE due to the use of the DP prior and the integration of clustering) and to establish performance bounds for our proposed clustering-aided estimator. These results are complemented by a detailed introduction to the theoretical foundations and statistical properties of the DP, which builds on [3] and, in contrast to the original definition in [4], does not require measure theory.

1.2 State of the Art

In recent years, statistical modeling has shifted towards more flexible methods to capture complex dataset structures. Dirichlet process mixtures (DPMs) have emerged as powerful tools across various domains, including machine learning and Bayesian statistics [5]. Introduced by Ferguson in the early 1970s [4], the DP provides a non-parametric framework for modeling uncertainty about the number of clusters in a dataset. Unlike fixed-size mixture models, DPMs adapt dynamically to data complexity by accommodating an infinite number of latent components, thereby eliminating the need to pre-specify the number of clusters. DPMs facilitate a robust modeling of complex real-world datasets, making them a compelling choice for tasks ranging from clustering and capturing latent structures to density estimation in diverse applications, such as target tracking [6] [7], robotics [8] [9], and medical imaging [10] [11].

Since the complexity of DPMs often renders direct analytical solutions intractable, Monte Carlo methods [2] or variational inference (VI) methods [12] are typically used to calculate approximations to the posterior distributions. In particular, MCMC algorithms like Gibbs sampling [13] and Metropolis-Hastings sampling [14] provide means to generate samples from the posterior distribution. A detailed review of MCMC sampling algorithms is provided in [6]. A VI-based coordinate ascent variational inference (CAVI) algorithm for the Gaussian estimation problem is derived in [15].

1.3 Thesis Outline

Following the introductory Chapter 1, we provide an overview of Bayesian estimation and consider the Gaussian distribution in Chapter 2.

In Chapter 3, we introduce and discuss the DP and the DPM. In particular, we focus on the properties and distributions associated with the DP and derive four different procedures that can be used for generating samples from the DP. Furthermore, we discuss the DP's relation to clustering and consider the Chinese restaurant process (CRP). Finally, we introduce the DPM and the distributions associated with it.

In Chapter 4, we present our general statistical model and derive closed-form expressions for the MMSE estimator and the MMSE for two benchmark scenarios that do not involve the DP nor any classification or clustering.

In Chapter 5, we consider two scenarios that involve the DP and inherent clustering.

We provide MCMC approximations for the MMSE estimator with inherent clustering and derive the distributions occurring in the Gibbs sampler. Furthermore, we derive closed-form expressions for the MMSE estimator and the MMSE for the case where the cluster association of each object is known.

Simulation results are presented in Chapter 6. We study the performance of the estimators for the four scenarios using artificial data generated according to our Gaussian models. Furthermore, we compare the performance of our estimators with the performance of estimators that use two standard clustering algorithms (*DBSCAN* [16] and *K-Means* ++ [17]).

Chapter 7 concludes the thesis with a summary of the main results and some suggestions for future research.

1.4 Notation

Table 1 provides an overview of the notation used in this thesis.

Table 1: Summary of notation

x	deterministic scalar
x	random scalar
$\boldsymbol{x}$	deterministic vector
$\mathbf{x}$	random vector
$\boldsymbol{X}$	deterministic matrix
$\mathbf{X}$	random matrix
$\mathcal{C}$	deterministic set
$\mathscr{C}$	random set
$\{\boldsymbol{x}\}$	set consisting of single vector $\boldsymbol{x}$
$\lvert\mathcal{C}\rvert$	cardinality of set $\mathcal{C}$
$\lVert\boldsymbol{x}\rVert$	Euclidean norm
$\mathrm{tr}[\boldsymbol{X}]$	matrix trace
$\boldsymbol{X}^{-1}$	matrix inverse
$\boldsymbol{x}^{\mathrm{T}}$, $\boldsymbol{X}^{\mathrm{T}}$	transpose
$\boldsymbol{X} \otimes \boldsymbol{Y}$	Kronecker product
$\mathrm{cov}(\mathbf{x}, \mathbf{y})$	cross-covariance matrix of $\mathbf{x}$ and $\mathbf{y}$
$\mathrm{var}(\mathbf{x})$	variance of $\mathbf{x}$
$\mathbf{I}_N$	identity matrix of size $N \times N$
$\mathbf{1}_N$	all-ones vector of size N
$\boldsymbol{x}_{1:N}$	column vector stacking vectors $\boldsymbol{x}_n$ for $n = 1, \ldots, N$, i.e., $(\boldsymbol{x}_1^{\mathrm{T}}, \ldots, \boldsymbol{x}_N^{\mathrm{T}})^{\mathrm{T}}$
$\boldsymbol{x}_{\neg n}$	column vector stacking vectors $\boldsymbol{x}_{n'}$ for $n' = 1, \ldots, n-1, n+1, \ldots, N$, i.e., $(\boldsymbol{x}_1^{\mathrm{T}}, \ldots, \boldsymbol{x}_{n-1}^{\mathrm{T}}, \boldsymbol{x}_{n+1}^{\mathrm{T}}, \ldots, \boldsymbol{x}_N^{\mathrm{T}})^{\mathrm{T}}$
$\bar{\boldsymbol{x}}_{1:N}$	sample mean of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$, i.e., $\frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n$
$f_{\mathsf{x}}(x)$, $f_{\mathbf{x}}(\boldsymbol{x})$	probability density function (pdf)
$p_{\mathsf{x}}(x)$, $p_{\mathbf{x}}(\boldsymbol{x})$	probability mass function (pmf)
$f_{\mathsf{x}\,\vert\,\mathsf{y}}(x \,\vert\, y)$, $f_{\mathbf{x}\,\vert\,\mathbf{y}}(\boldsymbol{x} \,\vert\, \boldsymbol{y})$	conditional pdf
$p_{\mathsf{x}\,\vert\,\mathsf{y}}(x \,\vert\, y)$, $p_{\mathbf{x}\,\vert\,\mathbf{y}}(\boldsymbol{x} \,\vert\, \boldsymbol{y})$	conditional pmf
$\mathsf{f}_{\mathrm{DP}}(\boldsymbol{x})$	random pdf
$\mathcal{N}(\,\cdot\,; \mu, \sigma^2)$	Gaussian pdf with mean μ and variance σ
$\mathcal{N}(\,\cdot\,; \boldsymbol{\mu}, \boldsymbol{\Sigma})$	multivariate Gaussian pdf with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
$\mathrm{DP}(\alpha, f_{\mathrm{H}})$	Dirichlet process with concentration parameter α and base pdf f_{H}
$\mathrm{GEM}(\,\cdot\,; \alpha)$	GEM distribution with parameter α
$\mathcal{U}(\,\cdot\,; a, b)$	continuous uniform distribution on $[a, b]$
$\delta(\,\cdot\,)$	Dirac delta function
$\mathbb{1}(\,\cdot\,)$	indicator function
$(f * g)(\,\cdot\,)$	convolution
$\mathbb{E}[\,\cdot\,]$	expectation
$\log(\,\cdot\,)$	natural logarithm
$\exp(\,\cdot\,)$	exponential function
$x^{\overline{n}}$	raising factorial (Pochhammer symbol)
$\mathbb{R}$	real numbers
$\mathbb{N}$	natural numbers

2 Bayesian Estimation

In this chapter, we introduce the basics of Bayesian inference. Our presentation is based on [1].

2.1 Fundamentals of Bayesian Estimation

We use observed data or measurements to infer an unknown quantity or parameter of interest. As opposed to the classical or frequentist approach, in which the parameter of interest is modeled as deterministic but unknown, both the parameter of interest and the measurement are modeled as continuous random variables, i.e., we assign to them a probability density function (pdf). This approach allows us to incorporate prior knowledge about the parameter of interest into the estimation. Thus, we consider the following:

- The random parameter of interest $\mathsf{x} \in \mathbb{R}$, with the pdf $f_{\mathsf{x}}(x)$, or random parameter vector $\mathbf{x} = (\mathsf{x}_1, \dots, \mathsf{x}_D)^{\mathrm{T}} \in \mathbb{R}^D$, with the pdf $f_{\mathbf{x}}(\boldsymbol{x})$, which is called the prior pdf.
- The random data or measurement $\mathsf{y} \in \mathbb{R}$, with the pdf $f_{\mathsf{y}}(y)$, or random measurement vector $\mathbf{y} = (\mathsf{y}_1, \dots, \mathsf{y}_N)^{\mathrm{T}} \in \mathbb{R}^N$, with the pdf $f_{\mathbf{y}}(\boldsymbol{y})$, which is called the evidence.
- The likelihood function $f_{\mathbf{y}|\mathbf{x}}(\boldsymbol{y} \,|\, \boldsymbol{x})$, which describes the statistical dependence of the measurement $\mathbf{y}$ given the parameter vector $\mathbf{x} = \boldsymbol{x}$.
- The estimator $\hat{\boldsymbol{x}}(\boldsymbol{y})$ is a function that assigns a value ("estimate") $\hat{\boldsymbol{x}} \in \mathbb{R}^D$ to each realization $\mathbf{y} = \boldsymbol{y}$. Here, $\hat{\boldsymbol{x}}$ is desired to be close to the true parameter $\mathbf{x}$. The estimate depends only on the observed data $\boldsymbol{y}$ and the statistical model. Because the estimate $\hat{\mathbf{x}}$ is a function of the random data $\mathbf{y}$, it is also random.
- The estimation error $\mathbf{e} = \hat{\mathbf{x}} - \mathbf{x}$ is the difference between the estimate $\hat{\mathbf{x}}$ and the true parameter of interest $\mathbf{x}$.

Our goal is to find an estimator that minimizes the estimation error $\mathbf{e}$, mapped to a nonnegative scalar value by a cost function $C(\mathbf{e})$ (sometimes also called loss function or objective function). We define the Bayes risk as

$$R \triangleq \mathbb{E}[C(\mathbf{e})] = \int_{\boldsymbol{y}} \int_{\boldsymbol{x}} C(\hat{\boldsymbol{x}}(\boldsymbol{y}) - \boldsymbol{x}) f_{\mathbf{y},\mathbf{x}}(\boldsymbol{y}, \boldsymbol{x}) d\boldsymbol{x} d\boldsymbol{y}, \tag{2.1}$$

which for fixed cost function $C(\boldsymbol{e})$ only depends on the estimator $\hat{\boldsymbol{x}}(\cdot)$. The Bayesian estimator $\hat{\boldsymbol{x}}_{\mathrm{B}}(\cdot)$, for the given cost function $C(\cdot)$, is now defined to minimize the Bayes risk R among all possible estimators, i.e.,

$$\hat{\boldsymbol{x}}_{\mathrm{B}}(\cdot) \triangleq \underset{\hat{\boldsymbol{x}}(\cdot)}{\arg\min}\, R = \underset{\hat{\boldsymbol{x}}(\cdot)}{\arg\min} \int_{\boldsymbol{y}} \int_{\boldsymbol{x}} C(\hat{\boldsymbol{x}}(\boldsymbol{y}) - \boldsymbol{x}) f_{\mathbf{y},\mathbf{x}}(\boldsymbol{y}, \boldsymbol{x}) d\boldsymbol{x} d\boldsymbol{y}. \tag{2.2}$$

Here, the joint pdf $f_{\mathbf{y},\mathbf{x}}(\boldsymbol{y}, \boldsymbol{x})$ can be obtained from the prior $f_{\mathbf{x}}(\boldsymbol{x})$ and the likelihood function $f_{\mathbf{y}\,|\,\mathbf{x}}(\boldsymbol{y}\,|\,\boldsymbol{x})$ according to

$$f_{\mathbf{y},\mathbf{x}}(\boldsymbol{y}, \boldsymbol{x}) = f_{\mathbf{y}\,|\,\mathbf{x}}(\boldsymbol{y}\,|\,\boldsymbol{x}) f_{\mathbf{x}}(\boldsymbol{x}). \tag{2.3}$$

An alternative factorization is

$$f_{\mathbf{y},\mathbf{x}}(\boldsymbol{y}, \boldsymbol{x}) = f_{\mathbf{x}\,|\,\mathbf{y}}(\boldsymbol{x}\,|\,\boldsymbol{y}) f_{\mathbf{y}}(\boldsymbol{y}). \tag{2.4}$$

2.2 The MMSE Estimator

We now derive the most commonly used Bayesian estimator, the minimum mean square error (MMSE) estimator, following the definitions in [1]. First we specify the cost function as $C(\boldsymbol{e}) = \frac{1}{D}\|\boldsymbol{e}\|^2$, i.e., the risk R becomes the mean square error (MSE),

$$R = \mathbb{E}[C(\mathbf{e})] = \frac{1}{D}\mathbb{E}[\|\mathbf{e}\|^2] = \frac{1}{D}\mathbb{E}[\|\hat{\mathbf{x}} - \mathbf{x}\|^2] = \mathrm{MSE}. \tag{2.5}$$

Using (2.4) in (2.2) and rearranging the terms, we obtain

$$\hat{\boldsymbol{x}}_{\mathrm{B}}(\cdot) = \underset{\hat{\boldsymbol{x}}(\cdot)}{\arg\min} \int_{\boldsymbol{y}} \left[\int_{\boldsymbol{x}} \|\hat{\boldsymbol{x}}(\boldsymbol{y}) - \boldsymbol{x}\|^2 f_{\mathbf{x}\,|\,\mathbf{y}}(\boldsymbol{x}\,|\,\boldsymbol{y}) d\boldsymbol{x} \right] f_{\mathbf{y}}(\boldsymbol{y}) d\boldsymbol{y}, \tag{2.6}$$

where we note that $f_{\mathbf{y}}(\boldsymbol{y}) \geq 0$. This means that if we minimize the term in brackets (conditional risk), the entire expression (2.6) will be minimized. Therefore for fixed $\mathbf{y} = \boldsymbol{y}$ we need to minimize the conditional risk with respect to the vector $\hat{\mathbf{x}} = \hat{\boldsymbol{x}}$ as opposed to the function $\hat{\boldsymbol{x}}(\cdot)$. To this end, we calculate[1] the gradient of the conditional risk with respect to $\hat{\boldsymbol{x}}$, i.e.,

$$\nabla_{\hat{\boldsymbol{x}}} \int_{\boldsymbol{x}} \|\hat{\boldsymbol{x}} - \boldsymbol{x}\|^2 f_{\mathbf{x}\,|\,\mathbf{y}}(\boldsymbol{x}\,|\,\boldsymbol{y}) d\boldsymbol{x} = \int_{\boldsymbol{x}} \nabla_{\hat{\boldsymbol{x}}} \|\hat{\boldsymbol{x}} - \boldsymbol{x}\|^2 f_{\mathbf{x}\,|\,\mathbf{y}}(\boldsymbol{x}\,|\,\boldsymbol{y}) d\boldsymbol{x}$$

[1] We can apply the Leibniz integral rule and interchange the order of integration and differentiation in (2.7) since both $\|\hat{\boldsymbol{x}} - \boldsymbol{x}\|^2 f_{\mathbf{x}\,|\,\mathbf{y}}(\boldsymbol{x}\,|\,\boldsymbol{y})$ and $\nabla_{\hat{\boldsymbol{x}}}\|\hat{\boldsymbol{x}} - \boldsymbol{x}\|^2 f_{\mathbf{x}\,|\,\mathbf{y}}(\boldsymbol{x}\,|\,\boldsymbol{y})$ are continuous in $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$.

$$
\begin{aligned}
&= \int_{\boldsymbol{x}} 2(\hat{\boldsymbol{x}} - \boldsymbol{x}) f_{\mathbf{x}\,|\,\mathbf{y}}(\boldsymbol{x}\,|\,\boldsymbol{y}) d\boldsymbol{x} \\
&= 2\hat{\boldsymbol{x}} \int_{\boldsymbol{x}} f_{\mathbf{x}\,|\,\mathbf{y}}(\boldsymbol{x}\,|\,\boldsymbol{y}) d\boldsymbol{x} - 2 \int_{\boldsymbol{x}} \boldsymbol{x} f_{\mathbf{x}\,|\,\mathbf{y}}(\boldsymbol{x}\,|\,\boldsymbol{y}) d\boldsymbol{x} \\
&= 2\left(\hat{\boldsymbol{x}} - \mathbb{E}[\mathbf{x}\,|\,\mathbf{y} = \boldsymbol{y}]\right), \qquad (2.7)
\end{aligned}
$$

where we used $\int_{\boldsymbol{x}} f_{\mathbf{x}\,|\,\mathbf{y}}(\boldsymbol{x}\,|\,\boldsymbol{y}) d\boldsymbol{x} = 1$. Setting the gradient in (2.7) equal to the zero vector yields the critical point of the function, i.e.,

$$\hat{\boldsymbol{x}} = \mathbb{E}[\mathbf{x}\,|\,\mathbf{y} = \boldsymbol{y}]. \qquad (2.8)$$

In order to prove that this point is a minimum we consider the Hessian matrix $\mathbf{H}$, i.e., the matrix of second order partial derivatives. It can be shown that the Hessian matrix is equal to

$$\mathbf{H} = 2\mathbf{I}_D, \qquad (2.9)$$

where $\mathbf{I}_D$ is the identity matrix of size $D \times D$. Since $\mathbf{H}$ is positive definite, the critical point $\hat{\boldsymbol{x}}$ is a minimum. Therefore, the MMSE estimator is finally seen to be

$$\hat{\boldsymbol{x}}_{\mathrm{MMSE}}(\boldsymbol{y}) = \mathbb{E}[\mathbf{x}\,|\,\mathbf{y} = \boldsymbol{y}] = \int_{\boldsymbol{x}} \boldsymbol{x} f_{\mathbf{x}\,|\,\mathbf{y}}(\boldsymbol{x}\,|\,\boldsymbol{y}) d\boldsymbol{x}, \qquad (2.10)$$

which is the posterior mean, i.e., the mean of the posterior pdf $f_{\mathbf{x}\,|\,\mathbf{y}}(\boldsymbol{x}\,|\,\boldsymbol{y})$. The posterior pdf $f_{\mathbf{x}\,|\,\mathbf{y}}(\boldsymbol{x}\,|\,\boldsymbol{y})$, which is the pdf of the parameter $\mathbf{x}$ *after* the data $\mathbf{y} = \boldsymbol{y}$ has been observed, can be obtained by using Bayes' theorem,

$$f_{\mathbf{x}\,|\,\mathbf{y}}(\boldsymbol{x}\,|\,\boldsymbol{y}) = \frac{f_{\mathbf{y}\,|\,\mathbf{x}}(\boldsymbol{y}\,|\,\boldsymbol{x}) f_{\mathbf{x}}(\boldsymbol{x})}{f_{\mathbf{y}}(\boldsymbol{y})}, \qquad (2.11)$$

where the prior pdf $f_{\mathbf{x}}(\boldsymbol{x})$ represents our knowledge of $\mathbf{x}$ *before* any data has been observed.

Before we discuss the choice of the prior pdf, we derive an expression for the MSE achieved by the MMSE estimator $\hat{\boldsymbol{x}}_{\mathrm{MMSE}}(\boldsymbol{y})$. Inserting (2.10) in (2.5) and briefly writing the posterior mean $\mathbb{E}[\mathbf{x}\,|\,\mathbf{y} = \boldsymbol{y}]$ as $\boldsymbol{\mu}_{\boldsymbol{x}\,|\,\boldsymbol{y}}$, we obtain

$$\mathrm{MSE}_{\min} = \frac{1}{D}\mathbb{E}[\|\hat{\mathbf{x}}_{\mathrm{MMSE}} - \mathbf{x}\|^2]$$

$$
\begin{aligned}
&= \frac{1}{D}\int_{\boldsymbol{y}}\int_{\boldsymbol{x}} \|\hat{\boldsymbol{x}}_{\mathrm{MMSE}}(\boldsymbol{y}) - \boldsymbol{x}\|^2 f_{\mathbf{y},\mathbf{x}}(\boldsymbol{y},\boldsymbol{x})d\boldsymbol{x}d\boldsymbol{y} \\
&= \frac{1}{D}\int_{\boldsymbol{y}}\left(\int_{\boldsymbol{x}} \left(\boldsymbol{x}-\boldsymbol{\mu}_{\boldsymbol{x}\,|\,\boldsymbol{y}}\right)^{\mathrm{T}}\left(\boldsymbol{x}-\boldsymbol{\mu}_{\boldsymbol{x}\,|\,\boldsymbol{y}}\right) f_{\mathbf{x}\,|\,\mathbf{y}}(\boldsymbol{x}\,|\,\boldsymbol{y})d\boldsymbol{x}\right) f_{\mathbf{y}}(\boldsymbol{y})d\boldsymbol{y} \\
&= \frac{1}{D}\int_{\boldsymbol{y}} \mathrm{tr}\left[\int_{\boldsymbol{x}} \left(\boldsymbol{x}-\boldsymbol{\mu}_{\boldsymbol{x}\,|\,\boldsymbol{y}}\right)\left(\boldsymbol{x}-\boldsymbol{\mu}_{\boldsymbol{x}\,|\,\boldsymbol{y}}\right)^{\mathrm{T}} f_{\mathbf{x}\,|\,\mathbf{y}}(\boldsymbol{x}\,|\,\boldsymbol{y})d\boldsymbol{x}\right] f_{\mathbf{y}}(\boldsymbol{y})d\boldsymbol{y} \\
&= \frac{1}{D}\int_{\boldsymbol{y}} \mathrm{tr}\left[\mathrm{cov}\left(\mathbf{x}\,|\,\mathbf{y}=\boldsymbol{y}\right)\right] f_{\mathbf{y}}(\boldsymbol{y})d\boldsymbol{y}, \qquad (2.12)
\end{aligned}
$$

where we used (2.4) and the identity $\boldsymbol{a}^{\mathrm{T}}\boldsymbol{b} = \mathrm{tr}\left[\boldsymbol{a}\boldsymbol{b}^{\mathrm{T}}\right]$. Thus, the MSE achieved by the MMSE estimator is the trace of the posterior covariance matrix averaged over the distribution $f_{\mathbf{y}}(\boldsymbol{y})$ of the measurements.

2.3 The Gaussian Distribution

We now discuss one of the most commonly chosen prior pdfs and likelihood functions, the Gaussian pdf (also called normal distribution). For a scalar random variable $\mathsf{z} \in \mathbb{R}$, the Gaussian pdf is defined as

$$
f_{\mathsf{z}}(z) = \frac{1}{\sqrt{2\pi\sigma_z^2}}\exp\left(-\frac{1}{2}\left(\frac{z-\mu_z}{\sigma_z}\right)^2\right), \qquad (2.13)
$$

where μ_z is the mean and σ_z^2 is the variance. In what follows, we will denote the Gaussian pdf as $\mathcal{N}(z;\mu_z,\sigma_z^2)$. In case of a D-dimensional random vector $\mathbf{z} \in \mathbb{R}^D$, the Gaussian pdf is defined as

$$
f_{\mathbf{z}}(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z};\boldsymbol{\mu}_{\boldsymbol{z}},\boldsymbol{\Sigma}_{\boldsymbol{z}}) = \frac{1}{\sqrt{(2\pi)^D \det(\boldsymbol{\Sigma}_{\boldsymbol{z}})}}\exp\left(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu}_{\boldsymbol{z}})^{\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{z}}^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_{\boldsymbol{z}})\right), \qquad (2.14)
$$

where $\boldsymbol{\Sigma}_{\boldsymbol{z}}$ is the covariance matrix. In subsequent calculations, it will sometimes be advantageous to use the precision matrix $\boldsymbol{\Lambda}_{\boldsymbol{z}} = \boldsymbol{\Sigma}_{\boldsymbol{z}}^{-1}$, i.e., the quadratic form in the exponent of (2.14) can be written as

$$
\begin{aligned}
(\boldsymbol{z}-\boldsymbol{\mu}_{\boldsymbol{z}})^{\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{z}}^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_{\boldsymbol{z}}) &= (\boldsymbol{z}-\boldsymbol{\mu}_{\boldsymbol{z}})^{\mathrm{T}}\boldsymbol{\Lambda}_{\boldsymbol{z}}(\boldsymbol{z}-\boldsymbol{\mu}_{\boldsymbol{z}}) \qquad (2.15) \\
&= \boldsymbol{z}^{\mathrm{T}}\boldsymbol{\Lambda}_{\boldsymbol{z}}\boldsymbol{z} - \boldsymbol{z}^{\mathrm{T}}\boldsymbol{\Lambda}_{\boldsymbol{z}}\boldsymbol{\mu}_{\boldsymbol{z}} - \boldsymbol{\mu}_{\boldsymbol{z}}^{\mathrm{T}}\boldsymbol{\Lambda}_{\boldsymbol{z}}\boldsymbol{z} + \boldsymbol{\mu}_{\boldsymbol{z}}^{\mathrm{T}}\boldsymbol{\Lambda}_{\boldsymbol{z}}\boldsymbol{\mu}_{\boldsymbol{z}} \qquad (2.16)
\end{aligned}
$$

2.3.1 Convolution Property

An important property of the Gaussian distribution is that the convolution of two Gaussian pdfs is also Gaussian [18, Sec. 7.14]. We consider two Gaussian pdfs, with generally different means and covariance matrices,

$$f_{\mathbf{x}}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x), \quad f_{\mathbf{y}}(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y). \tag{2.17}$$

The convolution of $f_{\mathbf{y}}(\boldsymbol{y})$ and $f_{\mathbf{x}}(\boldsymbol{x})$ is defined as

$$(f_{\mathbf{y}} * f_{\mathbf{x}})(\boldsymbol{z}) = \int_{\boldsymbol{x}} f_{\mathbf{y}}(\boldsymbol{z}-\boldsymbol{x}) f_{\mathbf{x}}(\boldsymbol{x}) d\boldsymbol{x} = \int_{\boldsymbol{x}} \mathcal{N}(\boldsymbol{z}-\boldsymbol{x}; \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x) d\boldsymbol{x}. \tag{2.18}$$

As shown in [18, Eq. 7.49], two Gaussians convolve to make another Gaussian, the means $\boldsymbol{\mu}$ and covariance matrices $\boldsymbol{\Sigma}$ being additive i.e.,

$$\int_{\boldsymbol{x}} \mathcal{N}(\boldsymbol{z}-\boldsymbol{x}; \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x) d\boldsymbol{x} = \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_x + \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_y). \tag{2.19}$$

Next, we consider two independent variables $\mathbf{x}$ and $\mathbf{y}$, distributed according to (2.17). We are interested in the distribution of the sum of $\mathbf{x}$ and $\mathbf{y}$, we define $\mathbf{z}$ as

$$\mathbf{z} \triangleq \mathbf{x} + \mathbf{y}. \tag{2.20}$$

The convolution of two, independent [2] random variables is equal to the pdf of the sum of the random variables [19, Eq. 6-43], i.e., we have

$$(f_{\mathbf{y}} * f_{\mathbf{x}})(\boldsymbol{z}) = f_{\mathbf{z}}(\boldsymbol{z}), \tag{2.21}$$

The mean of $\mathbf{z}$ is equal to the sum of the means of the Gaussian random variables $\mathbf{x}$ and $\mathbf{y}$, i.e.,

$$\boldsymbol{\mu}_z = \boldsymbol{\mu}_x + \boldsymbol{\mu}_y. \tag{2.22}$$

Since we assume $\mathbf{x}$ and $\mathbf{y}$ are independent, the covariance matrix of $\mathbf{z}$ can be given by the

[2] If $\mathbf{x}$ and $\mathbf{y}$ are not independent then $f_{\mathbf{z}}(z) = \int_{\boldsymbol{x}} f_{\mathbf{x},\mathbf{y}}(\boldsymbol{z}-\boldsymbol{x}, \boldsymbol{x}) d\boldsymbol{x}$

sum of the respective covariance matrices, i.e.,

$$\boldsymbol{\Sigma}_{\boldsymbol{z}} = \boldsymbol{\Sigma}_{\boldsymbol{x}} + \boldsymbol{\Sigma}_{\boldsymbol{y}}. \tag{2.23}$$

Using (2.22) and (2.23) we can write the pdf of $\mathbf{z}$ as

$$f_{\mathbf{z}}(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_{\boldsymbol{z}}, \boldsymbol{\Sigma}_{\boldsymbol{z}}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_{\boldsymbol{x}} + \boldsymbol{\mu}_{\boldsymbol{y}}, \boldsymbol{\Sigma}_{\boldsymbol{x}} + \boldsymbol{\Sigma}_{\boldsymbol{y}}). \tag{2.24}$$

2.3.2 Posterior pdf for Jointly Gaussian Measurement and Parameter

The choice of a prior is a critical part of Bayesian estimation, as shall be illustrated in the following example. In the expression (2.10) for the MMSE estimator, it may be impossible to calculate the posterior pdf $f_{\mathbf{x}\,|\,\mathbf{y}}(\boldsymbol{x}\,|\,\boldsymbol{y})$ in closed form, or we may have to compute the integral using numerical methods. From (2.11), we conclude that the ability to calculate the posterior pdf depends on the prior pdf $f_{\mathbf{x}}(\boldsymbol{x})$, the likelihood function $f_{\mathbf{y}\,|\,\mathbf{x}}(\boldsymbol{y}\,|\,\boldsymbol{x})$, and the evidence $f_{\mathbf{y}}(\boldsymbol{y}) = \int_{\boldsymbol{x}} f_{\mathbf{y}\,|\,\mathbf{x}}(\boldsymbol{y}\,|\,\boldsymbol{x}) f_{\mathbf{x}}(\boldsymbol{x}) d\boldsymbol{x}$.

We now derive the posterior pdf $f_{\mathbf{x}\,|\,\mathbf{y}}(\boldsymbol{x}\,|\,\boldsymbol{y})$ under the assumption that the parameter $\mathbf{x}$ and the data $\mathbf{y}$ are jointly Gaussian. Our development is based on [20]. We consider the stacked Gaussian random vector

$$\mathbf{z} \triangleq \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}. \tag{2.25}$$

The mean $\boldsymbol{\mu}_{\boldsymbol{z}}$ and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{z}}$ of $\mathbf{z}$ can be partitioned according to

$$\boldsymbol{\mu}_{\boldsymbol{z}} = \begin{pmatrix} \boldsymbol{\mu}_{\boldsymbol{x}} \\ \boldsymbol{\mu}_{\boldsymbol{y}} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma}_{\boldsymbol{z}} = \begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{xx}} & \boldsymbol{\Sigma}_{\boldsymbol{xy}} \\ \boldsymbol{\Sigma}_{\boldsymbol{yx}} & \boldsymbol{\Sigma}_{\boldsymbol{yy}} \end{pmatrix}. \tag{2.26}$$

We note that $\boldsymbol{\Sigma}_{\boldsymbol{z}}$ is a positive definite, symmetric matrix, which implies that both $\boldsymbol{\Sigma}_{\boldsymbol{xx}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{yy}}$ are positive definite symmetric matrices and $\boldsymbol{\Sigma}_{\boldsymbol{yx}} = \boldsymbol{\Sigma}_{\boldsymbol{xy}}^{\mathrm{T}}$. The precision matrix $\boldsymbol{\Lambda}_{\boldsymbol{z}} = \boldsymbol{\Sigma}_{\boldsymbol{z}}^{-1}$ can be partitioned similarly to $\boldsymbol{\Sigma}_{\boldsymbol{z}}$, i.e.,

$$\boldsymbol{\Lambda}_{\boldsymbol{z}} = \begin{pmatrix} \boldsymbol{\Lambda}_{\boldsymbol{xx}} & \boldsymbol{\Lambda}_{\boldsymbol{xy}} \\ \boldsymbol{\Lambda}_{\boldsymbol{yx}} & \boldsymbol{\Lambda}_{\boldsymbol{yy}} \end{pmatrix}. \tag{2.27}$$

As the inverse of the positive definite symmetric matrix $\boldsymbol{\Sigma}_{\boldsymbol{z}}$, the precision matrix $\boldsymbol{\Lambda}_{\boldsymbol{z}}$ is also positive definite symmetric, which implies that $\boldsymbol{\Lambda}_{\boldsymbol{xx}}$ and $\boldsymbol{\Lambda}_{\boldsymbol{yy}}$ are positive definite symmetric

matrices and $\boldsymbol{\Lambda}_{\boldsymbol{yx}} = \boldsymbol{\Lambda}_{\boldsymbol{xy}}^{\mathrm{T}}$.

The posterior pdf $f_{\mathbf{x}\,|\,\mathbf{y}}(\boldsymbol{x}\,|\,\boldsymbol{y})$ can be obtained from the joint distribution $f_{\mathbf{x},\mathbf{y}}(\boldsymbol{x},\boldsymbol{y})$ by fixing $\boldsymbol{y}$ and normalizing the expression to a valid pdf, i.e., by using Bayes' theorem (2.11) and the joint distribution $f_{\mathbf{x},\mathbf{y}}(\boldsymbol{x},\boldsymbol{y}) = f_{\mathbf{y}\,|\,\mathbf{x}}(\boldsymbol{y}\,|\,\boldsymbol{x})f_{\mathbf{x}}(\boldsymbol{x})$ we have

$$f_{\mathbf{x}\,|\,\mathbf{y}}(\boldsymbol{x}\,|\,\boldsymbol{y}) = \frac{f_{\mathbf{x},\mathbf{y}}(\boldsymbol{x},\boldsymbol{y})}{f_{\mathbf{y}}(\boldsymbol{y})} = \alpha f_{\mathbf{x},\mathbf{y}}(\boldsymbol{x},\boldsymbol{y}), \tag{2.28}$$

with normalizing factor α so that $\int_{\boldsymbol{x}} f_{\mathbf{x}\,|\,\mathbf{y}}(\boldsymbol{x}\,|\,\boldsymbol{y})d\boldsymbol{x} = 1$. Since we assume that $\mathbf{x}$ and $\mathbf{y}$ are jointly Gaussian, using the stacked Gaussian random vector $\mathbf{z}$ (see (2.25)), the joint pdf $f_{\mathbf{x},\mathbf{y}}(\boldsymbol{x},\boldsymbol{y})$ is given by (2.14), therefore we have

$$f_{\mathbf{x}\,|\,\mathbf{y}}(\boldsymbol{x}\,|\,\boldsymbol{y}) = \alpha f_{\mathbf{z}}(\boldsymbol{z}), \tag{2.29}$$

which according to (2.14) is Gaussian. If $\mathbf{x}$ and $\mathbf{y}$ are jointly Gaussian, then the posterior distribution $f_{\mathbf{x}\,|\,\mathbf{y}}(\boldsymbol{x}\,|\,\boldsymbol{y})$ is also Gaussian, i.e.,

$$f_{\mathbf{x}\,|\,\mathbf{y}}(\boldsymbol{x}\,|\,\boldsymbol{y}) = \mathcal{N}(\boldsymbol{x};\boldsymbol{\mu}_{\boldsymbol{x}\,|\,\boldsymbol{y}},\boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y}}), \tag{2.30}$$

with some posterior mean $\boldsymbol{\mu}_{\boldsymbol{x}\,|\,\boldsymbol{y}}$ and posterior covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y}}$.

We now take up the joint pdf $f_{\mathbf{x},\mathbf{y}}(\boldsymbol{x},\boldsymbol{y})$ given by (2.14). First, we consider the quadratic form in the exponent of (2.14), given in (2.15). Using (2.26) and (2.27), the quadratic form in (2.15) becomes

$$\begin{aligned}(\boldsymbol{z}-\boldsymbol{\mu}_{\boldsymbol{z}})^{\mathrm{T}}\boldsymbol{\Lambda}_{\boldsymbol{z}}(\boldsymbol{z}-\boldsymbol{\mu}_{\boldsymbol{z}}) = {}& (\boldsymbol{x}-\boldsymbol{\mu}_{\boldsymbol{x}})^{\mathrm{T}}\boldsymbol{\Lambda}_{\boldsymbol{xx}}(\boldsymbol{x}-\boldsymbol{\mu}_{\boldsymbol{x}}) + (\boldsymbol{x}-\boldsymbol{\mu}_{\boldsymbol{x}})^{\mathrm{T}}\boldsymbol{\Lambda}_{\boldsymbol{xy}}(\boldsymbol{y}-\boldsymbol{\mu}_{\boldsymbol{y}}) \\ & + (\boldsymbol{y}-\boldsymbol{\mu}_{\boldsymbol{y}})^{\mathrm{T}}\boldsymbol{\Lambda}_{\boldsymbol{yx}}(\boldsymbol{x}-\boldsymbol{\mu}_{\boldsymbol{x}}) + (\boldsymbol{y}-\boldsymbol{\mu}_{\boldsymbol{y}})^{\mathrm{T}}\boldsymbol{\Lambda}_{\boldsymbol{yy}}(\boldsymbol{y}-\boldsymbol{\mu}_{\boldsymbol{y}}).\end{aligned} \tag{2.31}$$

We now view the expression in (2.31), i.e., the exponent of the joint Gaussian pdf $f_{\mathbf{x},\mathbf{y}}(\boldsymbol{x},\boldsymbol{y})$ (up to the factor of $-\frac{1}{2}$), as a function of $\boldsymbol{x}$ and consider $\boldsymbol{y}$ fixed, i.e.,

$$\begin{aligned}(\boldsymbol{z}-\boldsymbol{\mu}_{\boldsymbol{z}})^{\mathrm{T}}\boldsymbol{\Lambda}_{\boldsymbol{z}}(\boldsymbol{z}-\boldsymbol{\mu}_{\boldsymbol{z}}) &= \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Lambda}_{\boldsymbol{xx}}\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{x}}^{\mathrm{T}}\boldsymbol{\Lambda}_{\boldsymbol{xx}}\boldsymbol{x} - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Lambda}_{\boldsymbol{xx}}\boldsymbol{\mu}_{\boldsymbol{x}} + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Lambda}_{\boldsymbol{xy}}\boldsymbol{y} - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Lambda}_{\boldsymbol{xy}}\boldsymbol{\mu}_{\boldsymbol{y}} \\ &\quad + \underbrace{\boldsymbol{y}^{\mathrm{T}}\boldsymbol{\Lambda}_{\boldsymbol{yx}}\boldsymbol{x}}_{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Lambda}_{\boldsymbol{yx}}^{\mathrm{T}}\boldsymbol{y}} - \underbrace{\boldsymbol{\mu}_{\boldsymbol{y}}^{\mathrm{T}}\boldsymbol{\Lambda}_{\boldsymbol{yx}}\boldsymbol{x}}_{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Lambda}_{\boldsymbol{yx}}^{\mathrm{T}}\boldsymbol{\mu}_{\boldsymbol{y}}}} + \text{const.} \\ &= \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Lambda}_{\boldsymbol{xx}}\boldsymbol{x} - 2\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Lambda}_{\boldsymbol{xx}}\boldsymbol{\mu}_{\boldsymbol{x}} - 2\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Lambda}_{\boldsymbol{xy}}\boldsymbol{\mu}_{\boldsymbol{y}} + 2\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Lambda}_{\boldsymbol{xy}}\boldsymbol{y} + \text{const.}\end{aligned}$$

$$= \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Lambda}_{\boldsymbol{xx}}\boldsymbol{x} - 2\boldsymbol{x}^{\mathrm{T}}\left(\boldsymbol{\Lambda}_{\boldsymbol{xx}}\boldsymbol{\mu}_{\boldsymbol{x}} - \boldsymbol{\Lambda}_{\boldsymbol{xy}}\left(\boldsymbol{y} - \boldsymbol{\mu}_{\boldsymbol{y}}\right)\right) + \text{const.} \tag{2.32}$$

The right hand side of (2.32) is recognized as a quadratic form in $\boldsymbol{x}$, which can be written as

$$(\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{x}\,|\,\boldsymbol{y}})^{\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y}}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{x}\,|\,\boldsymbol{y}}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y}}^{-1}\boldsymbol{x} - 2\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y}}^{-1}\boldsymbol{\mu}_{\boldsymbol{x}\,|\,\boldsymbol{y}} + \text{const.} \tag{2.33}$$

Comparing the quadratic terms in $\boldsymbol{x}$ occurring in (2.32) and (2.33), we obtain the posterior covariance matrix as

$$\boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y}} = \boldsymbol{\Lambda}_{\boldsymbol{xx}}^{-1}. \tag{2.34}$$

Furthermore, by comparing the linear terms in $\boldsymbol{x}$ occurring in (2.32) and (2.33), we obtain for the posterior mean $\boldsymbol{\mu}_{\boldsymbol{x}\,|\,\boldsymbol{y}}$

$$\boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y}}^{-1}\boldsymbol{\mu}_{\boldsymbol{x}\,|\,\boldsymbol{y}} = \boldsymbol{\Lambda}_{\boldsymbol{xx}}\boldsymbol{\mu}_{\boldsymbol{x}} - \boldsymbol{\Lambda}_{\boldsymbol{xy}}(\boldsymbol{y} - \boldsymbol{\mu}_{\boldsymbol{y}}), \tag{2.35}$$

and therefore

$$\begin{aligned}\boldsymbol{\mu}_{\boldsymbol{x}\,|\,\boldsymbol{y}} &= \boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y}}\left(\boldsymbol{\Lambda}_{\boldsymbol{xx}}\boldsymbol{\mu}_{\boldsymbol{x}} - \boldsymbol{\Lambda}_{\boldsymbol{xy}}(\boldsymbol{y} - \boldsymbol{\mu}_{\boldsymbol{y}})\right) \\ &= \boldsymbol{\mu}_{\boldsymbol{x}} - \boldsymbol{\Lambda}_{\boldsymbol{xx}}^{-1}\boldsymbol{\Lambda}_{\boldsymbol{xy}}(\boldsymbol{y} - \boldsymbol{\mu}_{\boldsymbol{y}}),\end{aligned} \tag{2.36}$$

where (2.34) was used. The posterior covariance $\boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y}}$ in (2.34) and the posterior mean in $\boldsymbol{\mu}_{\boldsymbol{x}\,|\,\boldsymbol{y}}$ (2.36) are expressed in terms of the partitioned precision matrix $\boldsymbol{\Lambda}_{\boldsymbol{z}}$ in (2.27); however, using linear algebra it is possible to express them in terms of the partitioned covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{z}}$ in (2.26). We have $\boldsymbol{\Sigma}_{\boldsymbol{z}}^{-1} = \boldsymbol{\Lambda}_{\boldsymbol{z}}$ or equivalently

$$\begin{pmatrix}\boldsymbol{\Sigma}_{\boldsymbol{xx}} & \boldsymbol{\Sigma}_{\boldsymbol{xy}} \\ \boldsymbol{\Sigma}_{\boldsymbol{yx}} & \boldsymbol{\Sigma}_{\boldsymbol{yy}}\end{pmatrix}^{-1} = \begin{pmatrix}\boldsymbol{\Lambda}_{\boldsymbol{xx}} & \boldsymbol{\Lambda}_{\boldsymbol{xy}} \\ \boldsymbol{\Lambda}_{\boldsymbol{yx}} & \boldsymbol{\Lambda}_{\boldsymbol{yy}}\end{pmatrix}. \tag{2.37}$$

One can show [20, Eq. 2.79] that,

$$\boldsymbol{\Lambda}_{\boldsymbol{xx}} = \left(\boldsymbol{\Sigma}_{\boldsymbol{xx}} - \boldsymbol{\Sigma}_{\boldsymbol{xy}}\boldsymbol{\Sigma}_{\boldsymbol{yy}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{yx}}\right)^{-1}, \tag{2.38}$$

Furthermore, using [20, Eq. 2.80] we have

$$\boldsymbol{\Lambda}_{\boldsymbol{xy}} = -\left(\boldsymbol{\Sigma}_{\boldsymbol{xx}} - \boldsymbol{\Sigma}_{\boldsymbol{xy}}\boldsymbol{\Sigma}_{\boldsymbol{yy}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{yx}}\right)^{-1}\boldsymbol{\Sigma}_{\boldsymbol{xy}}\boldsymbol{\Sigma}_{\boldsymbol{yy}}^{-1}. \tag{2.39}$$

Using (2.38) in (2.34) yields for the posterior covariance matrix

$$\boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y}} = \boldsymbol{\Sigma}_{\boldsymbol{xx}} - \boldsymbol{\Sigma}_{\boldsymbol{xy}}\boldsymbol{\Sigma}_{\boldsymbol{yy}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{yx}}. \tag{2.40}$$

Furthermore, using (2.38) together with (2.39) in (2.36) yields for the posterior mean

$$\boldsymbol{\mu}_{\boldsymbol{x}\,|\,\boldsymbol{y}} = \boldsymbol{\mu}_{\boldsymbol{x}} + \boldsymbol{\Sigma}_{\boldsymbol{xy}}\boldsymbol{\Sigma}_{\boldsymbol{yy}}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_{\boldsymbol{y}}). \tag{2.41}$$

To summarize, if $\mathbf{x}$ and $\mathbf{y}$ are jointly Gaussian, then the posterior distribution $f_{\mathbf{x}\,|\,\mathbf{y}}(\boldsymbol{x}\,|\,\boldsymbol{y})$ is also Gaussian, i.e.,

$$f_{\mathbf{x}\,|\,\mathbf{y}}(\boldsymbol{x}\,|\,\boldsymbol{y}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_{\boldsymbol{x}\,|\,\boldsymbol{y}}, \boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y}}), \tag{2.42}$$

with posterior mean $\boldsymbol{\mu}_{\boldsymbol{x}\,|\,\boldsymbol{y}}$ given by (2.41) and posterior covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y}}$ given by (2.40).

2.3.3 Posterior pdf for Multiple Jointly Gaussian Measurements and Parameter

We now consider multiple measurements $\mathbf{y}_n \in \mathbb{R}^D, n = 1, \ldots, N$, that are drawn independently from identical Gaussian distribution with mean $\mathbf{x} \in \mathbb{R}^D$ and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{y}}$, i.e., the likelihood function is given by

$$f_{\mathbf{y}_n\,|\,\mathbf{x}}(\boldsymbol{y}_n\,|\,\boldsymbol{x}) = \mathcal{N}(\boldsymbol{y}_n; \boldsymbol{x}, \boldsymbol{\Sigma}_{\boldsymbol{y}}) \quad \text{for} \quad n = 1, \ldots, N. \tag{2.43}$$

Let $\mathbf{y}_{1:N} = (\mathbf{y}_1^{\mathrm{T}}, \ldots, \mathbf{y}_N^{\mathrm{T}})^{\mathrm{T}}$ denote the stacked vector of measurements. Furthermore, the prior pdf is also Gaussian, with mean $\boldsymbol{\mu}_{\boldsymbol{x}}$ and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}}$, i.e.,

$$f_{\mathbf{x}}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_{\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{x}}). \tag{2.44}$$

We are interested in the posterior pdf $f_{\mathbf{x}\,|\,\mathbf{y}_{1:N}}(\boldsymbol{x}\,|\,\boldsymbol{y}_{1:N})$, i.e., the pdf of $\mathbf{x}$ conditioned on the measurements $\mathbf{y}_{1:N}$. First, since we assumed the observations to be independent and

identically distributed (i.i.d.), using (2.43) we have

$$f_{\mathbf{y}_{1:N}|\mathbf{x}}(\boldsymbol{y}_{1:N} \,|\, \boldsymbol{x}) = \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{y}_n; \boldsymbol{x}, \boldsymbol{\Sigma}_{\boldsymbol{y}}). \tag{2.45}$$

Using Bayes theorem (2.11) together with (2.44) and (2.45), we obtain

$$f_{\mathbf{x}|\mathbf{y}_{1:N}}(\boldsymbol{x} \,|\, \boldsymbol{y}_{1:N}) = \frac{f_{\mathbf{x}}(\boldsymbol{x}) f_{\mathbf{y}_{1:N}|\mathbf{x}}(\boldsymbol{y}_{1:N} \,|\, \boldsymbol{x})}{f_{\mathbf{y}_{1:N}}(\boldsymbol{y}_{1:N})} \tag{2.46}$$

$$= \gamma(\boldsymbol{y}_{1:N}) \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_{\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{x}}) \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{y}_n; \boldsymbol{x}, \boldsymbol{\Sigma}_{\boldsymbol{y}}). \tag{2.47}$$

The expression in (2.47) is recognized to be a product of Gaussian pdfs, which is another Gaussian [18, 7.14]. We will now show that the posterior distribution is Gaussian. Using the precision matrices $\boldsymbol{\Lambda}_{\boldsymbol{y}} = \boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}$ and $\boldsymbol{\Lambda}_{\boldsymbol{x}} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}$, (2.47) can be written as

$$\begin{aligned} f_{\mathbf{x}|\mathbf{y}_{1:N}}(\boldsymbol{x} \,|\, \boldsymbol{y}_{1:N}) &\propto \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_{\boldsymbol{x}})^{\mathrm{T}} \boldsymbol{\Lambda}_{\boldsymbol{x}} (\boldsymbol{x}-\boldsymbol{\mu}_{\boldsymbol{x}})\right) \prod_{n=1}^{N} \exp\left(-\frac{1}{2}(\boldsymbol{y}_n-\boldsymbol{x})^{\mathrm{T}} \boldsymbol{\Lambda}_{\boldsymbol{y}} (\boldsymbol{y}_n-\boldsymbol{x})\right]) \\ &= \exp\left(-\frac{1}{2}\left((\boldsymbol{x}-\boldsymbol{\mu}_{\boldsymbol{x}})^{\mathrm{T}} \boldsymbol{\Lambda}_{\boldsymbol{x}} (\boldsymbol{x}-\boldsymbol{\mu}_{\boldsymbol{x}}) + \sum_{n=1}^{N} (\boldsymbol{y}_n-\boldsymbol{x})^{\mathrm{T}} \boldsymbol{\Lambda}_{\boldsymbol{y}} (\boldsymbol{y}_n-\boldsymbol{x})\right)\right) \\ &= \exp\left(-\frac{1}{2} E\right), \end{aligned} \tag{2.48}$$

with

$$E \triangleq (\boldsymbol{x}-\boldsymbol{\mu}_{\boldsymbol{x}})^{\mathrm{T}} \boldsymbol{\Lambda}_{\boldsymbol{x}} (\boldsymbol{x}-\boldsymbol{\mu}_{\boldsymbol{x}}) + \sum_{n=1}^{N} (\boldsymbol{y}_n-\boldsymbol{x})^{\mathrm{T}} \boldsymbol{\Lambda}_{\boldsymbol{y}} (\boldsymbol{y}_n-\boldsymbol{x}). \tag{2.49}$$

We claim that E can be written as a quadratic form in $\boldsymbol{x}$, while considering all remaining terms as constant, i.e., it is equal to

$$\tilde{E} = (\boldsymbol{x}-\boldsymbol{\mu}_{\boldsymbol{x}|\boldsymbol{y}_{1:N}})^{\mathrm{T}} \boldsymbol{\Sigma}_{\boldsymbol{x}|\boldsymbol{y}_{1:N}}^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_{\boldsymbol{x}|\boldsymbol{y}_{1:N}}) = \boldsymbol{x}^{\mathrm{T}} \boldsymbol{\Sigma}_{\boldsymbol{x}|\boldsymbol{y}_{1:N}}^{-1} \boldsymbol{x} - 2\boldsymbol{x}^{\mathrm{T}} \boldsymbol{\Sigma}_{\boldsymbol{x}|\boldsymbol{y}_{1:N}}^{-1} \boldsymbol{\mu}_{\boldsymbol{x}|\boldsymbol{y}_{1:N}} + \text{const.}, \tag{2.50}$$

with some mean $\boldsymbol{\mu}_{\boldsymbol{x}|\boldsymbol{y}_{1:N}}$ and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}|\boldsymbol{y}_{1:N}}$. To show that $E = \tilde{E}$ and find $\boldsymbol{\mu}_{\boldsymbol{x}|\boldsymbol{y}_{1:N}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{x}|\boldsymbol{y}_{1:N}}$, we expand (2.49) as

$$E = \boldsymbol{x}^{\mathrm{T}} \boldsymbol{\Lambda}_{\boldsymbol{x}} \boldsymbol{x} - \boldsymbol{x}^{\mathrm{T}} \boldsymbol{\Lambda}_{\boldsymbol{x}} \boldsymbol{\mu}_{\boldsymbol{x}} - \boldsymbol{\mu}_{\boldsymbol{x}}^{\mathrm{T}} \boldsymbol{\Lambda}_{\boldsymbol{x}} \boldsymbol{x} + \boldsymbol{\mu}_{\boldsymbol{x}}^{\mathrm{T}} \boldsymbol{\Lambda}_{\boldsymbol{x}} \boldsymbol{\mu}_{\boldsymbol{x}} + \sum_{n=1}^{N} (\boldsymbol{y}_n^{\mathrm{T}} \boldsymbol{\Lambda}_{\boldsymbol{y}} \boldsymbol{y}_n - \boldsymbol{x}^{\mathrm{T}} \boldsymbol{\Lambda}_{\boldsymbol{y}} \boldsymbol{y}_n - \boldsymbol{y}_n^{\mathrm{T}} \boldsymbol{\Lambda}_{\boldsymbol{y}} \boldsymbol{x} + \boldsymbol{x}^{\mathrm{T}} \boldsymbol{\Lambda}_{\boldsymbol{y}} \boldsymbol{x})$$

$$
\begin{aligned}
&= \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Lambda}_{\boldsymbol{x}}\boldsymbol{x} - 2\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Lambda}_{\boldsymbol{x}}\boldsymbol{\mu}_{\boldsymbol{x}} + \sum_{n=1}^{N}(\boldsymbol{y}_n^{\mathrm{T}}\boldsymbol{\Lambda}_{\boldsymbol{y}}\boldsymbol{y}_n - 2\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Lambda}_{\boldsymbol{y}}\boldsymbol{y}_n + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Lambda}_{\boldsymbol{y}}\boldsymbol{x}) + \text{const.} \\
&= \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Lambda}_{\boldsymbol{x}}\boldsymbol{x} + N\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Lambda}_{\boldsymbol{y}}\boldsymbol{x} - 2\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Lambda}_{\boldsymbol{x}}\boldsymbol{\mu}_{\boldsymbol{x}} - 2\sum_{n=1}^{N}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Lambda}_{\boldsymbol{y}}\boldsymbol{y}_n + \text{const.} \\
&= \boldsymbol{x}^{\mathrm{T}}(\boldsymbol{\Lambda}_{\boldsymbol{x}} + N\boldsymbol{\Lambda}_{\boldsymbol{y}})\boldsymbol{x} - 2\boldsymbol{x}^{\mathrm{T}}(\boldsymbol{\Lambda}_{\boldsymbol{x}}\boldsymbol{\mu}_{\boldsymbol{x}} + \boldsymbol{\Lambda}_{\boldsymbol{y}}\sum_{n=1}^{N}\boldsymbol{y}_n) + \text{const.} \qquad (2.51)
\end{aligned}
$$

Comparing the quadratic term in $\boldsymbol{x}$ in (2.51) with the quadratic term in (2.50), we obtain $\boldsymbol{\Sigma}^{-1}_{\boldsymbol{x}\,|\,\boldsymbol{y}_{1:N}} = \boldsymbol{\Lambda}_{\boldsymbol{x}} + N\boldsymbol{\Lambda}_{\boldsymbol{y}}$ or equivalently

$$
\boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y}_{1:N}} = (\boldsymbol{\Lambda}_{\boldsymbol{x}} + N\boldsymbol{\Lambda}_{\boldsymbol{y}})^{-1} = \left(\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1} + N\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}\right)^{-1}. \qquad (2.52)
$$

The last expression can be simplified using matrix identity (B.1) and an alternative expression can be obtained using (B.3), i.e.,

$$
\boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y}_{1:N}} \overset{\text{(B.1)}}{=} \boldsymbol{\Sigma}_{\boldsymbol{y}}\left(\boldsymbol{\Sigma}_{\boldsymbol{y}} + N\boldsymbol{\Sigma}_{\boldsymbol{x}}\right)^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}} \overset{\text{(B.3)}}{=} \boldsymbol{\Sigma}_{\boldsymbol{x}}\left(\boldsymbol{\Sigma}_{\boldsymbol{y}} + N\boldsymbol{\Sigma}_{\boldsymbol{x}}\right)^{-1}\boldsymbol{\Sigma}_{\boldsymbol{y}} \qquad (2.53)
$$

Similarly, comparing the linear terms in $\boldsymbol{x}$ occurring in (2.51) and in (2.50) yields

$$
\boldsymbol{\Sigma}^{-1}_{\boldsymbol{x}\,|\,\boldsymbol{y}_{1:N}}\boldsymbol{\mu}_{\boldsymbol{x}\,|\,\boldsymbol{y}_{1:N}} = \boldsymbol{\Lambda}_{\boldsymbol{x}}\boldsymbol{\mu}_{\boldsymbol{x}} + \boldsymbol{\Lambda}_{\boldsymbol{y}}\sum_{n=1}^{N}\boldsymbol{y}_n \qquad (2.54)
$$

and thus

$$
\begin{aligned}
\boldsymbol{\mu}_{\boldsymbol{x}\,|\,\boldsymbol{y}_{1:N}} &= \boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y}_{1:N}}(\boldsymbol{\Lambda}_{\boldsymbol{x}}\boldsymbol{\mu}_{\boldsymbol{x}} + \boldsymbol{\Lambda}_{\boldsymbol{y}}\sum_{n=1}^{N}\boldsymbol{y}_n) \\
&= \boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y}_{1:N}}(\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\mu}_{\boldsymbol{x}} + N\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}\bar{\boldsymbol{y}}_{1:N}) \\
&= \boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y}_{1:N}}\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\mu}_{\boldsymbol{x}} + N\boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y}_{1:N}}\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}\bar{\boldsymbol{y}}_{1:N}, \qquad (2.55)
\end{aligned}
$$

with the sample mean $\bar{\boldsymbol{y}}_{1:N} \triangleq \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{y}_n$. Finally, inserting the two alternative expressions in (2.53) into (2.55) yields

$$
\boldsymbol{\mu}_{\boldsymbol{x}\,|\,\boldsymbol{y}_{1:N}} = \boldsymbol{\Sigma}_{\boldsymbol{y}}\left(\boldsymbol{\Sigma}_{\boldsymbol{y}} + N\boldsymbol{\Sigma}_{\boldsymbol{x}}\right)^{-1}\boldsymbol{\mu}_{\boldsymbol{x}} + N\boldsymbol{\Sigma}_{\boldsymbol{x}}\left(\boldsymbol{\Sigma}_{\boldsymbol{y}} + N\boldsymbol{\Sigma}_{\boldsymbol{x}}\right)^{-1}\bar{\boldsymbol{y}}_{1:N}, \qquad (2.56)
$$

which can be interpreted as a matrix-weighted mean of the prior mean $\boldsymbol{\mu}_{\boldsymbol{x}}$ and the sample mean (mean of the measurements) $\bar{\boldsymbol{y}}_{1:N}$. Using the covariance matrix in (2.53) and the mean

in (2.56), we can finally write the posterior pdf in (2.47) as

$$f_{\mathbf{x}\,|\,\mathbf{y}_{1:N}} = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_{\boldsymbol{x}\,|\,\boldsymbol{y}_{1:N}}, \boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y}_{1:N}}). \tag{2.57}$$

Note that by completing the square in (2.51) in terms of $\boldsymbol{x}$ and considering all remaining terms as constant, the resulting expression in (2.57) is a normalized Gaussian pdf. We now insert (2.57) into (2.47) and obtain

$$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_{\boldsymbol{x}\,|\,\boldsymbol{y}_{1:N}}, \boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y}_{1:N}}) = \gamma(\boldsymbol{y}_{1:N}) \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_{\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{x}}) \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{y}_n; \boldsymbol{x}, \boldsymbol{\Sigma}_{\boldsymbol{y}}), \tag{2.58}$$

or equivalently,

$$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_{\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{x}}) \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{y}_n; \boldsymbol{x}, \boldsymbol{\Sigma}_{\boldsymbol{y}}) = \tilde{\gamma}(\boldsymbol{y}_{1:N}) \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_{\boldsymbol{x}\,|\,\boldsymbol{y}_{1:N}}, \boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y}_{1:N}}) \tag{2.59}$$

with

$$\tilde{\gamma}(\boldsymbol{y}_{1:N}) = f_{\mathbf{y}_{1:N}}(\boldsymbol{y}_{1:N}) = \int_{\boldsymbol{x}} \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_{\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{x}}) \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{y}_n; \boldsymbol{x}, \boldsymbol{\Sigma}_{\boldsymbol{y}}) d\boldsymbol{x}. \tag{2.60}$$

which as shown in Appendix C.2 is Gaussian and only depends on $\boldsymbol{y}_{1:N}$. Using $\boldsymbol{A} \to \mathbf{I}_N$, $\boldsymbol{B} \to \mathbf{I}_N$ and $\boldsymbol{\mu}_{\boldsymbol{y}_n} \to \mathbf{0}$, (C.35) reduces to (2.60), therefore we can use the result (C.36) and obtain

$$\tilde{\gamma}(\boldsymbol{y}_{1:N}) = \mathcal{N}\left(\boldsymbol{y}_{1:N}; \tilde{\boldsymbol{\mu}}_{\boldsymbol{y}_{1:N}}, \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{y}\boldsymbol{y}}\right) \tag{2.61}$$

with mean $\tilde{\boldsymbol{\mu}}_{\boldsymbol{y}_{1:N}}$

$$\tilde{\boldsymbol{\mu}}_{\boldsymbol{y}_{1:N}} = \begin{pmatrix} \boldsymbol{\mu}_{\boldsymbol{x}} \\ \vdots \\ \boldsymbol{\mu}_{\boldsymbol{x}} \end{pmatrix} \tag{2.62}$$

and covariance matrix $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{y}\boldsymbol{y}}$

$$\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{y}\boldsymbol{y}} = \begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{y}} + \boldsymbol{\Sigma}_{\boldsymbol{x}} & \boldsymbol{\Sigma}_{\boldsymbol{x}} & \dots & \boldsymbol{\Sigma}_{\boldsymbol{x}} \\ \boldsymbol{\Sigma}_{\boldsymbol{x}} & \boldsymbol{\Sigma}_{\boldsymbol{y}} + \boldsymbol{\Sigma}_{\boldsymbol{x}} & \ddots & \boldsymbol{\Sigma}_{\boldsymbol{x}} \\ \vdots & \ddots & \ddots & \vdots \\ \boldsymbol{\Sigma}_{\boldsymbol{x}} & \dots & \boldsymbol{\Sigma}_{\boldsymbol{x}} & \boldsymbol{\Sigma}_{\boldsymbol{y}} + \boldsymbol{\Sigma}_{\boldsymbol{x}} \end{pmatrix}. \tag{2.63}$$

Lastly, we discuss a special case, where only one measurement is available ($N = 1$). The

likelihood function is still given by

$$f_{\mathbf{y}\,|\,\mathbf{x}}(\boldsymbol{y}\,|\,\boldsymbol{x}) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{x}, \boldsymbol{\Sigma}_{\boldsymbol{y}}) \tag{2.64}$$

and the prior pdf is also Gaussian, with mean $\boldsymbol{\mu}_{\boldsymbol{x}}$ and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}}$, i.e.,

$$f_{\mathbf{x}}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_{\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{x}}). \tag{2.65}$$

Since this is a special case ($N = 1$) of the general result in (2.57), (2.69) is also Gaussian, i.e.,

$$f_{\mathbf{x}\,|\,\mathbf{y}}(\boldsymbol{x}\,|\,\boldsymbol{y}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_{\boldsymbol{x}\,|\,\boldsymbol{y}}, \boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y}}) \tag{2.66}$$

with posterior covariance matrix

$$\boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y}} = \boldsymbol{\Sigma}_{\boldsymbol{y}} \left(\boldsymbol{\Sigma}_{\boldsymbol{y}} + \boldsymbol{\Sigma}_{\boldsymbol{x}}\right)^{-1} \boldsymbol{\Sigma}_{\boldsymbol{x}} = \boldsymbol{\Sigma}_{\boldsymbol{x}} \left(\boldsymbol{\Sigma}_{\boldsymbol{y}} + \boldsymbol{\Sigma}_{\boldsymbol{x}}\right)^{-1} \boldsymbol{\Sigma}_{\boldsymbol{y}} \tag{2.67}$$

and posterior mean

$$\boldsymbol{\mu}_{\boldsymbol{x}\,|\,\boldsymbol{y}} = \boldsymbol{\Sigma}_{\boldsymbol{y}} \left(\boldsymbol{\Sigma}_{\boldsymbol{y}} + \boldsymbol{\Sigma}_{\boldsymbol{x}}\right)^{-1} \boldsymbol{\mu}_{\boldsymbol{x}} + \boldsymbol{\Sigma}_{\boldsymbol{x}} \left(\boldsymbol{\Sigma}_{\boldsymbol{y}} + \boldsymbol{\Sigma}_{\boldsymbol{x}}\right)^{-1} \boldsymbol{y}, \tag{2.68}$$

where we adapted (2.53) and (2.56) using $N = 1$. Similar to (2.59), we have

$$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_{\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{x}})\mathcal{N}(\boldsymbol{y}; \boldsymbol{x}, \boldsymbol{\Sigma}_{\boldsymbol{y}}) = \tilde{\gamma}(\boldsymbol{y})\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_{\boldsymbol{x}\,|\,\boldsymbol{y}}, \boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y}}) \tag{2.69}$$

with

$$\tilde{\gamma}(\boldsymbol{y}) = \mathcal{N}\left(\boldsymbol{y}; \boldsymbol{\mu}_{\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{y}} + \boldsymbol{\Sigma}_{\boldsymbol{x}}\right). \tag{2.70}$$

3 Dirichlet Process and Dirichlet Process Mixture

We now introduce the Dirichlet process (DP). The DP is a stochastic process whose realizations (outcomes) are discrete *probability distributions*, i.e., the DP is a distribution over distributions, as opposed to, e.g., the Gaussian distribution in (2.13), whose realizations are *real numbers*. This means that drawing from the DP produces a discrete distribution.

The DP is an instance of a Bayesian nonparametric model. Let us consider a simple example to illustrate the difference between a parametric model and a nonparametric model. In order to model the weight distribution of people, we could choose the Gaussian distribution and estimate the mean and variance of the distribution from a set of measurements. If we model the mean and variance as random, with a prior distribution, we are considering a Bayesian model. Unfortunately, choosing only one Gaussian distribution would not yield an accurate model, since the weight distributions of women and men differ. We could expand the model by using a mixture of two Gaussian distributions so that the mean weights of women and men are modeled separately; however, there are still many other cases to consider, such as athletes or children. Using this approach, it is necessary to know the number of distributions beforehand. By contrast, if we decide to use the DP as a prior for the mean and variance, we do not need to know the number of distributions. Using the DP as a prior distribution in a mixture model is referred to as Dirichlet process mixture (DPM). The DPM, just as the DP, is an example of a Bayesian nonparametric model, since it cannot be parameterized by a finite number of parameters. This is a difference from the Gaussian mixture model considered above, which is an example of a parametric model, since it is parameterized by a finite number of parameters (the mean and variance for each distribution).

In what follows, we first introduce the DP and study some of its properties. Subsequently, we introduce and briefly discuss the properties of the DPM. The presentation is based on [21] and [6].

3.1 Dirichlet Process Construction

The DP was introduced in [4] as a random probability measure. The formal definition of the DP thus requires measure theory, which is beyond the scope of this thesis. We will therefore adopt the definition in [3], which does not require measure theory and has the advantage of

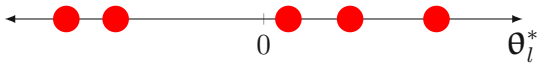


Figure 1: Random positions θ_l^* visualized for $P = 1$ and represented by red bullets.

being constructive. We will show presently what this means.

3.1.1 Positions and Weights

We start by defining two random sequences: a random sequence of *positions* θ_l^*, $l \in \mathbb{N}$ and a random sequence of *weights* Q_l, $l \in \mathbb{N}$. The sequence of random positions $(\theta_l^*)_{l=1}^{\infty}$ is defined to be independent of the sequence of the weights $(Q_l)_{l=1}^{\infty}$. In the literature, this construction of positions and weights is also referred to as *homogeneous* [21].

Random Positions

First, we consider a sequence of i.i.d. random vectors $\theta_l^* \in \mathbb{R}^P$ that are individually distributed according to a continuous distribution with pdf f_{H}, i.e.,

$$\theta_1^*, \theta_2^*, \ldots \sim_{\text{i.i.d.}} f_{\mathrm{H}}, \tag{3.1}$$

where f_{H} is referred to as the base distribution of the DP. The random variables θ_l^* can be thought of as random positions, as shown in Figure 1 for the one-dimensional case ($P = 1$).

Random Weights: Stick-Breaking Process

We consider an auxiliary sequence $V_l \in [0, 1]$, $l \in \mathbb{N}$, where the V_l are independent and individually Beta distributed with shape parameters $(1, \alpha)$, i.e.,

$$V_l \sim_{\text{i.i.d.}} \mathrm{Beta}(1, \alpha), \tag{3.2}$$

where $\alpha > 0$. We note that the support of the Beta distribution is $[0, 1]$, and for $\alpha = 1$, the uniform distribution on $[0, 1]$ is obtained. Using the sequence V_l, we define the weight sequence $Q_l \in [0, 1]$ by the following recursive construction:

$$Q_1 = V_1, \qquad Q_l = V_l \prod_{l'=1}^{l-1} (1 - V_{l'}), \quad l = 2, 3, \ldots. \tag{3.3}$$

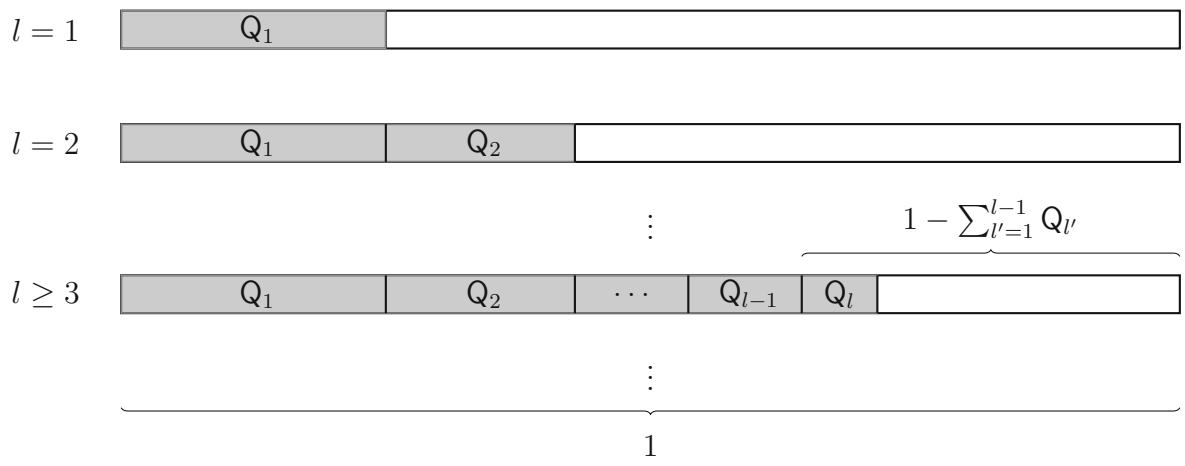


Figure 2: The stick-breaking process. The broken-off parts of the stick are represented by gray rectangles.

It can be shown that $\sum_{l=1}^{\infty} \mathsf{Q}_l = 1$ almost surely. The distribution of the Q_l resulting from this construction is called the GEM distribution (after Griffiths, Engen, and McCloskey [22]). We shall briefly write

$$(\mathsf{Q}_l)_{l=1}^{\infty} \sim \mathrm{GEM}\big((Q_l)_{l=1}^{\infty}; \alpha\big). \tag{3.4}$$

The construction in (3.3) is commonly referred to as the stick-breaking process [13]. Indeed, as shown in Appendix A.1, Eq. (3.3) can be reformulated as

$$\mathsf{Q}_1 = \mathsf{V}_1, \qquad \mathsf{Q}_l = \mathsf{V}_l \left(1 - \sum_{l'=1}^{l-1} \mathsf{Q}_{l'}\right), \quad l = 2, 3, \ldots. \tag{3.5}$$

Recalling that $\mathsf{V}_l \in [0, 1]$, this can be interpreted as follows. Consider a stick with length 1. Initially, for $l = 1$, we break off a part of that stick whose length Q_1 is equal to V_1, i.e., $\mathsf{Q}_1 = \mathsf{V}_1 \in [0, 1]$. Subsequently, for $l = 2, 3, \ldots$, we consider the currently remaining part of the stick, whose length is given by $1 - \sum_{l'=1}^{l-1} \mathsf{Q}_{l'}$, and we break off a part whose length Q_l is proportional to V_l, i.e., $\mathsf{Q}_l = \mathsf{V}_l \left(1 - \sum_{l'=1}^{l-1} \mathsf{Q}_{l'}\right)$. Thus, the random weights Q_l are constructed by repeatedly breaking off parts of a stick with length 1. This stick-breaking process is visualized in Figure 2.

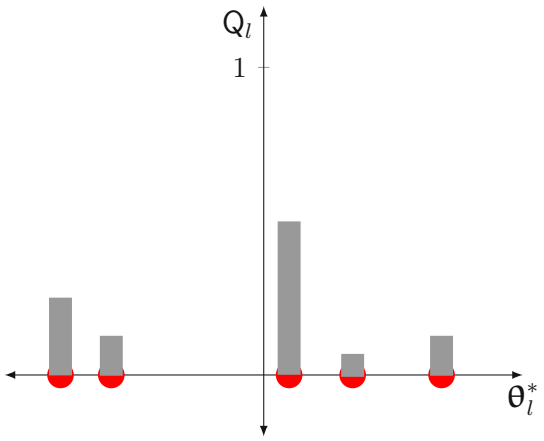


Figure 3: Visualization of the broken-off parts Q_l of the stick as vertical bars located at the random positions θ_l^* (red bullets).

Dirichlet Process Definition

Using the random positions $\theta_l^* \in \mathbb{R}^P$ and the random weights $\mathsf{Q}_l \in [0, 1]$ defined by the stick breaking process, we now define the DP as a random pdf f_{DP}, as follows:

$$\mathsf{f}_{\mathrm{DP}}(\boldsymbol{\theta}) = \sum_{l=1}^{\infty} \mathsf{Q}_l \delta_{\theta_l^*}(\boldsymbol{\theta}). \tag{3.6}$$

Here, $\delta_{\theta_l^*}(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_l^*)$ denotes the Dirac delta function at random position θ_l^*, the weight sequence $(\mathsf{Q}_l)_{l=1}^{\infty}$ is distributed according to $\mathrm{GEM}\big((Q_l)_{l=1}^{\infty}; \alpha\big)$ with $\alpha > 0$, and the position sequence $(\theta_l^*)_{l=1}^{\infty}$ is distributed i.i.d. according to the base distribution f_{H} defined on $\mathbb{R}^P$. Thus, $\mathsf{f}_{\mathrm{DP}}(\boldsymbol{\theta})$ is an infinite sum of Dirac delta functions at random positions θ_l^* weighted by random weights Q_l, as schematically shown in Figure 3. Accordingly, each realization of the random pdf actually describes a *discrete* distribution of θ: the random variable θ assumes the values θ_l^* with probabilities Q_l. We will denote the distribution of the random pdf f_{DP} as

$$\mathsf{f}_{\mathrm{DP}} \sim \mathrm{DP}(\alpha, f_{\mathrm{H}}), \tag{3.7}$$

where we call α the concentration parameter and f_{H} the base distribution.

Lastly, we note that the Dirichlet Process is *homogeneous* [21]. This means that the random positions θ_l^* are i.i.d. and independent of the weights Q_l, i.e.,

$$f_{\theta_l^* \,|\, \mathsf{Q}_l}(\boldsymbol{\theta}_l^* \,|\, Q_l) = f_{\theta_l^*}(\boldsymbol{\theta}_l^*) = f_{\mathrm{H}}(\boldsymbol{\theta}_l^*). \tag{3.8}$$

3.1.2 Generation of Random Vectors $\boldsymbol{\theta}_n$

Consider the following two-step generation of random vectors (samples) $\boldsymbol{\theta}_n \in \mathbb{R}^P$, $n \in \mathbb{N}$.

1. First, we define a DP by generating a random sequence of positions $(\boldsymbol{\theta}_l^*)_{l=1}^{\infty}$ that are distributed i.i.d. according to a base distribution f_{H} (see (3.1)), and a random sequence of weights $(\mathsf{Q}_l)_{l=1}^{\infty}$ (see (3.3)) that are distributed according to the GEM distribution with parameter α. Hence, the DP is given by the random pdf (see (3.6))

$$\mathsf{f}_{\mathrm{DP}}(\boldsymbol{\theta}) = \sum_{l=1}^{\infty} \mathsf{Q}_l \delta_{\boldsymbol{\theta}_l^*}(\boldsymbol{\theta}). \tag{3.9}$$

2. Then, given a realization $f_{\mathrm{DP}}(\boldsymbol{\theta})$ of the DP $\mathsf{f}_{\mathrm{DP}}(\boldsymbol{\theta})$ (equivalently, given realizations $(\boldsymbol{\theta}_l^*)_{l=1}^{\infty}$ and $(Q_l)_{l=1}^{\infty}$), we define the random vectors $\boldsymbol{\theta}_n$, $n \in \mathbb{N}$, to be distributed conditionally i.i.d. according to $f_{\mathrm{DP}}(\boldsymbol{\theta})$, i.e.,

$$\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots \,|\, (\mathsf{f}_{\mathrm{DP}} = f_{\mathrm{DP}}) \sim_{\text{i.i.d.}} f_{\mathrm{DP}}. \tag{3.10}$$

   From (3.10) and the fact that $f_{\mathrm{DP}}(\boldsymbol{\theta}) = \sum_{l=1}^{\infty} Q_l \delta_{\boldsymbol{\theta}_l^*}(\boldsymbol{\theta})$, it follows that for all $n \in \mathbb{N}$ we have

$$\boldsymbol{\theta}_n = \boldsymbol{\theta}_l^* \text{ with probability } Q_l. \tag{3.11}$$

   That is, conditioned on $\mathsf{f}_{\mathrm{DP}} = f_{\mathrm{DP}}$ or, equivalently, on $(\boldsymbol{\theta}_l^*)_{l=1}^{\infty} = (\boldsymbol{\theta}_l^*)_{l=1}^{\infty}$ and $(\mathsf{Q}_l)_{l=1}^{\infty} = (Q_l)_{l=1}^{\infty}$, $\boldsymbol{\theta}_n$ is chosen as $\boldsymbol{\theta}_l^*$ with probability Q_l, for $l \in \mathbb{N}$.

Empirical Reordering

Once the random vectors $\boldsymbol{\theta}_n$ are generated, we can perform a reordering $(\boldsymbol{\theta}_l^*)_{l=1}^{\infty} = (\boldsymbol{\vartheta}_s^*)_{s=1}^{\infty}$ of the random positions $\boldsymbol{\theta}_l^*$ as follows. For $s = 1$, we set

$$\boldsymbol{\vartheta}_1^* = \boldsymbol{\theta}_1. \tag{3.12}$$

For $s = 2$, we set $\boldsymbol{\vartheta}_2^* = \boldsymbol{\theta}_{n_2}$, where $\boldsymbol{\theta}_{n_2}$ is the next sample within the remaining sequence $(\boldsymbol{\theta}_n)_{n=2}^{\infty}$ that is not equal to $\boldsymbol{\theta}_1$. Similarly, for all subsequent new indices $s = 3, 4, \ldots$, we set $\boldsymbol{\vartheta}_s^* = \boldsymbol{\theta}_{n_s}$, where $\boldsymbol{\theta}_{n_s}$ is the next random position that is not equal to any previously used sample $\boldsymbol{\theta}_n$. Since the $\boldsymbol{\theta}_n$ were randomly drawn from the sequence $(\boldsymbol{\theta}_l^*)_{l=1}^{\infty}$, the sequence

$(\vartheta_s^*)_{s=1}^{\infty}$ is a permuted version of the sequence of random positions $(\theta_l^*)_{l=1}^{\infty}$, i.e.,

$$(\theta_l^*)_{l=1}^{\infty} = (\vartheta_{\sigma(l)}^*)_{l=1}^{\infty} \tag{3.13}$$

or individually

$$\theta_l^* = \vartheta_{\sigma(l)}^*, \quad l = 1, 2, \ldots, \tag{3.14}$$

corresponding to some index transformation (permutation)

$$s = \sigma(l), \quad l = 1, 2, \ldots. \tag{3.15}$$

Since this index transformation does not change the i.i.d. property and individual distribution of the θ_l^* given by (3.1), the permuted sequence of random positions $(\vartheta_s^*)_{s=1}^{\infty}$ is also distributed according to

$$\vartheta_1^*, \vartheta_2^*, \ldots \sim_{\text{i.i.d.}} f_{\text{H}}. \tag{3.16}$$

3.2 Properties of the Dirichlet Process

We now discuss the most important properties of the DP and the distributions associated with it.

3.2.1 The *The Rich Get Richer* Property and the Pólya Urn Model

The *the rich get richer* property of the DP can be explained by the following procedure of drawing colored balls from an urn [23], which is based on the Pólya Urn model [24]. Consider an urn that at time n contains various colored balls and $\alpha \in \mathbb{N}$ black balls. Balls are drawn from the urn at random. If a black ball is drawn, we return it back to the urn, together with an additional colored ball with a new color that is not contained in the urn; for example, the color can be sampled from a continuous color distribution. If a colored ball is drawn, we return it back to the urn, together with an additional ball of exactly the same color, thereby increasing the number of balls with this color.

We assume that at time $n = 0$ the urn only contains α black balls, therefore the first draw at time $n = 1$ results in a black ball, and a new colored ball is added to the urn (for example orange). Next, at time $n = 2$, we can either pick the orange-colored ball or one of the black balls. For larger α, we are more likely to pick a black ball, thus adding a new ball

with a distinct new color to the urn. This means that for a large α, the number of uniquely colored balls in the urn (i.e., the number of different colors) increases, as drawing a black ball also adds into the urn a ball whose color is different from those of the already existing balls. Conversely, for a small α, we are more likely to pick the orange ball, which results in two orange balls and α black balls in the urn. This means that for a small α, we are more likely to draw a colored ball from the urn and return it together with an additional ball of the same color. As a consequence, for a small α, the balls in the urn are less likely to have different colors. This is a manifestation of the *the rich get richer* property: for a small α, we are most likely to draw a colored ball from the urn, more specifically, the ball whose color is represented the most in the urn, i.e., an abundant or "rich" color. Drawing a colored ball adds an additional ball of exactly the same color into the urn, thereby increasing the probability that a ball with this abundant or "rich" color will be drawn again and further increasing the number of such colored balls in the urn.

This is closely related to the DP. The colored balls inside the urn can be thought of as random vectors $\boldsymbol{\theta}_n$, whereas the color of the ball corresponds to the distinct positions $\boldsymbol{\vartheta}_s^*$ and the number of black balls is related to the concentration parameter α, which is the only parameter of the GEM distribution (see (3.4)) of the weights Q_l.

In the next subsection, we will discuss DP-associated distributions that describe the drawing of the samples $\boldsymbol{\theta}_n$ from the DP, and in Section 3.3, we will further discuss the *the rich get richer* property. An expanded model presented in [23] also allows for $\alpha \in \mathbb{R}^+$, which is what we assume in the definition of the DP.

3.2.2 Distributions Associated with the Dirichlet Process

We will now discuss the marginal pdf $f_{\boldsymbol{\theta}_n}(\boldsymbol{\theta}_n)$ of an individual $\boldsymbol{\theta}_n$, the predictive pdf $f_{\boldsymbol{\theta}_n \,|\, \boldsymbol{\theta}_{1:n-1}}(\boldsymbol{\theta}_n \,|\, \boldsymbol{\theta}_{1:n-1})$, and the joint pdf $f_{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_N}(\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_N)$ of the length-N sequence of random vectors $(\boldsymbol{\theta}_n)_{n=1}^N$, equivalently written as a vector $\boldsymbol{\theta}_{1:N} = (\boldsymbol{\theta}_1^{\mathrm{T}},\ldots,\boldsymbol{\theta}_N^{\mathrm{T}})^{\mathrm{T}}$. Furthermore, we will discuss the posterior distribution and the conjugate posterior property of the DP.

Marginal Distribution

According to our generation model described in Section 3.1.2 (see (3.10)), the random vectors $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_N$ are conditionally i.i.d. given the realization of the DP, i.e., given $\mathsf{f}_{\mathrm{DP}} = f_{\mathrm{DP}}$

or equivalently given the sequences of positions Θ_l^* and weights Q_l, $l \in \mathbb{N}$. Note that in a practical sampling scheme, fixing the weights and positions is not feasible, as there are infinitely many of them. In (3.1), we defined the random positions Θ_l^* as i.i.d. with pdf $f_{\mathrm{H}}(\boldsymbol{\theta}_l^*)$.

To derive the marginal pdf $f_{\Theta_n}(\boldsymbol{\theta}_n)$, we recall that Θ_n is chosen as $\boldsymbol{\theta}_l^*$ with probability Q_l. Let $\mathcal{A}_{n,l}$ denote the event that Θ_n is chosen as $\boldsymbol{\theta}_l^*$, i.e., $\mathcal{A}_{n,l} \triangleq \{\Theta_n = \boldsymbol{\theta}_l^*\}$, and note that $P(\mathcal{A}_{n,l}) = Q_l$ for $n \in \mathbb{N}$. Therefore, we can write

$$f_{\Theta_n}(\boldsymbol{\theta}_n) = \sum_{l=1}^{\infty} f_{\Theta_n \,|\, \mathcal{A}_{n,l}}(\boldsymbol{\theta}_n \,|\, \mathcal{A}_{n,l}) P(\mathcal{A}_{n,l}). \tag{3.17}$$

Now $f_{\Theta_n \,|\, \mathcal{A}_{n,l}}(\boldsymbol{\theta}_n \,|\, \mathcal{A}_{n,l}) = f_{\mathrm{H}}(\boldsymbol{\theta}_n)$ since $\mathcal{A}_{n,l} = \{\Theta_n = \boldsymbol{\theta}_l^*\}$, i.e., given $\mathcal{A}_{n,l}$, $\boldsymbol{\theta}_n$ equals $\boldsymbol{\theta}_l^*$ and thus the pdf of Θ_n (still given $\mathcal{A}_{n,l}$) equals the pdf of Θ_l^*, which is $f_{\mathrm{H}}(\boldsymbol{\theta}_n)$. Furthermore, $P(\mathcal{A}_{n,l}) = Q_l$. Therefore, (3.17) becomes

$$f_{\Theta_n}(\boldsymbol{\theta}_n) = \sum_{l=1}^{\infty} f_{\mathrm{H}}(\boldsymbol{\theta}_n) Q_l = f_{\mathrm{H}}(\boldsymbol{\theta}_n) \underbrace{\sum_{l=1}^{\infty} Q_l}_{1} = f_{\mathrm{H}}(\boldsymbol{\theta}_n). \tag{3.18}$$

Thus, the marginal pdf of each Θ_n, $f_{\Theta_n}(\boldsymbol{\theta}_n)$, equals the base pdf $f_{\mathrm{H}}(\boldsymbol{\theta}_n)$.

As an alternative to the above derivation, the marginal distribution can be derived as follows. As discussed in [22, Theorem 14] and in [13, Sec. 2.2.1], the sequence of random vectors $(\Theta_n)_{n=1}^N$ generated according to (3.10) is *exchangeable*. This means that for any permutation $\sigma(1), \ldots, \sigma(N)$ of the indices $1, \ldots, N$, the joint distribution of $\Theta_1, \ldots, \Theta_N$ is equal to the joint distribution of $\Theta_{\sigma(1)}, \ldots, \Theta_{\sigma(N)}$, i.e., we have

$$f_{\Theta_1,\ldots,\Theta_N}(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N) = f_{\Theta_{\sigma(1)},\ldots,\Theta_{\sigma(N)}}(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N). \tag{3.19}$$

(We note that the joint distribution of $\Theta_1, \ldots, \Theta_N$ will be discussed in detail in a later subsection.) As a consequence of de Finetti's representation theorem, the exchangeability of samples Θ_n implies that the marginal distribution of any Θ_n equals the marginal distribution of any specific $\Theta_{n'}$, say Θ_1, i.e., $f_{\Theta_n}(\boldsymbol{\theta}_n) = f_{\Theta_1}(\boldsymbol{\theta}_n)$ [25]. Furthermore, from the fact that Θ_1 equals ϑ_1^*, we conclude that $f_{\Theta_1}(\boldsymbol{\theta}_1)$ equals $f_{\mathrm{H}}(\vartheta_1^*)$ which is the base distribution f_{H} (see

(3.16)). Thus, we obtain

$$f_{\boldsymbol{\theta}_n}(\boldsymbol{\theta}_n) = f_{\boldsymbol{\theta}_1}(\boldsymbol{\theta}_n) = f_{\boldsymbol{\vartheta}_1^*}(\boldsymbol{\theta}_n) = f_{\mathrm{H}}(\boldsymbol{\theta}_n), \quad \text{for all} \quad n \in \{1, \ldots, N\}. \tag{3.20}$$

In other words, the marginal distribution of any individual $\boldsymbol{\theta}_n$ is the base distribution f_{H}.

The random vectors $\boldsymbol{\theta}_n$ are identically distributed (see (3.18) or (3.20)) but, in contrast to the $\boldsymbol{\theta}_l^*$ or $\boldsymbol{\vartheta}_s^*$, they are not independent. Consider drawing $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_1$ from the DP, which consequently fixes $\boldsymbol{\vartheta}_1^* = \boldsymbol{\theta}_1$ as well (see (3.12)). Then, when drawing $\boldsymbol{\theta}_2$, we either obtain $\boldsymbol{\theta}_2 = \boldsymbol{\vartheta}_1^* = \boldsymbol{\theta}_1$ or another randomly chosen position $\boldsymbol{\vartheta}_2^*$. This shows that $\boldsymbol{\theta}_2$ is not independent of $\boldsymbol{\theta}_1$. The pdf of $\boldsymbol{\theta}_2$ given $\boldsymbol{\theta}_1$ will be discussed in the following subsection.

Predictive Distribution

Let us now consider the pdf of $\boldsymbol{\theta}_n$ given the vector of previously drawn samples $\boldsymbol{\theta}_{1:n-1} = (\boldsymbol{\theta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\theta}_{n-1}^{\mathrm{T}})^{\mathrm{T}}$, i.e., $f_{\boldsymbol{\theta}_n \,|\, \boldsymbol{\theta}_{1:n-1}}(\boldsymbol{\theta}_n \,|\, \boldsymbol{\theta}_{1:n-1})$, which is sometimes referred to as the predictive pdf. In [4], it was shown that

$$f_{\boldsymbol{\theta}_n \,|\, \boldsymbol{\theta}_{1:n-1}}(\boldsymbol{\theta}_n \,|\, \boldsymbol{\theta}_{1:n-1}) = \frac{\alpha}{\alpha + n - 1} f_{\mathrm{H}}(\boldsymbol{\theta}_n) + \frac{1}{\alpha + n - 1} \sum_{n'=1}^{n-1} \delta_{\boldsymbol{\theta}_{n'}}(\boldsymbol{\theta}_n). \tag{3.21}$$

This is a mixture distribution involving the continuous base distribution $f_{\mathrm{H}}(\boldsymbol{\theta}_n)$ and up to $n-1$ distinct discrete components $\delta_{\boldsymbol{\theta}_{n'}}(\boldsymbol{\theta}_n)$, $n' = 1, \ldots, n-1$. We note that the $\boldsymbol{\theta}_{n'}$ are not necessarily distinct. The predictive pdf $f_{\boldsymbol{\theta}_n \,|\, \boldsymbol{\theta}_{1:n-1}}(\boldsymbol{\theta}_n \,|\, \boldsymbol{\theta}_{1:n-1})$ in (3.21) can be interpreted as follows: given the previously drawn $\boldsymbol{\theta}_{1:n-1}$, the new $\boldsymbol{\theta}_n$ is either drawn from the base pdf $f_{\mathrm{H}}(\boldsymbol{\theta}_n)$ with probability $\frac{\alpha}{\alpha+n-1}$ or is equal to one of the previously drawn $\boldsymbol{\theta}_{n'}$, $n' \in \{1, \ldots, n-1\}$, with probability $\frac{1}{\alpha+n-1}$ (i.e., equal to $\boldsymbol{\theta}_1$ with probability $\frac{1}{\alpha+n-1}$, equal to $\boldsymbol{\theta}_2$ with probability $\frac{1}{\alpha+n-1}$, etc.). Note however, that $\boldsymbol{\theta}_{n'}$ may take on identical values. We will further discuss the predictive pdf $f_{\boldsymbol{\theta}_n \,|\, \boldsymbol{\theta}_{1:n-1}}(\boldsymbol{\theta}_n \,|\, \boldsymbol{\theta}_{1:n-1})$ in more detail, including a simulated example, in Section 3.3.

Lastly, we consider a sequence of samples $(\boldsymbol{\theta}_n)_{n=1}^N$ from the DP and discuss the conditional pdf of one sample $\boldsymbol{\theta}_n$, $n \in \{1, \ldots, N\}$, given a subset of the remaining samples, i.e., $\{\boldsymbol{\theta}_m\}_{m \in \mathcal{M}}$ with $\mathcal{M} = \{n_{m(1)}, \ldots, n_{m(|\mathcal{M}|)}\} \subseteq \{1, \ldots, N\} \backslash \{n\}$, with $|\mathcal{M}|$ denoting the cardinality of the set $\mathcal{M}$. As previously stated, the samples $\boldsymbol{\theta}_n$, $n = 1, \ldots, N$ are exchangeable. This implies that also the samples $\boldsymbol{\theta}_m$, $m \in \mathcal{M}$ with $\mathcal{M} \subset \{1, \ldots, N\}$ are exchange-

able [26, Theorem 1]. Thus, in analogy to (3.19), the joint pdf of all $\mathsf{\theta}_m$, $m \in \mathcal{M}$ satisfies

$$f_{\mathsf{\theta}_{m(1)},\ldots,\mathsf{\theta}_{m(|\mathcal{M}|)}}(\boldsymbol{\theta}_{m(1)},\ldots,\boldsymbol{\theta}_{m(|\mathcal{M}|)}) = f_{\mathsf{\theta}_{m(\sigma(1))},\ldots,\mathsf{\theta}_{m(\sigma(|\mathcal{M}|))}}(\boldsymbol{\theta}_{m(1)},\ldots,\boldsymbol{\theta}_{m(|\mathcal{M}|)}). \tag{3.22}$$

Using [27, Proposition 6.7] and the exchangeability of the subset $\{\mathsf{\theta}_m\}_{m\in\mathcal{M}}$, we can generalize (3.21) to obtain

$$f_{\mathsf{\theta}_n \,|\, (\mathsf{\theta}_m)_{m\in\mathcal{M}}}(\boldsymbol{\theta}_n \,|\, (\boldsymbol{\theta}_m)_{m\in\mathcal{M}}) = \frac{\alpha}{\alpha+|\mathcal{M}|} f_{\mathrm{H}}(\boldsymbol{\theta}_n) + \frac{1}{\alpha+|\mathcal{M}|}\sum_{m\in\mathcal{M}} \delta_{\boldsymbol{\theta}_m}(\boldsymbol{\theta}_n), \tag{3.23}$$

for any $n \in \{1,\ldots,N\}$ and $\mathcal{M} \subseteq \{1,\ldots,N\}\backslash\{n\}$.

Generation of the Samples $\mathsf{\theta}_n$ Using the Predictive pdf

Let us consider the following recursive construction of a sequence of samples $\mathsf{\theta}_n$, (for $n = 1, 2, \ldots$):

1. Draw the first sample $\mathsf{\theta}_1$ from the base distribution f_{H}.
2. For $n = 2, 3, \ldots$, draw the next sample $\mathsf{\theta}_n$ from the predictive pdf $f_{\mathsf{\theta}_n \,|\, \mathsf{\theta}_{1:n-1}}(\boldsymbol{\theta}_n \,|\, \boldsymbol{\theta}_{1:n-1})$ in (3.21).

We note this generation procedure of the samples $\mathsf{\theta}_n$ does not involve the weights Q_l. The resulting sequence of random vectors $\mathsf{\theta}_n$, $n \in \mathbb{N}$ is called a Pólya sequence. A Pólya sequence can be shown to be exchangeable, i.e., it satisfies the generalized permutation invariance property (3.19) for $N \to \infty$. Therefore, as was shown in [5, Sec. 4.2.4], by de Finetti's theorem there exists a random pdf f_{DP} such that

$$\mathsf{\theta}_1, \mathsf{\theta}_2, \ldots \,|\, (\mathsf{f}_{\mathrm{DP}} = f_{\mathrm{DP}}) \sim_{\text{i.i.d.}} f_{\mathrm{DP}}, \tag{3.24}$$

with $\mathsf{f}_{\mathrm{DP}} \sim \mathrm{DP}(\alpha, f_{\mathrm{H}})$. A comparison with (3.10) shows that the considered recursive generation based on the predictive pdf is equivalent to our previous generation procedure described in Section 3.1.2.

Joint Distribution

The joint pdf $f_{\theta_{1:N}}(\boldsymbol{\theta}_{1:N}) = f_{\theta_1,\dots,\theta_N}(\boldsymbol{\theta}_1,\dots,\boldsymbol{\theta}_N)$ has a complicated mixture structure; however, it can be derived by applying the chain rule [5, Sec. 4.1.4]. Indeed, we have

$$\begin{aligned} f_{\theta_1,\dots,\theta_N}(\boldsymbol{\theta}_1,\dots,\boldsymbol{\theta}_N) &= f_{\theta_1}(\boldsymbol{\theta}_1) f_{\theta_2\,|\,\theta_1}(\boldsymbol{\theta}_2\,|\,\boldsymbol{\theta}_1)\dots f_{\theta_N\,|\,\theta_1,\dots,\theta_{N-1}}(\boldsymbol{\theta}_N\,|\,\boldsymbol{\theta}_1,\dots,\boldsymbol{\theta}_{N-1}) \\ &= f_{\theta_1}(\boldsymbol{\theta}_1)\prod_{n=2}^{N} f_{\theta_n\,|\,\theta_{1:n-1}}(\boldsymbol{\theta}_n\,|\,\boldsymbol{\theta}_{1:n-1}). \end{aligned} \tag{3.25}$$

According to (3.20), we have

$$f_{\theta_1}(\boldsymbol{\theta}_1) = f_{\mathrm{H}}(\boldsymbol{\theta}_1). \tag{3.26}$$

Using (3.26) and (3.21) in (3.25) finally yields

$$f_{\theta_1,\dots,\theta_N}(\boldsymbol{\theta}_1,\dots,\boldsymbol{\theta}_N) = f_{\mathrm{H}}(\boldsymbol{\theta}_1)\prod_{n=2}^{N}\left(\frac{\alpha}{\alpha+n-1}f_{\mathrm{H}}(\boldsymbol{\theta}_n) + \frac{1}{\alpha+n-1}\sum_{n'=1}^{n-1}\delta_{\boldsymbol{\theta}_{n'}}(\boldsymbol{\theta}_n)\right). \tag{3.27}$$

This is recognized to be a product mixture of the base distribution $f_{\mathrm{H}}(\boldsymbol{\theta}_n)$ and discrete components $\delta_{\boldsymbol{\theta}_{n'}}(\boldsymbol{\theta}_n)$ for $n' = 1,\dots,n-1$, weighted by $\frac{\alpha}{\alpha+n-1}$ and $\frac{1}{\alpha+n-1}$ respectively.

We recall from (3.19) that the samples θ_n are exchangeable, i.e., for any permutation $\sigma(1),\dots,\sigma(N)$ of the indices $1,\dots,N$, we have $f_{\theta_1,\dots,\theta_N}(\boldsymbol{\theta}_1,\dots,\boldsymbol{\theta}_N) = f_{\theta_{\sigma(1)},\dots,\theta_{\sigma(N)}}(\boldsymbol{\theta}_1,\dots,\boldsymbol{\theta}_N)$, even though this is not obvious from expression (3.27). For further discussion of the joint pdf, we refer to [28].

Posterior Distribution

Let us consider the random pdf $\mathsf{f}_{\mathrm{DP}} \sim \mathrm{DP}(\alpha, f_{\mathrm{H}})$ and the random variables $\theta_{1:N}$. With an abuse of language, the random variables $\theta_{1:N}$ are often referred to as samples from the Dirichlet process, or as observations. The observations are sampled conditionally i.i.d. (given $\mathsf{f}_{\mathrm{DP}} = f_{\mathrm{DP}}$) from the distribution f_{DP}, as stated in (3.10). The distribution f_{DP} was previously drawn from $\mathrm{DP}(\alpha, f_{\mathrm{H}})$. We will refer to the distribution of the random pdf $\mathsf{f}_{\mathrm{DP}} \sim \mathrm{DP}(\alpha, f_{\mathrm{H}})$ as *DP prior* and to the distribution of the random f_{DP} *after* we observed $\theta_{1:N}$, i.e., $\mathsf{f}_{\mathrm{DP}}\,|\,(\theta_{1:N} = \boldsymbol{\theta}_{1:N})$, as *DP posterior*. An important property of the DP prior is its conjugate posterior [21]. Indeed, it was shown in [4] that the DP posterior is given by

$$\mathsf{f}_{\mathrm{DP}}\,|\,(\theta_{1:N} = \boldsymbol{\theta}_{1:N}) \sim \mathrm{DP}(\alpha, \tilde{f}_{\mathrm{H}}), \tag{3.28}$$

where the base pdf $\tilde{f}_{\mathrm{H}}$ is given by

$$\tilde{f}_{\mathrm{H}}(\boldsymbol{\theta}) = \frac{\alpha}{\alpha + N} f_{\mathrm{H}}(\boldsymbol{\theta}) + \frac{1}{\alpha + N} \sum_{n=1}^{N} \delta_{\boldsymbol{\theta}_n}(\boldsymbol{\theta}). \tag{3.29}$$

We see that the posterior $\mathsf{f}_{\mathrm{DP}} \,|\, (\boldsymbol{\Theta}_{1:N} = \boldsymbol{\theta}_{1:N})$ in (3.28) is again a DP; however, the base distribution is no longer f_{H} but $\tilde{f}_{\mathrm{H}}$ as given in (3.29). That is, given $\boldsymbol{\Theta}_{1:N} = \boldsymbol{\theta}_{1:N}$, the random positions $\boldsymbol{\Theta}_l^*$ are now drawn from f_{H} with probability $\frac{\alpha}{\alpha+N}$ and equal to $\boldsymbol{\theta}_n$ for $n \in \{1, \ldots, N\}$ with probability $\frac{1}{\alpha+N}$. The conjugacy is similar to a Gaussian prior having a Gaussian posterior (provided the likelihood function is Gaussian). In our case, the random pdf $\mathsf{f}_{\mathrm{DP}} \sim \mathrm{DP}(\alpha, f_{\mathrm{H}})$ is the DP prior, whereas the likelihood function is given by $\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_N \,|\, (\mathsf{f}_{\mathrm{DP}} = f_{\mathrm{DP}}) \sim_{\text{i.i.d.}} f_{\mathrm{DP}}$ (see (3.10)).

3.3 Clustering Property and Chinese Restaurant Process

We now discuss the clustering property of the DP and the effect of the value of the concentration parameter α on the samples $\boldsymbol{\Theta}_{1:N}$.

3.3.1 Clustering Property and Random Partition

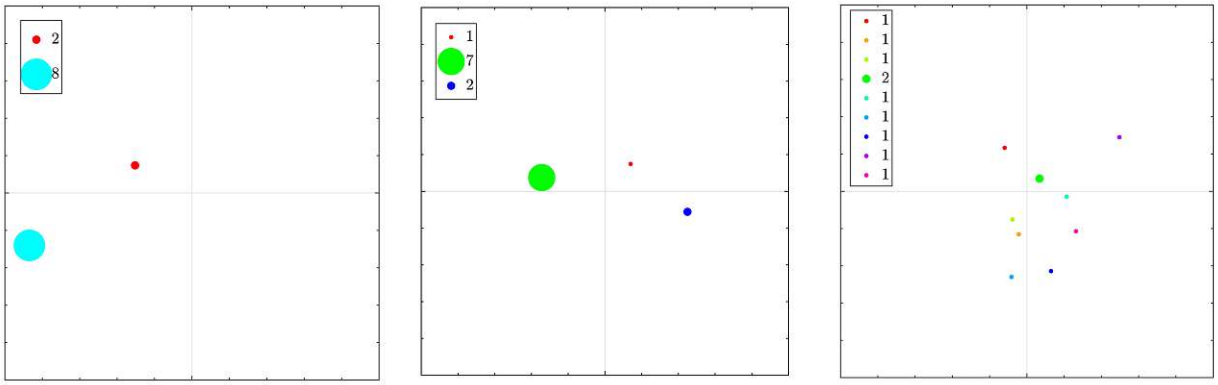
In what follows, let $S(N)$ be the number of unique positions $\boldsymbol{\vartheta}_s^*$ within the N samples $\boldsymbol{\Theta}_{1:N}$, and let $\boldsymbol{\vartheta}_{1:S(N)}^* = (\boldsymbol{\vartheta}_1^{*\mathrm{T}}, \ldots, \boldsymbol{\vartheta}_{S(N)}^{*\mathrm{T}})^{\mathrm{T}}$ be the vector composed of these unique positions $\boldsymbol{\vartheta}_s^*$ we observe within $\boldsymbol{\Theta}_{1:N}$. We note that $1 \leq S(N) \leq N$. Furthermore, $\tilde{m}_s(N)$ denotes the number of times $\boldsymbol{\vartheta}_s^*$ is observed within $\boldsymbol{\theta}_{1:N}$, i.e.,

$$\tilde{m}_s(N) \triangleq \sum_{n=1}^{N} \mathbb{1}(\boldsymbol{\theta}_n = \boldsymbol{\vartheta}_s^*), \quad s = 1, \ldots, S(N), \tag{3.30}$$

where $\mathbb{1}$ denotes the indicator function, i.e., $\mathbb{1}(\boldsymbol{\theta}_n = \boldsymbol{\vartheta}_s^*) = 1$ if $\boldsymbol{\theta}_n = \boldsymbol{\vartheta}_s^*$ and 0 otherwise. Since $\boldsymbol{\vartheta}_{1:S(N)}^*$ comprises only positions $\boldsymbol{\vartheta}_s^*$ that have been observed within $\boldsymbol{\Theta}_{1:N}$ and $s \in \{1, \ldots, S(N)\}$, we note that $\tilde{m}_s(N) = 0$ is not possible. Using (3.30), we have

$$\sum_{s=1}^{S(N)} \tilde{m}_s(N) = N. \tag{3.31}$$

We consider a simple simulated example, with a DP with dimension $P = 2$, i.e., $\boldsymbol{\Theta}_l^* \in \mathbb{R}^2$, and Gaussian base distribution $f_{\mathrm{H}}(\boldsymbol{\theta}_l^*) = \mathcal{N}(\boldsymbol{\theta}_l^*; \mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_2$, where $\sigma^2 = 2$ and $\mathbf{I}_2$



(a) $\alpha = 0.1$ (b) $\alpha = 1$ (c) $\alpha = 10$

Figure 4: Example of $N = 10$ samples $\boldsymbol{\theta}_n$ drawn from the DP. The observed distinct positions $\boldsymbol{\vartheta}_s^*$ are represented by circles whose radius is proportional to the number $\tilde{m}_s(10)$ of samples $\boldsymbol{\theta}_n$ that are equal to $\boldsymbol{\vartheta}_s^*$.

is the identity matrix of size 2×2. To demonstrate the role of the concentration parameter α, we consider three different values $\alpha = 0.1, 1,$ and 10. We draw $N = 10$ samples $\boldsymbol{\theta}_n$, $n = 1, \ldots, 10$ from the DP. Figure 4 shows realizations of the distinct $\boldsymbol{\vartheta}_s^*$ (or equivalently, of the $\boldsymbol{\theta}_l^*$, see (3.13)) observed within our $N = 10$ samples $\boldsymbol{\theta}_n$. Each realization of $\boldsymbol{\vartheta}_s^*$ is represented by a circle whose radius is proportional to $\tilde{m}_s(10)$, i.e. the number of samples $\boldsymbol{\theta}_n$ that are equal to $\boldsymbol{\vartheta}_s^*$. Evidently, for each s, this number is at least 1 and at most $N = 10$. For $\alpha = 0.1$, we observed only two distinct positions $\boldsymbol{\vartheta}_s^*$, i.e., $\boldsymbol{\vartheta}_1^*$ and $\boldsymbol{\vartheta}_2^*$, whereas for $\alpha = 10$ there are nine distinct positions $\boldsymbol{\vartheta}_1^*, \ldots, \boldsymbol{\vartheta}_9^*$. This also shows that a strong concentration of the $\boldsymbol{\theta}_n$, corresponding to a small number $\tilde{m}_s(10)$ of distinct values $\boldsymbol{\vartheta}_s^*$, is obtained for a *small* concentration parameter α.

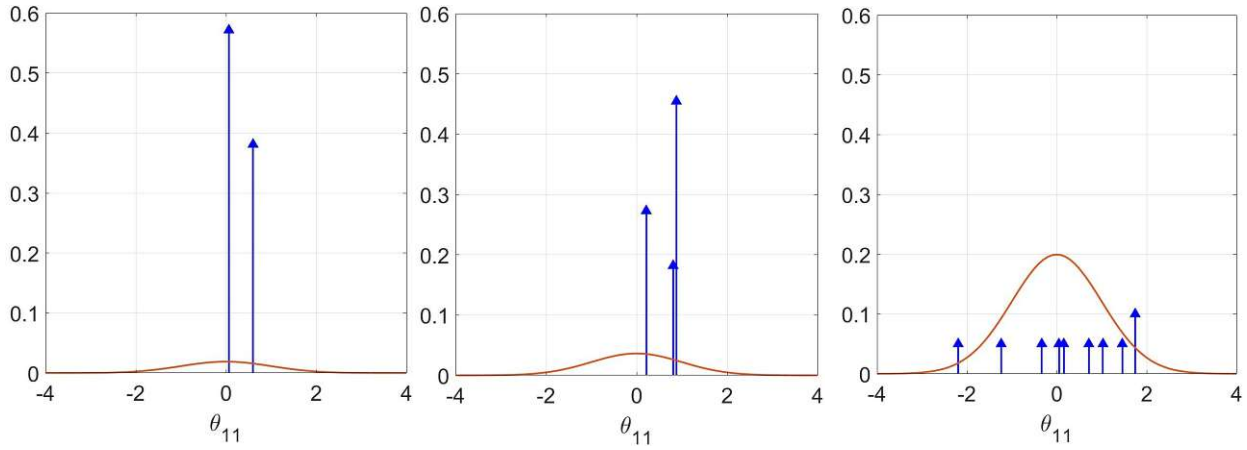
The joint pdf in (3.27) can be rewritten using (3.30) as

$$f_{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_N}(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N) = f_{\mathrm{H}}(\boldsymbol{\theta}_1) \prod_{n=2}^{N} \left(\frac{\alpha}{\alpha + n - 1} f_{\mathrm{H}}(\boldsymbol{\theta}_n) + \frac{1}{\alpha + n - 1} \sum_{s=1}^{S(n-1)} \tilde{m}_s(n-1) \delta_{\boldsymbol{\vartheta}_s^*}(\boldsymbol{\theta}_n) \right), \tag{3.32}$$

with $S(n-1)$ indicating the number of unique positions $\boldsymbol{\vartheta}_s^*$ observed within the sequence $\boldsymbol{\theta}_{1:n-1}$, for $2 \leq n \leq N$. Similarly, we can rewrite the predictive pdf (3.21) as

$$f_{\boldsymbol{\theta}_n \,|\, \boldsymbol{\theta}_{1:n-1}}(\boldsymbol{\theta}_n \,|\, \boldsymbol{\theta}_{1:n-1}) = \frac{\alpha}{\alpha + n - 1} f_{\mathrm{H}}(\boldsymbol{\theta}_n) + \frac{1}{\alpha + n - 1} \sum_{s=1}^{S(n-1)} \tilde{m}_s(n-1) \delta_{\boldsymbol{\vartheta}_s^*}(\boldsymbol{\theta}_n). \tag{3.33}$$

Note that in this mixture distribution, there are $S(n-1)$ discrete components at the posi-



(a) $\alpha = 0.5$ (b) $\alpha = 1$ (c) $\alpha = 10$

Figure 5: Example of the predictive pdf $f_{\theta_{11}\,|\,\boldsymbol{\theta}_{1:10}}(\theta_{11}\,|\,\boldsymbol{\theta}_{1:10})$ for $P = 1$, after $n-1 = 10$ samples $\boldsymbol{\theta}_{1:10}$ have been drawn from the DP. The discrete components of the pdf, i.e., the Dirac delta functions $\delta_{\vartheta_s^*}(\theta_{11})$ located at the ϑ_s^*, are represented graphically by their weights $\frac{\tilde{m}_s(10)}{\alpha+10}$, shown in blue, whereas the continuous component (the Gaussian base pdf weighted by $\frac{\alpha}{\alpha+10}$), is shown in red.

tions $\boldsymbol{\vartheta}_s^*$, $s = 1, \ldots, S(n-1)$, and the probability that $\boldsymbol{\theta}_n = \boldsymbol{\vartheta}_s^*$ (given $\boldsymbol{\theta}_{1:n-1}$) is $\frac{\tilde{m}_s(n-1)}{\alpha+n-1}$, i.e., proportional to the number $\tilde{m}_s(n-1)$ of times $\boldsymbol{\vartheta}_s^*$ was observed within the sequence $\boldsymbol{\theta}_{1:n-1}$. This is a manifestation of the *the rich get richer* property. Each time we observe $\boldsymbol{\theta}_{n'} = \boldsymbol{\vartheta}_s^*$, the number $\tilde{m}_s(n-1)$ grows, therefore also the probability that the subsequent sample $\boldsymbol{\theta}_n$ equals $\boldsymbol{\vartheta}_s^*$, which is proportional to $\tilde{m}_s(n-1)$, also increases. Figure 5 visualizes expression (3.33) for $P = 1$, i.e., $\theta_l^* \in \mathbb{R}$, and a Gaussian base distribution $f_{\mathrm{H}}(\theta_l^*) = \mathcal{N}(\theta_l^*; 0, \sigma^2)$ with $\sigma^2 = 1$. We consider three different concentration parameters $\alpha = 0.5$, 1, and 10 and draw samples $\boldsymbol{\theta}_{n'}$, $n' = 1, \ldots, 10$. Figure 5 shows the predictive pdf $f_{\boldsymbol{\theta}_n\,|\,\boldsymbol{\theta}_{1:n-1}}(\boldsymbol{\theta}_n\,|\,\boldsymbol{\theta}_{1:n-1}) = f_{\boldsymbol{\theta}_{11}\,|\,\boldsymbol{\theta}_{1:10}}(\theta_{11}\,|\,\boldsymbol{\theta}_{1:10})$, i.e., the pdf of the sample $\boldsymbol{\theta}_{11}$ drawn after the first $n-1 = 10$ samples $\boldsymbol{\theta}_{1:10}$ have been observed. For the small concentration parameter $\alpha = 0.5$ (shown in Figure 5a), $\boldsymbol{\theta}_{11}$ will most likely be equal to either ϑ_1^* or ϑ_2^*, as the weighted base pdf (shown in red) is very small compared to the weights of the Dirac delta functions, located at the ϑ_s^*. On the other hand, for the large concentration parameter $\alpha = 10$ (shown in Figure 5c), $\boldsymbol{\theta}_{11}$ will most likely be sampled from the base distribution, which means it will be different from the previously drawn samples ϑ_s^*. Furthermore, we can also see that for $\alpha = 0.5$, there are only $S(10) = 2$ distinct positions ϑ_s^*, whereas for $\alpha = 10$, there are $S(10) = 9$ distinct positions ϑ_s^*. This is another manifestation of the *the rich get richer* property.

Number of Unique Positions

We consider N samples $\boldsymbol{\theta}_{1:N}$ from a DP with concentration parameter α. As before, $S(N)$ denotes the number of the unique positions $\boldsymbol{\vartheta}_s^*$ within the samples $\boldsymbol{\theta}_{1:N}$. Figure 4 suggests that $S(N)$ depends on the parameter α; however, Figure 4 only shows one realization of $S(N)$. Furthermore, $S(N)$ also depends on the number of samples N. For a random sequence of samples $\boldsymbol{\theta}_{1:N}$, $\mathsf{S}(N)$ is itself a random number. It can be shown [5, Prop. 4.8] that the expected value of $\mathsf{S}(N)$ can be approximated by

$$\mathbb{E}[\mathsf{S}(N)] \approx \alpha \log(N). \tag{3.34}$$

This approximation is asymptotically exact as $N \to \infty$.

Random Partition

Let us consider an ordered random partition of $\mathbb{N}$,

$$\boldsymbol{\Phi} = (\Phi_1, \Phi_2, \ldots), \tag{3.35}$$

with an infinite number of subsets $\Phi_l \subset \mathbb{N}$. Each $n \in \mathbb{N}$ is contained in exactly one subset Φ_l, which we refer to as the l^{th} *cluster*. This means that

$$\mathbb{N} = \Phi_1 \cup \Phi_2 \cup \ldots \quad \text{and} \quad \Phi_i \cap \Phi_j = \emptyset \;\text{ for }\; i \neq j. \tag{3.36}$$

The random partition $\boldsymbol{\Phi}$ is constructed from an infinite sequence of DP samples $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots$ by including in Φ_l all $n \in \mathbb{N}$ for which $\boldsymbol{\theta}_n = \boldsymbol{\theta}_l^*$, i.e.,

$$n \in \Phi_l \;\text{ if }\; \boldsymbol{\theta}_n = \boldsymbol{\theta}_l^*. \tag{3.37}$$

By this construction, each $n \in \mathbb{N}$ belongs to exactly one subset Φ_l, and thus all Φ_l are disjoint and their union is $\mathbb{N}$, i.e., we obtain a partition of $\mathbb{N}$; furthermore, all the samples $\boldsymbol{\theta}_n$, $n \in \mathbb{N}$ that equal $\boldsymbol{\theta}_l^*$ are associated with the l^{th} cluster Φ_l. Thus, our clustering of the indices $n \in \mathbb{N}$ into subsets Φ_l corresponds to a clustering of the DP samples $\boldsymbol{\theta}_n$, $n \in \mathbb{N}$. Since both $\boldsymbol{\theta}_l^*$ and $\boldsymbol{\theta}_n$ are random, the partition $\boldsymbol{\Phi}$ defined by (3.37) is random as well. For

a characterization of its probability distribution, we obtain using (3.11)

$$\mathbb{P}\big(n \in \phi_l \,|\, (\theta_l^*)_{l=1}^{\infty} = (\boldsymbol{\theta}_l^*)_{l=1}^{\infty}, (\mathsf{Q}_{l'})_{l'=1}^{\infty} = (Q_{l'})_{l'=1}^{\infty}\big) = Q_l. \tag{3.38}$$

This means that the probability distribution of $\boldsymbol{\Phi}$ is induced only by the weights $(\mathsf{Q}_l)_{l=1}^{\infty}$ and is independent of the values of the positions $(\theta_l^*)_{l=1}^{\infty}$ [21, pg.17].

3.3.2 Cluster Assignment Variables

The assignment of the samples θ_n to distinct positions θ_l^* (or equivalently the association of the θ_n with clusters ϕ_l) can be expressed in terms of cluster assignment variables $\mathsf{C}_1, \mathsf{C}_2, \ldots \in \mathbb{N}$ by setting

$$\mathsf{C}_n = l \quad \text{if} \quad \theta_n = \theta_l^*. \tag{3.39}$$

Note that we can also write

$$\theta_n = \theta_{\mathsf{C}_n}^*. \tag{3.40}$$

By (3.37), $\mathsf{C}_n = l$ implies $n \in \phi_l$, and vice versa, i.e.,

$$n \in \phi_l \quad \text{if and only if} \quad \mathsf{C}_n = l. \tag{3.41}$$

In other words, for each $n \in \mathbb{N}$, C_n equals the label l of the subset (cluster) ϕ_l to which n (or by association, θ_n) belongs. This implies that the random sequence of cluster assignment variables $\mathsf{C}_1, \mathsf{C}_2, \ldots$ is equivalent to the random partition $\boldsymbol{\Phi} = (\phi_1, \phi_2, \ldots)$ of $\mathbb{N}$. We note that the order of the C_n is different from the order of the partition $\boldsymbol{\Phi} = (\phi_1, \phi_2, \ldots)$ since according to (3.39) we have $\mathsf{C}_n = l$ if $\theta_n = \theta_l^*$. This means that since θ_1 is not necessarily equal to θ_1^*, C_1 is not necessarily equal to 1, but may be any $l \in \mathbb{N}$. In the last paragraph of the current subsection, we will introduce an alternative definition of the cluster assignment variables using the ordering implied by the ϑ_s^*.

We recall from (3.11) that $\theta_n = \theta_l^*$ with probability Q_l. Furthermore, $\theta_n = \theta_l^*$ is equivalent to $\mathsf{C}_n = l$. Therefore, conditioned on $\mathsf{f}_{\mathrm{DP}} = f_{\mathrm{DP}}$, or equivalently conditioned on a sequence of positions $(\theta_{l'}^*)_{l'=1}^{\infty} = (\boldsymbol{\theta}_{l'}^*)_{l'=1}^{\infty}$ and weights $(\mathsf{Q}_{l'})_{l'=1}^{\infty} = (Q_{l'})_{l'=1}^{\infty}$, we have

$$\mathbb{P}\big(\mathsf{C}_n = l \,|\, (\theta_{l'}^*)_{l'=1}^{\infty} = (\boldsymbol{\theta}_{l'}^*)_{l'=1}^{\infty}, (\mathsf{Q}_{l'})_{l'=1}^{\infty} = (Q_{l'})_{l'=1}^{\infty}\big) = Q_l, \tag{3.42}$$

or equivalently formulated in terms of probability mass function (pmf),

$$p_{\mathsf{C}_n \,|\, (\theta^*_{l'})_{l'=1}^{\infty}, (\mathsf{Q}_{l'})_{l'=1}^{\infty}}(l \,|\, (\boldsymbol{\theta}^*_{l'})_{l'=1}^{\infty}, (Q_{l'})_{l'=1}^{\infty}) = Q_l. \tag{3.43}$$

Moreover, we conclude from (3.40) that, together with the sequence of positions $(\theta^*_l)_{l=1}^{\infty}$, the cluster assignment variables C_n are probabilistically equivalent to the DP samples θ_n. Therefore, since the θ_n are conditionally i.i.d. given $(\theta^*_l)_{l=1}^{\infty} = (\boldsymbol{\theta}^*_l)_{l=1}^{\infty}$ and $(\mathsf{Q}_l)_{l=1}^{\infty} = (Q_l)_{l=1}^{\infty}$, also the cluster assignment variables C_n are conditionally i.i.d. given $(\theta^*_l)_{l=1}^{\infty} = (\boldsymbol{\theta}^*_l)_{l=1}^{\infty}$ and $(\mathsf{Q}_l)_{l=1}^{\infty} = (Q_l)_{l=1}^{\infty}$ (equivalently, given $\mathsf{f}_{\mathrm{DP}} = f_{\mathrm{DP}}$).

Lastly, as evidenced by the right-hand side of (3.43), the cluster assignment variables C_n are conditionally independent of the positions $(\theta^*_l)_{l=1}^{\infty}$ given the weights $(\mathsf{Q}_l)_{l=1}^{\infty} = (Q_l)_{l=1}^{\infty}$. The weights $(\mathsf{Q}_l)_{l=1}^{\infty}$ are by definition independent of the positions $(\theta^*_l)_{l=1}^{\infty}$ (see 3.1.1).

Generation of Random Vectors θ_n

The two-step generation of the random vectors (samples) θ_n, $n \in \mathbb{N}$ described in Section 3.1.2 can be reformulated in terms of the cluster assignment variables C_n as follows:

1. As before, we generate a DP by generating a random sequence of positions $(\theta^*_l)_{l=1}^{\infty}$ that are distributed i.i.d. according to a base distribution f_{H} (see (3.1)), and a random sequence of weights $(\mathsf{Q}_l)_{l=1}^{\infty}$ that are distributed according to the GEM distribution with parameter α (see (3.4)). Hence, the DP is given by the random pdf (see (3.6))

$$\mathsf{f}_{\mathrm{DP}}(\boldsymbol{\theta}) = \sum_{l=1}^{\infty} \mathsf{Q}_l \delta_{\theta^*_l}(\boldsymbol{\theta}). \tag{3.44}$$

2. Next, we generate cluster assignment variables C_n, $n \in \mathbb{N}$ conditionally i.i.d. given $\mathsf{f}_{\mathrm{DP}} = f_{\mathrm{DP}}$ (or equivalently given $(\theta^*_l)_{l=1}^{\infty} = (\boldsymbol{\theta}^*_l)_{l=1}^{\infty}$ and $(\mathsf{Q}_l)_{l=1}^{\infty} = (Q_l)_{l=1}^{\infty}$), using the conditional pmf $p_{\mathsf{C}_n \,|\, (\theta^*_{l'})_{l'=1}^{\infty}, (\mathsf{Q}_{l'})_{l'=1}^{\infty}}(l \,|\, (\boldsymbol{\theta}^*_{l'})_{l'=1}^{\infty}, (Q_{l'})_{l'=1}^{\infty}) = Q_l$ (see (3.43)).

3. Finally, for each $n \in \mathbb{N}$, we set the random vector θ_n equal to the position θ^*_l with $l = \mathsf{C}_n$, i.e., the position determined by the cluster assignment variable C_n:

$$\theta_n = \theta^*_{\mathsf{C}_n}, \quad n \in \mathbb{N}. \tag{3.45}$$

Similarly to $\tilde{m}_s(N)$ defined in (3.30), we define $m_l(N)$ as the number of occurrences of

$\boldsymbol{\theta}_l^*$ in the sample sequence $\boldsymbol{\theta}_{1:N}$, i.e.,

$$m_l(N) \triangleq \sum_{n=1}^{N} \mathbb{1}(\boldsymbol{\theta}_n = \boldsymbol{\theta}_l^*). \tag{3.46}$$

Using the cluster assignment variables C_n, this can be expressed as

$$m_l(N) = \sum_{n=1}^{N} \mathbb{1}(C_n = l), \quad l \in \{C_1, \ldots, C_N\}. \tag{3.47}$$

Note that in the sequence $(C_1, \ldots, C_N)$, some of the C_n may be equal and that $m_l(N) = 0$ is not possible since $l \in \{C_1, \ldots, C_N\}$. Furthermore, $m_l(N)$ is a permutated version of $\tilde{\mathsf{m}}_s(N)$:

$$\mathsf{m}_l(N) = \tilde{\mathsf{m}}_{\sigma(l)}(N) = \tilde{\mathsf{m}}_s(N), \tag{3.48}$$

with same permutation $s = \sigma(l)$ as in (3.15). It will be convenient to define the random vector of cluster assignment variables

$$\mathbf{C}_{1:N} \triangleq (\mathsf{C}_1, \ldots, \mathsf{C}_N)^{\mathrm{T}} \tag{3.49}$$

and the random set[3]

$$\mathscr{C}(N) = \{\mathbf{C}_{1:N}\} \triangleq \{\mathsf{C}_1, \ldots, \mathsf{C}_N\}, \tag{3.50}$$

which is the set containing all unique cluster assignment variables C_n, for $n = 1, \ldots, N$. Similarly, for $\mathbf{C}_{1:N} = \boldsymbol{C}_{1:N}$ we define

$$\mathcal{C}(N) = \{\boldsymbol{C}_{1:N}\} \triangleq \{C_1, \ldots, C_N\}, \tag{3.51}$$

as the set containing all unique observed cluster assignment variables C_n, for $n = 1, \ldots, N$. In analogy to (3.31), we obtain

$$\sum_{l \in \mathcal{C}(N)} m_l(N) = N. \tag{3.52}$$

Furthermore, based on (3.50), we define

$$\mathbf{m}_{\mathscr{C}(N)} \triangleq (\mathsf{m}_l)_{l \in \mathscr{C}(N)} \tag{3.53}$$

[3]Note that in (3.50), and also subsequently, the notation $\{\boldsymbol{a}\}$ expresses the set composed of the components of the vector $\boldsymbol{a}$.

as the vector of all cluster sizes m_l, ordered in any convenient way, for example ascending. Similarly, we also define

$$\boldsymbol{\theta}^*_{\mathscr{C}(N)} \triangleq (\boldsymbol{\theta}^*_l)_{l\in\mathscr{C}(N)} \tag{3.54}$$

as the vector of all distinct $\boldsymbol{\theta}^*_l$ within the samples $\boldsymbol{\theta}_{1:N}$, again ordered in any convenient way. We note that since the positions $\boldsymbol{\theta}^*_l$ are i.i.d. (see (3.1)), the conditional joint pdf of the positions $\boldsymbol{\theta}^*_{\mathscr{C}(N)}$ given $\mathbf{C}_{1:N} = \boldsymbol{C}_{1:N}$ or, equivalently, $\mathscr{C}(N) = \mathcal{C}(N)$, can be written as

$$f_{\boldsymbol{\theta}^*_{\mathscr{C}(N)}\,|\,\mathbf{C}_{1:N}}(\boldsymbol{\theta}^*_{\mathcal{C}(N)}\,|\,\boldsymbol{C}_{1:N}) = f_{\boldsymbol{\theta}^*_{\mathscr{C}(N)}\,|\,\mathcal{C}(N)}(\boldsymbol{\theta}^*_{\mathcal{C}(N)}\,|\,\mathcal{C}(N)) = \prod_{l\in\mathcal{C}(N)} f_{\mathrm{H}}(\boldsymbol{\theta}^*_l). \tag{3.55}$$

The predictive pdf in (3.33) can be rewritten in terms of the set of cluster assignment variables $\mathcal{C}(n-1) = \{C_1, \ldots, C_{n-1}\}$ as

$$f_{\boldsymbol{\theta}_n\,|\,\boldsymbol{\theta}_{1:n-1}}(\boldsymbol{\theta}_n\,|\,\boldsymbol{\theta}_{1:n-1}) = \frac{\alpha}{\alpha+n-1} f_{\mathrm{H}}(\boldsymbol{\theta}_n) + \frac{1}{\alpha+n-1}\sum_{l\in\mathcal{C}(n-1)} m_l(n-1)\delta_{\boldsymbol{\theta}^*_l}(\boldsymbol{\theta}_n). \tag{3.56}$$

Similarly, the joint pdf in (3.32) can be rewritten as

$$f_{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_N}(\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_N) = f_{\mathrm{H}}(\boldsymbol{\theta}_1)\prod_{n=2}^{N}\left(\frac{\alpha}{\alpha+n-1} f_{\mathrm{H}}(\boldsymbol{\theta}_n) + \frac{1}{\alpha+n-1}\sum_{l\in\mathcal{C}(n-1)} m_l(n-1)\delta_{\boldsymbol{\theta}^*_l}(\boldsymbol{\theta}_n)\right). \tag{3.57}$$

Cluster Assignment Variables Based on Empirical Reordering

Finally, we introduce a modified definition of the cluster assignment variables C_n in which the indices l are replaced by the empirically ordered indices s. That is, similarly to (3.39), we now define the cluster assignment variable $\tilde{\mathsf{C}}_n$ using $\boldsymbol{\vartheta}^*_s$, i.e.,

$$\tilde{\mathsf{C}}_n = s \quad \text{if} \quad \boldsymbol{\theta}_n = \boldsymbol{\vartheta}^*_s. \tag{3.58}$$

Note that this can equivalently be written as (see (3.45))

$$\boldsymbol{\theta}_n = \boldsymbol{\vartheta}^*_{\tilde{\mathsf{C}}_n}, \quad n\in\mathbb{N}. \tag{3.59}$$

The number $\tilde{m}_s(N)$ of occurrences of $\boldsymbol{\vartheta}_s^*$ in the sample sequence $\boldsymbol{\theta}_{1:N}$, previously defined in (3.30), can now be expressed using the cluster assignment variables $\tilde{\mathsf{C}}_n$ as (see (3.47))

$$\tilde{m}_s(N) = \sum_{n=1}^{N} \mathbb{1}(\tilde{C}_n = s), \quad s = 1, \ldots, S(N). \tag{3.60}$$

For completeness, we define the random vector of cluster assignment variables (see (3.49))

$$\tilde{\mathbf{C}}_{1:N} \triangleq (\tilde{\mathsf{C}}_1, \ldots, \tilde{\mathsf{C}}_N)^{\mathrm{T}} \tag{3.61}$$

and the random set (see (3.50))

$$\tilde{\mathscr{C}}(N) = \{\tilde{\mathbf{C}}_{1:N}\} \triangleq \{\tilde{\mathsf{C}}_1, \ldots, \tilde{\mathsf{C}}_N\} = \{1, \ldots, \mathsf{S}(N)\}, \tag{3.62}$$

which is the set containing all unique cluster assignment variables $\tilde{\mathsf{C}}_n$, for $n = 1, \ldots, N$. Finally, based on (3.54), we define

$$\boldsymbol{\vartheta}^*_{\tilde{\mathscr{C}}(N)} \triangleq (\boldsymbol{\vartheta}_s^*)_{s \in \tilde{\mathscr{C}}(N)} \tag{3.63}$$

3.3.3 Chinese Restaurant Process

We have mentioned that the DP, with continuous f_{H}, induces a random partition of $\mathbb{N}$. The distribution over the partitions is called a Chinese restaurant process (CRP). The name comes from the analogy to seating customers in a restaurant that has an infinite number of tables and space for an infinite number of customers. Each time a new customer enters the restaurant, he/she sits either at an already occupied table or at an empty table.

Let us reconsider the discrete part $\frac{\alpha}{\alpha+n-1} \sum_{s=1}^{S(n-1)} \tilde{m}_s(n-1)\delta_{\boldsymbol{\vartheta}_s^*}(\boldsymbol{\theta}_n)$ of the predictive pdf $f_{\boldsymbol{\theta}_n \,|\, \boldsymbol{\theta}_{1:n-1}}(\boldsymbol{\theta}_n \,|\, \boldsymbol{\theta}_{1:n-1})$ in (3.33). We see that the probability that the next sample $\boldsymbol{\theta}_n$ equals $\boldsymbol{\vartheta}_s^*$ increases with $\tilde{m}_s(n-1)$, i.e., the number of times $\boldsymbol{\vartheta}_s^*$ has already been observed. For example, in Figure 5a, out of the $n - 1 = 10$ samples that were observed, there are only two distinct positions $\boldsymbol{\vartheta}_s^*$ or, equivalently, clusters, whereas in Figure 5c, there are nine distinct positions $\boldsymbol{\vartheta}_s^*$ or clusters, i.e., only one cluster contains more than one sample. If we were to sample from the predictive distribution shown in Figure 5a, the next sample would most likely be equal to one of the two previous samples. On the other hand, in the case of Figure 5c, the next sample would most likely be sampled from the base distribution. This

is again a manifestation of the *the rich get richer* property of the DP.

Suppose now that $\boldsymbol{\theta}_1$ represents a customer that enters an empty restaurant ($n-1=0$) and sits down at the first table. This customer is therefore assigned to cluster $\tilde{\mathsf{C}}_1 = 1$. Next, the second customer $\boldsymbol{\theta}_2$ can either sit down at the already occupied table or at a previously empty table, therefore there are two possible cluster assignments, $\tilde{\mathsf{C}}_2 = 1$ or $\tilde{\mathsf{C}}_2 = 2$. Continuing in this manner, let $\boldsymbol{\theta}_n$ represent a customer entering the restaurant with $n-1$ customers $\boldsymbol{\theta}_{1:n-1}$ already inside and $S(n-1)$ tables occupied. The new customer chooses to sit at the already occupied table $s \in \{1, \ldots, S(n-1)\}$ with probability $\frac{\tilde{m}_s(n-1)}{\alpha+n-1}$, or he/she sits at an empty table $s = S(n-1)+1$ with probability $\frac{\alpha}{\alpha+n-1}$. The customer is assigned a cluster $\tilde{\mathsf{C}}_n = s$ according to his/her table. The predictive pmf of $\tilde{\mathsf{C}}_n$, i.e., given the previous cluster assignment variables $\tilde{\mathbf{C}}_{1:n-1} = (\tilde{\mathsf{C}}_1, \ldots, \tilde{\mathsf{C}}_{n-1})^{\mathrm{T}}$, is thus obtained as

$$p_{\tilde{\mathsf{C}}_n \,|\, \tilde{\mathbf{C}}_{1:n-1}}(s \,|\, \tilde{\boldsymbol{C}}_{1:n-1}) = \begin{cases} \dfrac{\sum_{n'=1}^{n-1} \mathbb{1}(\tilde{C}_{n'} = s)}{\alpha + n - 1} & \text{for} \quad s = 1, \ldots, S(n-1) \quad \text{and} \quad n \geq 2 \\ \dfrac{\alpha}{\alpha + n - 1} & \text{for} \quad s = S(n-1)+1 \qquad \text{and} \quad n \geq 2, \end{cases} \tag{3.64}$$

where (3.60) was used. For the case $n = 1$, we obtain $p_{\tilde{\mathsf{C}}_1}(s) = 1$.

CRP-based Generation of the Samples $\boldsymbol{\theta}_n$

The recursive construction presented in Section 3.2.2 can similarly be formulated using the cluster assignment variables $\tilde{\mathsf{C}}_n$ and the predictive pmf of the cluster assignment variables in (3.64), thus establishing a relation to the CRP:

1. Draw the distinct random positions $\boldsymbol{\vartheta}_s^*$, $s \in \mathbb{N}$ i.i.d. from the base distribution f_{H}, i.e.,

$$\boldsymbol{\vartheta}_1^*, \boldsymbol{\vartheta}_2^*, \ldots \sim_{\text{i.i.d.}} f_{\mathrm{H}}. \tag{3.65}$$

2. Initialize the recursion by choosing the first cluster assignment $\tilde{\mathsf{C}}_1$ as

$$\tilde{\mathsf{C}}_1 = 1, \tag{3.66}$$

and for $n = 2, 3 \ldots$, draw the next cluster assignment variable $\tilde{\mathsf{C}}_n$ from the predictive pmf $p_{\tilde{\mathsf{C}}_n \,|\, \tilde{\mathbf{C}}_{1:n-1}}(s \,|\, \tilde{\boldsymbol{C}}_{1:n-1})$ in (3.64).

3. For all $n \in \mathbb{N}$, set

$$\boldsymbol{\theta}_n = \boldsymbol{\vartheta}^*_{\tilde{\mathsf{C}}_n}. \tag{3.67}$$

We note that using this construction, we have $\tilde{\mathsf{C}}_1 = 1$ and therefore also $\boldsymbol{\theta}_1 = \boldsymbol{\vartheta}^*_1$, as postulated previously (see (3.12)). A remarkable feature of this recursive construction is that it does not involve the weights Q_l.

CRP-induced Random Partition

Using an infinite sequence of DP samples $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots$ generated according to the above-described recursion, we can construct an ordered random partition $\tilde{\boldsymbol{\phi}} = (\tilde{\phi}_1, \tilde{\phi}_2, \ldots)$ of $\mathbb{N}$ by including in the subset (cluster) $\tilde{\phi}_s$ all $n \in \mathbb{N}$ for which $\boldsymbol{\theta}_n = \boldsymbol{\vartheta}^*_s$, i.e.,

$$n \in \tilde{\phi}_s \quad \text{if} \quad \boldsymbol{\theta}_n = \boldsymbol{\vartheta}^*_s. \tag{3.68}$$

This random partition $\tilde{\boldsymbol{\phi}}$ is equivalent to the sequence of cluster assignment variables $(\tilde{\mathsf{C}}_n)_{n=1}^{\infty}$ because $\tilde{\phi}_s$ comprises all $n \in \mathbb{N}$ for which $\tilde{\mathsf{C}}_n = s$, i.e.,

$$n \in \tilde{\phi}_s \quad \text{if and only if} \quad \tilde{\mathsf{C}}_n = s. \tag{3.69}$$

Note the analogy of (3.68) and (3.69) to (3.37) and (3.41), respectively. For $n = 1$ and $s = 1$, (3.69) reads $1 \in \tilde{\phi}_1$ if and only if $\tilde{\mathsf{C}}_1 = 1$, and thus we conclude from (3.66) that $1 \in \tilde{\phi}_1$.

According to (3.69), the CRP induces a partition of $\mathbb{N}$ that is independent of the positions $\boldsymbol{\vartheta}^*_s$. This is because the predictive pmf $p_{\tilde{\mathsf{C}}_n \,|\, \tilde{\mathsf{c}}_{1:n-1}}(s \,|\, \tilde{\boldsymbol{C}}_{1:n-1})$ in (3.64) that is used in the recursive generation of the infinite sequence of cluster assignment variables $(\tilde{\mathsf{C}}_n)_{n=1}^{\infty}$ does not depend on the positions $\boldsymbol{\vartheta}^*_s$.

Joint pmf of the Cluster Assignment Variables

Next, we consider the joint pmf of the cluster assignment variables $\tilde{\mathbf{C}}_{1:N}$. By applying the chain rule, this can be formulated as a product of the conditional pmfs $p_{\tilde{\mathsf{C}}_n | \tilde{\mathsf{c}}_{1:n-1}}(s | \tilde{\boldsymbol{C}}_{1:n-1})$ (see (3.64)), i.e.,

$$p_{\tilde{\mathbf{C}}_{1:N}}(\tilde{\boldsymbol{C}}_{1:N}) = p_{\tilde{\mathsf{C}}_1}(\tilde{C}_1) \prod_{n=2}^{N} p_{\tilde{\mathsf{C}}_n \,|\, \tilde{\mathsf{c}}_{1:n-1}}(\tilde{C}_n \,|\, \tilde{\boldsymbol{C}}_{1:n-1}). \tag{3.70}$$

Following the results in [29, Eq. 8], the joint pmf of the cluster assignment variables, conditioned on the respective cluster sizes $\tilde{\mathbf{m}}_{1:S(N)} \triangleq (\tilde{\mathsf{m}}_1(N), \ldots, \tilde{\mathsf{m}}_{S(N)}(N))^{\mathrm{T}}$ is given by

$$p_{\tilde{\mathbf{C}}_{1:N} \,|\, \tilde{\mathbf{m}}_{1:S(N)}}(\tilde{\boldsymbol{C}}_{1:N} \,|\, \tilde{\boldsymbol{m}}_{1:S(N)}) = \frac{\alpha^{S(N)} \prod_{s=1}^{S(N)} (\tilde{m}_s(N) - 1)!}{\prod_{n=1}^{N} (\alpha + n - 1)}. \tag{3.71}$$

We note that this pmf depends only on the respective cluster sizes $\tilde{\mathbf{m}}_{1:S(N)}$, the total number of distinct objects $S(N)$, and the concentration parameter α. On the other hand, it does not depend on the order of the cluster assignment variables $\tilde{C}_n$, which means the cluster assignment variables are conditionally exchangeable given $\tilde{\mathbf{m}}_{1:S(N)}$ [29].

Exchangeable Random Partition

Let us consider the ordered random partition $\boldsymbol{\Phi} = (\Phi_1, \Phi_2, \ldots)$ of $\mathbb{N}$ constructed from an infinite sequence of DP samples $(\boldsymbol{\Theta}_n)_{n=1}^{\infty}$ according to (3.37). From the fact that the DP samples $(\boldsymbol{\Theta}_n)_{n=1}^{\infty}$ are exchangeable, it follows that the random partition $\boldsymbol{\Phi}$ is also exchangeable. This is proven using Kingman's representation [30], which shows that for any exchangeable infinite random partition, there exists an exchangeable sequence defined by (3.10) that generates the infinite partition (see (3.37)) [5, Theorem 14.7].

Next, we reconsider the index transformation $s = \sigma(l), l = 1, 2, \ldots$ in (3.15). Using (3.14) in (3.68), we have

$$n \in \tilde{\Phi}_{\sigma(l)} \quad \text{if} \quad \boldsymbol{\Theta}_n = \boldsymbol{\vartheta}^*_{\sigma(l)} = \boldsymbol{\Theta}^*_l. \tag{3.72}$$

Comparing (3.72) with (3.37) we conclude that

$$\tilde{\Phi}_{\sigma(l)} = \Phi_l, \tag{3.73}$$

which implies that the random partition $\tilde{\boldsymbol{\Phi}}$ is a reindexed version of $\boldsymbol{\Phi}$, i.e.,

$$(\tilde{\Phi}_{\sigma(1)}, \tilde{\Phi}_{\sigma(2)}, \ldots) = (\Phi_1, \Phi_2, \ldots). \tag{3.74}$$

Furthermore, using (3.73) in (3.69) and recalling that $s = \sigma(l)$, we obtain

$$n \in \Phi_l \quad \text{if and only if} \quad \tilde{\mathsf{C}}_n = \sigma(l). \tag{3.75}$$

Comparing with (3.41), we conclude that

$$\tilde{\mathsf{C}}_n = \sigma(l) = \sigma(\mathsf{C}_n). \tag{3.76}$$

Thus, the cluster assignment variable $\tilde{\mathsf{C}}_n$ is related to the cluster assignment variable C_n by the permutation function $\sigma(\cdot)$.

The random partition $\boldsymbol{\Phi} = (\Phi_1, \Phi_2, \ldots)$ implied by the random sequence of cluster assignment variables $(\mathsf{C}_n)_{n=1}^{\infty}$ (see (3.41)) is (up to reindexing) equivalent to the random partition $\tilde{\boldsymbol{\Phi}} = (\tilde{\Phi}_1, \tilde{\Phi}_2, \ldots)$ implied by the random sequence of cluster assignment variables $(\tilde{\mathsf{C}}_n)_{n=1}^{\infty}$ (see (3.69)).

We note that the process of generating the cluster assignments by sampling from the predictive pmf in (3.64) is called *recursive partitioning* [5, caption of Figure 14.1], and the conditional pmf of the cluster assignment variables in (3.71) is also called the *exchangeable partition probability function* [5, Eq. 14.6].

3.4 Dirichlet Process Mixture

The DP is widely used across a great variety of applications in Bayesian analysis, e.g., in density estimation and data clustering. Due to its discrete nature, i.e., realizations of the DP are discrete probability distributions, the DP alone is not suitable as a prior for estimating a continuous density. However, the DP can be used as a prior distribution in a mixture model. The resulting mixture model is referred to as a Dirichlet process mixture (DPM).

At the beginning of this chapter, we considered the DPM in the context of the simple example of estimating the weight distribution of people from a set of measurements. Let us denote the weight of person n by $\mathsf{x}_n \in \mathbb{R}^+$. We argued previously that in a given population, there are several different groups of people (*clusters*). Since it is very unlikely that two people have exactly the same weight, even within the same cluster, the DP cannot be used as a prior for the weight distribution.

Let the random position $\theta_l^* \in \mathbb{R}^+$ represent the mean weight of the l^{th} cluster. Using the cluster assignment variable C_n to assign person n to a cluster l, i.e., $\mathsf{l} = \mathsf{C}_n$, we have $\theta_n = \theta^*_{\mathsf{C}_n}$ for all n (here, θ_n denotes the mean weight of the cluster to which person n belongs). Since the weight distribution we want to estimate is continuous whereas the DP is discrete, we do not use θ_n to model the weight x_n, but instead, we use a continuous

pdf $f_{\mathsf{x}_n \mid \theta_n}(x_n \mid \theta_n) = \phi(x_n \mid \theta_n)$ to *smooth out* the DP. For example, we might consider using $\phi(x_n \mid \theta_n) = \mathcal{N}(x_n; \theta_n, \sigma^2)$. However, since this allows $x_n \leq 0$, we use for $\phi(x_n \mid \theta_n)$ a truncated Gaussian distribution, with mean θ_n and some variance $\sigma^2 > 0$, where the truncation enforces $\mathsf{x}_n > 0$. The advantage of the DPM model is that we do not need to specify the number of clusters beforehand, i.e., the number of distinct weight means θ_l^*, as the underlying DP offers an infinite number of θ_l^*, which are automatically chosen depending on the observed data and the concentration parameter α.

Definition of the DPM

We now formally define the DPM, following [6] and [5, Sec. 5.1]. Consider a random position $\boldsymbol{\theta}_n \in \mathbb{R}^P$ distributed according to the DP, i.e.,

$$\boldsymbol{\theta}_n \mid (\mathsf{f}_{\mathrm{DP}} = f_{\mathrm{DP}}) \sim_{\text{i.i.d.}} f_{\mathrm{DP}}, \tag{3.77}$$

where $f_{\mathrm{DP}} \sim \mathrm{DP}(\alpha, f_{\mathrm{H}})$ with some concentration parameter $\alpha > 0$ and base pdf $f_{\mathrm{H}}(\boldsymbol{\theta})$ for $\boldsymbol{\theta} \in \mathbb{R}^P$. In addition, consider a random vector $\mathbf{x}_n \in \mathbb{R}^D$, and let the conditional distribution of $\mathbf{x}_n$ given $\boldsymbol{\theta}_n$ be described by a continuous pdf $\phi(\boldsymbol{x} \mid \boldsymbol{\theta})$ on $\mathbb{R}^D$ for each $\boldsymbol{\theta} \in \mathbb{R}^P$, i.e.,

$$f_{\mathbf{x}_n \mid \boldsymbol{\theta}_n}(\boldsymbol{x}_n \mid \boldsymbol{\theta}_n) = \phi(\boldsymbol{x}_n \mid \boldsymbol{\theta}_n). \tag{3.78}$$

The DPM is now defined as the random pdf $\phi(\boldsymbol{x}_n \mid \boldsymbol{\theta}_n)$ with random condition variable $\boldsymbol{\theta}_n$ distributed according to (3.77). Consistent with this hierarchical model for generating $\boldsymbol{\theta}_n$ from f_{DP} and $\mathbf{x}_n$ from $\boldsymbol{\theta}_n$, we further assume that $\mathbf{x}_n$ is conditionally independent of f_{DP} given $\boldsymbol{\theta}_n$, i.e.,

$$f_{\mathbf{x}_n \mid \mathsf{f}_{\mathrm{DP}}, \boldsymbol{\theta}_n}(\boldsymbol{x}_n \mid f_{\mathrm{DP}}, \boldsymbol{\theta}_n) = f_{\mathbf{x}_n \mid \boldsymbol{\theta}_n}(\boldsymbol{x}_n \mid \boldsymbol{\theta}_n). \tag{3.79}$$

Note that this corresponds to a *Markov chain* $\mathsf{f}_{\mathrm{DP}} \to \boldsymbol{\theta}_n \to \mathbf{x}_n$.

The conditional distribution of $\mathbf{x}_n$ given f_{DP} can be calculated as [6, Eq. 3.34]

$$\begin{aligned} f_{\mathbf{x}_n \mid \mathsf{f}_{\mathrm{DP}}}(\boldsymbol{x}_n \mid f_{\mathrm{DP}}) &= \int_{\boldsymbol{\theta}_n} f_{\mathbf{x}_n \mid \mathsf{f}_{\mathrm{DP}}, \boldsymbol{\theta}_n}(\boldsymbol{x}_n \mid f_{\mathrm{DP}}, \boldsymbol{\theta}_n) f_{\boldsymbol{\theta}_n \mid \mathsf{f}_{\mathrm{DP}}}(\boldsymbol{\theta}_n \mid f_{\mathrm{DP}}) d\boldsymbol{\theta}_n \\ &= \int_{\boldsymbol{\theta}_n} f_{\mathbf{x}_n \mid \boldsymbol{\theta}_n}(\boldsymbol{x}_n \mid \boldsymbol{\theta}_n) f_{\mathrm{DP}}(\boldsymbol{\theta}_n) d\boldsymbol{\theta}_n \\ &= \int_{\boldsymbol{\theta}_n} \phi(\boldsymbol{x}_n \mid \boldsymbol{\theta}_n) f_{\mathrm{DP}}(\boldsymbol{\theta}_n) d\boldsymbol{\theta}_n, \end{aligned} \tag{3.80}$$

where we used the law of total probability as well as (3.79), (3.77), and (3.78). We recall that a realization $\mathsf{f}_{\mathrm{DP}} = f_{\mathrm{DP}}$ corresponds to a realization of the position sequence $(\boldsymbol{\theta}_l^*)_{l=1}^{\infty} = (\boldsymbol{\theta}_l^*)_{l=1}^{\infty}$ and a realization of the weight sequence $(\mathsf{Q}_l)_{l=1}^{\infty} = (Q_l)_{l=1}^{\infty}$, and by (3.6) we have

$$f_{\mathrm{DP}}(\boldsymbol{\theta}) = \sum_{l=1}^{\infty} Q_l \delta_{\boldsymbol{\theta}_l^*}(\boldsymbol{\theta}), \tag{3.81}$$

for some $(\boldsymbol{\theta}_l^*)_{l=1}^{\infty}$ and $(Q_l)_{l=1}^{\infty}$. Using (3.81) in (3.80) finally yields

$$\begin{aligned} f_{\mathbf{x}_n \,|\, \mathsf{f}_{\mathrm{DP}}}(\boldsymbol{x}_n \,|\, f_{\mathrm{DP}}) &= \int_{\boldsymbol{\theta}_n} \phi(\boldsymbol{x}_n \,|\, \boldsymbol{\theta}_n) \left(\sum_{l=1}^{\infty} Q_l \delta_{\boldsymbol{\theta}_l^*}(\boldsymbol{\theta}_n) \right) d\boldsymbol{\theta}_n \\ &= \sum_{l=1}^{\infty} Q_l \int_{\mathbb{R}^P} \phi(\boldsymbol{x}_n \,|\, \boldsymbol{\theta}_n) \delta_{\boldsymbol{\theta}_l^*}(\boldsymbol{\theta}_n) d\boldsymbol{\theta}_n \\ &= \sum_{l=1}^{\infty} Q_l \phi(\boldsymbol{x}_n \,|\, \boldsymbol{\theta}_l^*). \end{aligned} \tag{3.82}$$

We note that (3.82) is an infinite mixture of continuous distributions $\phi(\boldsymbol{x}_n \,|\, \boldsymbol{\theta}_l^*)$, weighted by Q_l.

We can equivalently express the random condition variable $\boldsymbol{\theta}_n$ using the cluster assignment variable C_n and the random position $\boldsymbol{\theta}_l^*$ as $\boldsymbol{\theta}_n = \boldsymbol{\theta}_{\mathsf{C}_n}^*$ (see (3.40)), therefore the right-hand side of (3.79) can equivalently be written as

$$f_{\mathbf{x}_n \,|\, \boldsymbol{\theta}_n}(\boldsymbol{x}_n \,|\, \boldsymbol{\theta}_n) = f_{\mathbf{x}_n \,|\, \boldsymbol{\theta}_{\mathsf{C}_n}^*, \mathsf{C}_n}(\boldsymbol{x}_n \,|\, \boldsymbol{\theta}_{C_n}^*, C_n). \tag{3.83}$$

Finally, we consider a length-N sequence of random vectors $(\boldsymbol{\theta}_n)_{n=1}^N$, equivalently written as a vector $\boldsymbol{\theta}_{1:N} = (\boldsymbol{\theta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\theta}_N^{\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{\theta}_n$ is conditionally i.i.d. given $\mathsf{f}_{\mathrm{DP}} = f_{\mathrm{DP}}$ (see (3.77)). Furthermore, we consider a length-N sequence of random vectors $(\mathbf{x}_n)_{n=1}^N$, equivalently written as a vector $\mathbf{x}_{1:N} = (\mathbf{x}_1^{\mathrm{T}}, \ldots, \mathbf{x}_N^{\mathrm{T}})^{\mathrm{T}}$, where $\mathbf{x}_n$ is conditionally i.i.d. given $\boldsymbol{\theta}_{1:N}$, i.e.,

$$\mathbf{x}_n \,|\, (\boldsymbol{\theta}_{1:N} = \boldsymbol{\theta}_{1:N}) \sim_{\text{i.i.d.}} f_{\mathbf{x}_n \,|\, \boldsymbol{\theta}_{1:N}}(\boldsymbol{x}_n \,|\, \boldsymbol{\theta}_{1:N}). \tag{3.84}$$

In addition, we assume that $\mathbf{x}_n$ is conditionally independent of all other $\boldsymbol{\theta}_{n'}$ given $\boldsymbol{\theta}_n$, with $n' \neq n$. Thus, the joint conditional pdf can also be factorized as

$$f_{\mathbf{x}_{1:N} \,|\, \boldsymbol{\theta}_{1:N}}(\boldsymbol{x}_{1:N} \,|\, \boldsymbol{\theta}_{1:N}) = \prod_{n=1}^{N} f_{\mathbf{x}_n \,|\, \boldsymbol{\theta}_{1:N}}(\boldsymbol{x}_n \,|\, \boldsymbol{\theta}_{1:N}) = \prod_{n=1}^{N} f_{\mathbf{x}_n \,|\, \boldsymbol{\theta}_n}(\boldsymbol{x}_n \,|\, \boldsymbol{\theta}_n). \tag{3.85}$$

Equivalently, (3.85) can be expressed using the cluster assignment variable C_n and the random position $\boldsymbol{\theta}_l^*$ by inserting (3.83) into (3.85), hence

$$f_{\mathbf{x}_{1:N}\,|\,\boldsymbol{\theta}_{1:N}}(\boldsymbol{x}_{1:N}\,|\,\boldsymbol{\theta}_{1:N}) = \prod_{n=1}^{N} f_{\mathbf{x}_n\,|\,\boldsymbol{\theta}^*_{\mathsf{C}_n},\mathsf{C}_n}(\boldsymbol{x}_n\,|\,\boldsymbol{\theta}^*_{C_n}, C_n) = \prod_{l\in\mathcal{C}(N)}\prod_{n:C_n=l} f_{\mathbf{x}_n\,|\,\boldsymbol{\theta}^*_l,\mathsf{C}_n}(\boldsymbol{x}_n\,|\,\boldsymbol{\theta}^*_l, C_n), \tag{3.86}$$

with $\mathcal{C}(N) = \{C_1, \ldots, C_N\}$ being the set of all unique cluster assignment variables (see (3.51)) for $n = 1, \ldots, N$.

4 General Gaussian Model, Benchmark Scenarios and Estimators

This chapter will first introduce our statistical model and some fundamental assumptions that are valid for each of the four scenarios. We then discuss two basic scenarios that do not make use of any classification or clustering. These results will later be used as performance bounds in Section 5, where we study two more sophisticated scenarios that make use of joint clustering and estimation.

In each scenario, multiple objects are indexed by $n \in \{1, \dots, N\}$, where N is the total number of objects. For each object, we want to estimate a random parameter vector of interest $\mathbf{x}_n = (\mathsf{x}_{n,1}, \dots, \mathsf{x}_{n,D})^{\mathrm{T}} \in \mathbb{R}^D$, which is statistically dependent on a hyperparameter vector $\boldsymbol{\theta}_n = (\theta_{n,1}, \dots, \theta_{n,D})^{\mathrm{T}} \in \mathbb{R}^D$. Also, for each object, we observe a noisy measurement $\mathbf{y}_n = (\mathsf{y}_{n,1}, \dots, \mathsf{y}_{n,D})^{\mathrm{T}} \in \mathbb{R}^D$, which in all four scenarios is considered known and generated according to a stochastic dependence on our parameter of interest $\mathbf{x}_n$.

Thus, object n is associated with a single $\mathbf{x}_n$, a single $\boldsymbol{\theta}_n$, and a single $\mathbf{y}_n$. The task is to estimate the parameter of interest $\mathbf{x}_n$ for each object n, given the vector of all measurements $\mathbf{y}_{1:N} = (\mathbf{y}_1^{\mathrm{T}}, \dots, \mathbf{y}_N^{\mathrm{T}})^{\mathrm{T}}$. We denote the vector containing all hyperparameter vectors as $\boldsymbol{\theta}_{1:N} = (\boldsymbol{\theta}_1^{\mathrm{T}}, \dots, \boldsymbol{\theta}_N^{\mathrm{T}})^{\mathrm{T}}$ and the vector of all parameters of interest as $\mathbf{x}_{1:N} = (\mathbf{x}_1^{\mathrm{T}}, \dots, \mathbf{x}_N^{\mathrm{T}})^{\mathrm{T}}$.

Representative Scenarios

For each scenario, we choose a different prior pdf of the hyperparameter $\boldsymbol{\theta}_n$ for all $n \in \{1, \dots, N\}$. In general, we do not necessarily assume independence of $\boldsymbol{\theta}_n$ across the object index n, which means that the hyperparameter $\boldsymbol{\theta}_n$ can be statistically related not only to $\mathbf{x}_n$ but also to $\mathbf{x}_j$ with $j \neq n$. More specifically, we consider the following four scenarios:

1. $\boldsymbol{\theta}_n$ is i.i.d. across n and distributed according to a given prior $f_{\boldsymbol{\theta}_n}(\boldsymbol{\theta}_n)$.
2. $\boldsymbol{\theta}_n$ is distributed according to $f_{\boldsymbol{\theta}_n}(\boldsymbol{\theta}_n)$ as in first scenario, but observed, i.e. $\boldsymbol{\theta}_n = \boldsymbol{\theta}_n$.
3. $\boldsymbol{\theta}_{1:N}$ is distributed according to a DP as introduced in (3.7).
4. Similar to the previous scenario, $\boldsymbol{\theta}_{1:N}$ is distributed according to a DP; however, we assume that the cluster assignments $\mathbf{C}_{1:N}$ (see Sec. 3.3) are known.

We consider the first two scenarios as benchmarks providing theoretical performance bounds relative to the third and fourth scenarios. As opposed to the first two scenarios,

the third and fourth scenarios involve the DP prior, and they enable joint clustering and estimation. Since in the first scenario the hyperparameter $\boldsymbol{\theta}_n$ is i.i.d. across n, we cannot make use of any underlying cluster structure, and thus we expect the estimator in the first scenario to have the worst performance in terms of MSE. On the other hand, in the second scenario, the hyperparameter $\boldsymbol{\theta}_n = \boldsymbol{\theta}_n$ is assumed to be known, therefore the estimator in this scenario should have the best performance in terms of MSE. In the third scenario, we make use of the cluster structure, therefore we expect better performance in terms of MSE than in the first scenario but still poorer performance than in the second scenario. Furthermore, as we will observe in Section 5.2, the MMSE estimator in the third scenario cannot be expressed in closed form; however, it can be approximated by a Monte Carlo (MC) evaluation of the posterior mean. Lastly, in the fourth scenario, we also make use of the cluster structure; however, since the cluster assignments $\mathbf{C}_{1:N}$ are given, the estimator is provided with additional knowledge compared to the third scenario. Hence, we expect the estimator in the fourth scenario to perform better than the estimator in the third scenario, but still worse than the estimator in the second scenario.

4.1 General Model and Assumptions

We assume that the parameter of interest of each object n, $\mathbf{x}_n$, is related to the hyperparameter $\boldsymbol{\theta}_n$ according to the additive-noise model

$$\mathbf{x}_n = \boldsymbol{\theta}_n + \mathbf{u}_n, \tag{4.1}$$

where $\mathbf{u}_n$ will be called the parameter noise. We also assume that $\mathbf{u}_n$ is zero-mean Gaussian, i.e.,

$$f_{\mathbf{u}_n}(\boldsymbol{u}_n) = \mathcal{N}(\boldsymbol{u}_n; \mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{u}}), \tag{4.2}$$

with some covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{u}}$. It follows from (4.1) and (4.2) that given $\boldsymbol{\theta}_n$, $\mathbf{x}_n$ is Gaussian distributed according to

$$f_{\mathbf{x}_n \mid \boldsymbol{\theta}_n}(\boldsymbol{x}_n \mid \boldsymbol{\theta}_n) = \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\theta}_n, \boldsymbol{\Sigma}_{\boldsymbol{u}}). \tag{4.3}$$

Thus, conditioned on the hyperparameter $\boldsymbol{\theta}_n$, the mean of $\mathbf{x}_n$ equals $\boldsymbol{\theta}_n$ and the covariance matrix of $\mathbf{x}_n$ is $\boldsymbol{\Sigma}_{\boldsymbol{u}}$.

Furthermore, we assume that the measurement $\mathbf{y}_n$ is $\mathbf{x}_n$ corrupted by additive Gaussian

noise $\mathbf{v}_n$, i.e.,

$$\mathbf{y}_n = \mathbf{x}_n + \mathbf{v}_n, \tag{4.4}$$

where $\mathbf{v}_n$ is zero-mean Gaussian, i.e.,

$$f_{\mathbf{v}_n}(\boldsymbol{v}_n) = \mathcal{N}(\boldsymbol{v}_n; \mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{v}}), \tag{4.5}$$

with some covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{v}}$. It follows from (4.4) and (4.5) that $\mathbf{y}_n$ given $\mathbf{x}_n$ is Gaussian with mean $\boldsymbol{x}_n$ and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{v}}$, i.e,

$$f_{\mathbf{y}_n \,|\, \mathbf{x}_n}(\boldsymbol{y}_n \,|\, \boldsymbol{x}_n) = \mathcal{N}(\boldsymbol{y}_n; \boldsymbol{x}_n, \boldsymbol{\Sigma}_{\boldsymbol{v}}). \tag{4.6}$$

Note also that combining (4.1) and (4.4) gives

$$\mathbf{y}_n = \boldsymbol{\theta}_n + \mathbf{u}_n + \mathbf{v}_n. \tag{4.7}$$

We can also write (4.1), (4.4) and (4.7) using vector-matrix notation as

$$\begin{pmatrix} \mathbf{y}_n \\ \mathbf{x}_n \\ \boldsymbol{\theta}_n \end{pmatrix} = \begin{pmatrix} \boldsymbol{\theta}_n + \mathbf{u}_n + \mathbf{v}_n \\ \boldsymbol{\theta}_n + \mathbf{u}_n \\ \boldsymbol{\theta}_n \end{pmatrix} = \begin{pmatrix} \mathbf{I}_D & \mathbf{I}_D & \mathbf{I}_D \\ \mathbf{I}_D & \mathbf{I}_D & \mathbf{0} \\ \mathbf{I}_D & \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\theta}_n \\ \mathbf{u}_n \\ \mathbf{v}_n \end{pmatrix}, \tag{4.8}$$

with identity matrix $\mathbf{I}_D$ of size $D \times D$.

Independence Assumptions

For all scenarios, we furthermore assume the following:

A1) The parameter noise $\mathbf{u}_n$ is i.i.d. across object index n. Therefore, the pdf of the parameter noise vector of all objects, $\mathbf{u}_{1:N} = (\mathbf{u}_1^{\mathrm{T}}, \ldots, \mathbf{u}_N^{\mathrm{T}})^{\mathrm{T}}$, is

$$f_{\mathbf{u}_{1:N}}(\boldsymbol{u}_{1:N}) = \prod_{n=1}^{N} f_{\mathbf{u}_n}(\boldsymbol{u}_n) = \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{u}_n; \mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{u}}). \tag{4.9}$$

A2) The measurement noise $\mathbf{v}_n$ is i.i.d. across object index n. Therefore, the pdf of the

measurement noise vector of all objects, $\mathbf{v}_{1:N} = (\mathbf{v}_1^{\mathrm{T}}, \ldots, \mathbf{v}_N^{\mathrm{T}})^{\mathrm{T}}$, is

$$f_{\mathbf{v}_{1:N}}(\boldsymbol{v}_{1:N}) = \prod_{n=1}^{N} f_{\mathbf{v}_n}(\boldsymbol{v}_n) = \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{v}_n; \mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{v}}). \tag{4.10}$$

A3) $\boldsymbol{\theta}_n$, $\mathbf{u}_j$, and $\mathbf{v}_k$ are mutually independent for any $n, j, k \in \{1, \ldots, N\}$. This means that the joint pdf $f_{\boldsymbol{\theta}_{1:N}, \mathbf{u}_{1:N}, \mathbf{v}_{1:N}}(\boldsymbol{\theta}_{1:N}, \boldsymbol{u}_{1:N}, \boldsymbol{v}_{1:N})$ can be factored as

$$\begin{aligned} f_{\boldsymbol{\theta}_{1:N}, \mathbf{u}_{1:N}, \mathbf{v}_{1:N}}(\boldsymbol{\theta}_{1:N}, \boldsymbol{u}_{1:N}, \boldsymbol{v}_{1:N}) &= f_{\boldsymbol{\theta}_{1:N}}(\boldsymbol{\theta}_{1:N}) \prod_{n=1}^{N} f_{\mathbf{u}_n}(\boldsymbol{u}_n) f_{\mathbf{v}_n}(\boldsymbol{v}_n) \\ &= f_{\boldsymbol{\theta}_{1:N}}(\boldsymbol{\theta}_{1:N}) \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{u}_n; \mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{u}})\, \mathcal{N}(\boldsymbol{v}_n; \mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{v}}). \end{aligned} \tag{4.11}$$

In general, except for the first scenario, we do not assume that $\boldsymbol{\theta}_n$ is independent across object index n, hence $f_{\boldsymbol{\theta}_{1:N}}(\boldsymbol{\theta}_{1:N})$ is not necessarily equal to $\prod_{n=1}^{N} f_{\boldsymbol{\theta}_n}(\boldsymbol{\theta}_n)$.

General Factorizations of PDFs

The hierarchical model for generating $\mathbf{x}_n$ from $\boldsymbol{\theta}_n$ (see (4.1)) and $\mathbf{y}_n$ from $\mathbf{x}_n$ (see (4.4)) is a Markov chain, denoted as $\boldsymbol{\theta}_n \to \mathbf{x}_n \to \mathbf{y}_n$. That is, given $\mathbf{x}_n$, $\mathbf{y}_n$ is conditionally independent of $\boldsymbol{\theta}_n$, or equivalently

$$f_{\mathbf{y}_n \,|\, \mathbf{x}_n, \boldsymbol{\theta}_n}(\boldsymbol{y}_n \,|\, \boldsymbol{x}_n, \boldsymbol{\theta}_n) = f_{\mathbf{y}_n \,|\, \mathbf{x}_n}(\boldsymbol{y}_n \,|\, \boldsymbol{x}_n). \tag{4.12}$$

The conditional pdf on the right-hand side of (4.12) is recognized as the likelihood function. We now factorize the joint likelihood function $f_{\mathbf{y}_{1:N} \,|\, \mathbf{x}_{1:N}}(\boldsymbol{y}_{1:N} \,|\, \boldsymbol{x}_{1:N})$ using, in turn, (4.4), (4.10), and (4.6) as follows:

$$\begin{aligned} f_{\mathbf{y}_{1:N} \,|\, \mathbf{x}_{1:N}}(\boldsymbol{y}_{1:N} \,|\, \boldsymbol{x}_{1:N}) &\overset{(4.4)}{=} f_{\mathbf{v}_{1:N} \,|\, \mathbf{x}_{1:N}}(\boldsymbol{y}_{1:N} - \boldsymbol{x}_{1:N} \,|\, \boldsymbol{x}_{1:N}) \\ &\overset{(4.10)}{=} \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{y}_n - \boldsymbol{x}_n; \mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{v}}) \\ &= \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{y}_n; \boldsymbol{x}_n, \boldsymbol{\Sigma}_{\boldsymbol{v}}) \qquad (4.13) \\ &\overset{(4.6)}{=} \prod_{n=1}^{N} f_{\mathbf{y}_n \,|\, \mathbf{x}_n}(\boldsymbol{y}_n \,|\, \boldsymbol{x}_n) \qquad (4.14) \end{aligned}$$

Hence, the joint likelihood function $f_{\mathbf{y}_{1:N}\,|\,\mathbf{x}_{1:N}}(\boldsymbol{y}_{1:N}\,|\,\boldsymbol{x}_{1:N})$ is given for all scenarios by

$$f_{\mathbf{y}_{1:N}\,|\,\mathbf{x}_{1:N}}(\boldsymbol{y}_{1:N}\,|\,\boldsymbol{x}_{1:N}) = \prod_{n=1}^{N} f_{\mathbf{y}_n\,|\,\mathbf{x}_n}(\boldsymbol{y}_n\,|\,\boldsymbol{x}_n) = \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{y}_n; \boldsymbol{x}_n, \boldsymbol{\Sigma}_{\boldsymbol{v}}). \tag{4.15}$$

Furthermore, we factorize $f_{\mathbf{y}_{1:N}\,|\,\mathbf{x}_{1:N},\boldsymbol{\theta}_{1:N}}(\boldsymbol{y}_{1:N}\,|\,\boldsymbol{x}_{1:N},\boldsymbol{\theta}_{1:N})$ using, in turn, (4.4), (4.1), Assumption A3, and (4.10), as follows:

$$\begin{aligned}
f_{\mathbf{y}_{1:N}\,|\,\mathbf{x}_{1:N},\boldsymbol{\theta}_{1:N}}(\boldsymbol{y}_{1:N}\,|\,\boldsymbol{x}_{1:N},\boldsymbol{\theta}_{1:N}) &\overset{(4.4)}{=} f_{\mathbf{v}_{1:N}\,|\,\mathbf{x}_{1:N},\boldsymbol{\theta}_{1:N}}(\boldsymbol{y}_{1:N}-\boldsymbol{x}_{1:N}\,|\,\boldsymbol{x}_{1:N},\boldsymbol{\theta}_{1:N}) \\
&\overset{(4.1)}{=} f_{\mathbf{v}_{1:N}\,|\,\boldsymbol{\theta}_{1:N}+\mathbf{u}_{1:N},\boldsymbol{\theta}_{1:N}}(\boldsymbol{y}_{1:N}-\boldsymbol{x}_{1:N}\,|\,\boldsymbol{\theta}_{1:N}+\boldsymbol{u}_{1:N},\boldsymbol{\theta}_{1:N}) \\
&\overset{\text{A3}}{=} f_{\mathbf{v}_{1:N}}(\boldsymbol{y}_{1:N}-\boldsymbol{x}_{1:N}) \\
&\overset{(4.10)}{=} \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{y}_n-\boldsymbol{x}_n; \mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{v}}) \\
&= \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{y}_n; \boldsymbol{x}_n, \boldsymbol{\Sigma}_{\boldsymbol{v}}).
\end{aligned} \tag{4.16}$$

Thus, using (4.6) and (4.15), we also have

$$f_{\mathbf{y}_{1:N}\,|\,\mathbf{x}_{1:N},\boldsymbol{\theta}_{1:N}}(\boldsymbol{y}_{1:N}\,|\,\boldsymbol{x}_{1:N},\boldsymbol{\theta}_{1:N}) = \prod_{n=1}^{N} f_{\mathbf{y}_n\,|\,\mathbf{x}_n}(\boldsymbol{y}_n\,|\,\boldsymbol{x}_n) \tag{4.17}$$

$$= f_{\mathbf{y}_{1:N}\,|\,\mathbf{x}_{1:N}}(\boldsymbol{y}_{1:N}\,|\,\boldsymbol{x}_{1:N}) \tag{4.18}$$

Next, the conditional pdf $f_{\mathbf{x}_{1:N}\,|\,\boldsymbol{\theta}_{1:N}}(\boldsymbol{x}_{1:N}\,|\,\boldsymbol{\theta}_{1:N})$ can be factorized using, in turn, (4.1), Assumption A3, and (4.9) as follows:

$$\begin{aligned}
f_{\mathbf{x}_{1:N}\,|\,\boldsymbol{\theta}_{1:N}}(\boldsymbol{x}_{1:N}\,|\,\boldsymbol{\theta}_{1:N}) &\overset{(4.1)}{=} f_{\mathbf{u}_{1:N}\,|\,\boldsymbol{\theta}_{1:N}}(\boldsymbol{x}_{1:N}-\boldsymbol{\theta}_{1:N}\,|\,\boldsymbol{\theta}_{1:N}) \\
&\overset{\text{A3}}{=} f_{\mathbf{u}_{1:N}}(\boldsymbol{x}_{1:N}-\boldsymbol{\theta}_{1:N}) \\
&\overset{(4.9)}{=} \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{x}_n-\boldsymbol{\theta}_n; \mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{u}}) \\
&= \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\theta}_n, \boldsymbol{\Sigma}_{\boldsymbol{u}}).
\end{aligned} \tag{4.19}$$

Using (4.3), we also have

$$f_{\mathbf{x}_{1:N}\,|\,\boldsymbol{\theta}_{1:N}}(\boldsymbol{x}_{1:N}\,|\,\boldsymbol{\theta}_{1:N}) = \prod_{n=1}^{N} f_{\mathbf{x}_n\,|\,\boldsymbol{\theta}_n}(\boldsymbol{x}_n\,|\,\boldsymbol{\theta}_n). \tag{4.20}$$

Lastly, the conditional pdf $f_{\mathbf{y}_{1:N}\,|\,\boldsymbol{\theta}_{1:N}}(\boldsymbol{y}_{1:N}\,|\,\boldsymbol{\theta}_{1:N})$ can be written using the law of total probability as

$$f_{\mathbf{y}_{1:N}\,|\,\boldsymbol{\theta}_{1:N}}(\boldsymbol{y}_{1:N}\,|\,\boldsymbol{\theta}_{1:N}) = \int_{\boldsymbol{x}_{1:N}} f_{\mathbf{y}_{1:N}\,|\,\mathbf{x}_{1:N},\boldsymbol{\theta}_{1:N}}(\boldsymbol{y}_{1:N}\,|\,\boldsymbol{x}_{1:N},\boldsymbol{\theta}_{1:N}) f_{\mathbf{x}_{1:N}\,|\,\boldsymbol{\theta}_{1:N}}(\boldsymbol{x}_{1:N}\,|\,\boldsymbol{\theta}_{1:N}) d\boldsymbol{x}_{1:N}. \tag{4.21}$$

Inserting (4.17) and (4.20) into (4.21), we obtain

$$\begin{aligned} f_{\mathbf{y}_{1:N}\,|\,\boldsymbol{\theta}_{1:N}}(\boldsymbol{y}_{1:N}\,|\,\boldsymbol{\theta}_{1:N}) &= \int_{\boldsymbol{x}_{1:N}} \left(\prod_{n=1}^{N} f_{\mathbf{y}_n\,|\,\mathbf{x}_n}(\boldsymbol{y}_n\,|\,\boldsymbol{x}_n) f_{\mathbf{x}_n\,|\,\boldsymbol{\theta}_n}(\boldsymbol{x}_n\,|\,\boldsymbol{\theta}_n) \right) d\boldsymbol{x}_{1:N} \\ &= \prod_{n=1}^{N} \int_{\boldsymbol{x}_n} f_{\mathbf{y}_n\,|\,\mathbf{x}_n}(\boldsymbol{y}_n\,|\,\boldsymbol{x}_n) f_{\mathbf{x}_n\,|\,\boldsymbol{\theta}_n}(\boldsymbol{x}_n\,|\,\boldsymbol{\theta}_n) d\boldsymbol{x}_n. \end{aligned} \tag{4.22}$$

Using (4.12) and again the law of total probability, this becomes further

$$\begin{aligned} f_{\mathbf{y}_{1:N}\,|\,\boldsymbol{\theta}_{1:N}}(\boldsymbol{y}_{1:N}\,|\,\boldsymbol{\theta}_{1:N}) &= \prod_{n=1}^{N} \int_{\boldsymbol{x}_n} f_{\mathbf{y}_n\,|\,\mathbf{x}_n,\boldsymbol{\theta}_n}(\boldsymbol{y}_n\,|\,\boldsymbol{x}_n,\boldsymbol{\theta}_n) f_{\mathbf{x}_n\,|\,\boldsymbol{\theta}_n}(\boldsymbol{x}_n\,|\,\boldsymbol{\theta}_n) d\boldsymbol{x}_n \\ &= \prod_{n=1}^{N} f_{\mathbf{y}_n\,|\,\boldsymbol{\theta}_n}(\boldsymbol{y}_n\,|\,\boldsymbol{\theta}_n). \end{aligned} \tag{4.23}$$

The equality of (4.23) and (4.22) also shows that

$$f_{\mathbf{y}_n\,|\,\boldsymbol{\theta}_n}(\boldsymbol{y}_n\,|\,\boldsymbol{\theta}_n) = \int_{\boldsymbol{x}_n} f_{\mathbf{y}_n\,|\,\mathbf{x}_n}(\boldsymbol{y}_n\,|\,\boldsymbol{x}_n) f_{\mathbf{x}_n\,|\,\boldsymbol{\theta}_n}(\boldsymbol{x}_n\,|\,\boldsymbol{\theta}_n) d\boldsymbol{x}_n. \tag{4.24}$$

Inserting (4.6) and (4.3) into (4.24) we obtain

$$\begin{aligned} f_{\mathbf{y}_n\,|\,\boldsymbol{\theta}_n}(\boldsymbol{y}_n\,|\,\boldsymbol{\theta}_n) &= \int_{\boldsymbol{x_n}} \mathcal{N}(\boldsymbol{y}_n;\boldsymbol{x}_n,\boldsymbol{\Sigma_v}) \mathcal{N}(\boldsymbol{x}_n;\boldsymbol{\theta}_n,\boldsymbol{\Sigma_u}) d\boldsymbol{x}_n \\ &= \int_{\boldsymbol{x_n}} \mathcal{N}(\boldsymbol{y}_n - \boldsymbol{x}_n;\mathbf{0},\boldsymbol{\Sigma_v}) \mathcal{N}(\boldsymbol{x}_n;\boldsymbol{\theta}_n,\boldsymbol{\Sigma_u}) d\boldsymbol{x}_n. \end{aligned} \tag{4.25}$$

This integral is the convolution of two Gaussian pdfs. Therefore according to (2.19), it is again a Gaussian pdf; its mean is the sum of the means, i.e., $\boldsymbol{\mu}_{\boldsymbol{y}_n\,|\,\boldsymbol{\theta}_n} = \boldsymbol{\theta}_n + \mathbf{0} = \boldsymbol{\theta}_n$ and its covariance matrix is the sum of the covariance matrices, i.e., $\boldsymbol{\Sigma}_{\boldsymbol{y}_n\,|\,\boldsymbol{\theta}_n} = \boldsymbol{\Sigma_u} + \boldsymbol{\Sigma_v}$. Thus, we obtain

$$f_{\mathbf{y}_n\,|\,\boldsymbol{\theta}_n}(\boldsymbol{y}_n\,|\,\boldsymbol{\theta}_n) = \mathcal{N}(\boldsymbol{y}_n;\boldsymbol{\theta}_n,\boldsymbol{\Sigma_u} + \boldsymbol{\Sigma_v}), \tag{4.26}$$

and by inserting (4.26) into (4.23)

$$f_{\mathbf{y}_{1:N}\,|\,\boldsymbol{\theta}_{1:N}}(\boldsymbol{y}_{1:N}\,|\,\boldsymbol{\theta}_{1:N}) = \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{y}_n; \boldsymbol{\theta}_n, \boldsymbol{\Sigma_u} + \boldsymbol{\Sigma_v}). \tag{4.27}$$

4.2 First Scenario

In the first scenario, we assume that the hyperparameter $\boldsymbol{\theta}_n$ is i.i.d. across object index n and Gaussian distributed.

4.2.1 Statistical Model

Because of this assumption, the pdf of the hyperparameter vector $\boldsymbol{\theta}_{1:N}$ is given by

$$f_{\boldsymbol{\theta}_{1:N}}(\boldsymbol{\theta}_{1:N}) = \prod_{n=1}^{N} f_{\boldsymbol{\theta}_n}(\boldsymbol{\theta}_n), \tag{4.28}$$

where

$$f_{\boldsymbol{\theta}_n}(\boldsymbol{\theta}_n) = \mathcal{N}(\boldsymbol{\theta}_n; \boldsymbol{\mu_\theta}, \boldsymbol{\Sigma_\theta}), \tag{4.29}$$

with some mean $\boldsymbol{\mu_\theta}$ and covariance matrix $\boldsymbol{\Sigma_\theta}$. Using (4.3) and (4.29) we obtain

$$\begin{aligned} f_{\mathbf{x}_n}(\boldsymbol{x_n}) &= \int_{\boldsymbol{\theta}_n} f_{\mathbf{x}_n\,|\,\boldsymbol{\theta}_n}(\boldsymbol{x}_n\,|\,\boldsymbol{\theta}_n) f_{\boldsymbol{\theta}_n}(\boldsymbol{\theta}_n) d\boldsymbol{\theta}_n \\ &= \int_{\boldsymbol{\theta}_n} \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\theta}_n, \boldsymbol{\Sigma_u}) \mathcal{N}(\boldsymbol{\theta}_n; \boldsymbol{\mu_\theta}, \boldsymbol{\Sigma_\theta}) d\boldsymbol{\theta}_n \\ &= \int_{\boldsymbol{\theta}_n} \mathcal{N}(\boldsymbol{x}_n - \boldsymbol{\theta}_n; \mathbf{0}, \boldsymbol{\Sigma_u}) \mathcal{N}(\boldsymbol{\theta}_n; \boldsymbol{\mu_\theta}, \boldsymbol{\Sigma_\theta}) d\boldsymbol{\theta}_n \\ &= \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu_\theta}, \boldsymbol{\Sigma_\theta} + \boldsymbol{\Sigma_u}), \end{aligned} \tag{4.30}$$

where (2.19) was used. We conclude that

$$f_{\mathbf{x}_n}(\boldsymbol{x_n}) = \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu_x}, \boldsymbol{\Sigma_x}) \tag{4.31}$$

with mean

$$\boldsymbol{\mu_x} = \boldsymbol{\mu_\theta} \tag{4.32}$$

and covariance matrix

$$\boldsymbol{\Sigma_x} = \boldsymbol{\Sigma_\theta} + \boldsymbol{\Sigma_u}. \tag{4.33}$$

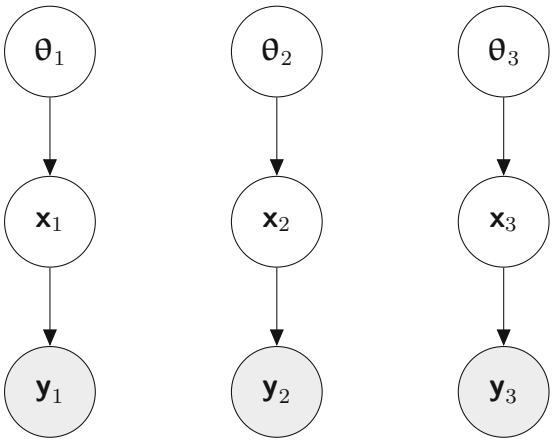


Figure 6: Bayesian network for the first scenario, assuming three objects ($N = 3$). Observed random variables are displayed in shaded disks.

Similarly, using (4.6), (4.30), and again (2.19), we obtain

$$\begin{aligned} f_{\mathbf{y}_n}(\boldsymbol{y}_n) &= \int_{\boldsymbol{x}_n} f_{\mathbf{y}_n \,|\, \mathbf{x}_n}(\boldsymbol{y}_n \,|\, \boldsymbol{x}_n) f_{\mathbf{x}_n}(\boldsymbol{x}_n) d\boldsymbol{x}_n \\ &= \int_{\boldsymbol{x_n}} \mathcal{N}(\boldsymbol{y}_n; \boldsymbol{x}_n, \boldsymbol{\Sigma_v}) \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu_\theta}, \boldsymbol{\Sigma_\theta} + \boldsymbol{\Sigma_u}) d\boldsymbol{x}_n \\ &= \int_{\boldsymbol{x_n}} \mathcal{N}(\boldsymbol{y}_n - \boldsymbol{x}_n; \boldsymbol{0}, \boldsymbol{\Sigma_v}) \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu_\theta}, \boldsymbol{\Sigma_\theta} + \boldsymbol{\Sigma_u}) d\boldsymbol{x}_n \\ &= \mathcal{N}(\boldsymbol{y}_n; \boldsymbol{\mu_\theta}, \boldsymbol{\Sigma_\theta} + \boldsymbol{\Sigma_u} + \boldsymbol{\Sigma_v}), \end{aligned} \tag{4.34}$$

and thus

$$f_{\mathbf{y}_n}(\boldsymbol{y}_n) = \mathcal{N}(\boldsymbol{y}_n; \boldsymbol{\mu_y}, \boldsymbol{\Sigma_y}) \tag{4.35}$$

with mean

$$\boldsymbol{\mu_y} = \boldsymbol{\mu_\theta} \tag{4.36}$$

and covariance matrix

$$\boldsymbol{\Sigma_y} = \boldsymbol{\Sigma_x} + \boldsymbol{\Sigma_v} = \boldsymbol{\Sigma_\theta} + \boldsymbol{\Sigma_u} + \boldsymbol{\Sigma_v}, \tag{4.37}$$

where (4.33) was used. We can visualize the dependencies via the Bayesian network shown in Figure 6.

4.2.2 MMSE Estimator

We will now derive the MMSE estimator of $\mathbf{x}_n$ for this scenario, denoted as $\hat{\boldsymbol{x}}_n^{(1)}(\boldsymbol{y}_{1:N})$. According to (2.10), we have

$$\hat{\boldsymbol{x}}_n^{(1)}(\boldsymbol{y}_{1:N}) = \mathbb{E}[\mathbf{x}_n \,|\, \mathbf{y}_{1:N} = \boldsymbol{y}_{1:N}] = \int_{\boldsymbol{x}_n} \boldsymbol{x}_n \, f_{\mathbf{x}_n \,|\, \mathbf{y}_{1:N}}(\boldsymbol{x}_n \,|\, \boldsymbol{y}_{1:N}) \, d\boldsymbol{x}_n. \tag{4.38}$$

The posterior pdf $f_{\mathbf{x}_n \,|\, \mathbf{y}_{1:N}}(\mathbf{x}_n \,|\, \boldsymbol{y}_{1:N})$ can be obtained from the joint posterior pdf $f_{\mathbf{x}_{1:N} \,|\, \mathbf{y}_{1:N}}(\boldsymbol{x}_{1:N} \,|\, \boldsymbol{y}_{1:N})$ as

$$f_{\mathbf{x}_n \,|\, \mathbf{y}_{1:N}}(\boldsymbol{x}_n \,|\, \boldsymbol{y}_{1:N}) = \int_{\boldsymbol{x}_{\neg n}} f_{\mathbf{x}_{1:N} \,|\, \mathbf{y}_{1:N}}(\boldsymbol{x}_{1:N} \,|\, \boldsymbol{y}_{1:N}) d\boldsymbol{x}_{\neg n}, \tag{4.39}$$

where

$$\boldsymbol{x}_{\neg n} = (\boldsymbol{x}_1^{\mathrm{T}}, \ldots, \boldsymbol{x}_{n-1}^{\mathrm{T}}, \boldsymbol{x}_{n+1}^{\mathrm{T}}, \ldots, \boldsymbol{x}_N^{\mathrm{T}})^{\mathrm{T}} \tag{4.40}$$

stands for all parameters of interest not associated with object n. Using Bayes' theorem, we can write the joint posterior as

$$f_{\mathbf{x}_{1:N} \,|\, \mathbf{y}_{1:N}}(\boldsymbol{x}_{1:N} \,|\, \boldsymbol{y}_{1:N}) = \frac{f_{\mathbf{y}_{1:N} \,|\, \mathbf{x}_{1:N}}(\boldsymbol{y}_{1:N} \,|\, \boldsymbol{x}_{1:N}) f_{\mathbf{x}_{1:N}}(\boldsymbol{x}_{1:N})}{f_{\mathbf{y}_{1:N}}(\boldsymbol{y}_{1:N})}. \tag{4.41}$$

In what follows, we will consider the individual factors in this expression.

- Using (4.20) and (4.28), the joint prior $f_{\mathbf{x}_{1:N}}(\boldsymbol{x}_{1:N})$ can be developed as

$$\begin{aligned} f_{\mathbf{x}_{1:N}}(\boldsymbol{x}_{1:N}) &= \int_{\boldsymbol{\theta}_{1:N}} f_{\mathbf{x}_{1:N} \,|\, \boldsymbol{\theta}_{1:N}}(\boldsymbol{x}_{1:N} \,|\, \boldsymbol{\theta}_{1:N}) f_{\boldsymbol{\theta}_{1:N}}(\boldsymbol{\theta}_{1:N}) d\boldsymbol{\theta}_{1:N} \\ &= \int_{\boldsymbol{\theta}_{1:N}} \left(\prod_{n=1}^{N} f_{\mathbf{x}_n \,|\, \boldsymbol{\theta}_n}(\boldsymbol{x}_n \,|\, \boldsymbol{\theta}_n) f_{\boldsymbol{\theta}_n}(\boldsymbol{\theta}_n) \right) d\boldsymbol{\theta}_{1:N} \\ &= \prod_{n=1}^{N} \int_{\boldsymbol{\theta}_n} f_{\mathbf{x}_n \,|\, \boldsymbol{\theta}_n}(\boldsymbol{x}_n \,|\, \boldsymbol{\theta}_n) f_{\boldsymbol{\theta}_n}(\boldsymbol{\theta}_n) d\boldsymbol{\theta}_n \\ &= \prod_{n=1}^{N} f_{\mathbf{x}_n}(\boldsymbol{x}_n). \end{aligned} \tag{4.42}$$

  Inserting (4.31) into (4.42) gives

$$f_{\mathbf{x}_{1:N}}(\boldsymbol{x}_{1:N}) = \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}_{\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{x}}) = \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}} + \boldsymbol{\Sigma}_{\boldsymbol{u}}). \tag{4.43}$$

- The likelihood function $f_{\mathbf{y}_{1:N}|\mathbf{x}_{1:N}}(\boldsymbol{y}_{1:N}\,|\,\boldsymbol{x}_{1:N})$ is by (4.15)

$$f_{\mathbf{y}_{1:N}|\mathbf{x}_{1:N}}(\boldsymbol{y}_{1:N}\,|\,\boldsymbol{x}_{1:N}) = \prod_{n=1}^{N} f_{\mathbf{y}_n|\mathbf{x}_n}(\boldsymbol{y}_n\,|\,\boldsymbol{x}_n) \tag{4.44}$$

$$= \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{y}_n; \boldsymbol{x}_n, \boldsymbol{\Sigma}_{\boldsymbol{v}}) \tag{4.45}$$

- The evidence $f_{\mathbf{y}_{1:N}}(\boldsymbol{y}_{1:N})$ can be written as

$$f_{\mathbf{y}_{1:N}}(\boldsymbol{y}_{1:N}) = \int_{\boldsymbol{x}_{1:N}} f_{\mathbf{y}_{1:N}|\mathbf{x}_{1:N}}(\boldsymbol{y}_{1:N}\,|\,\boldsymbol{x}_{1:N}) f_{\mathbf{x}_{1:N}}(\boldsymbol{x}_{1:N}) d\boldsymbol{x}_{1:N}. \tag{4.46}$$

Inserting (4.42) and (4.44) into (4.46), we obtain further

$$\begin{aligned} f_{\mathbf{y}_{1:N}}(\boldsymbol{y}_{1:N}) &= \int_{\boldsymbol{x}_{1:N}} \left(\prod_{n=1}^{N} f_{\mathbf{y}_n|\mathbf{x}_n}(\boldsymbol{y}_n\,|\,\boldsymbol{x}_n) f_{\mathbf{x}_n}(\boldsymbol{x}_n) \right) d\boldsymbol{x}_{1:N} \\ &= \prod_{n=1}^{N} \int_{\boldsymbol{x}_n} f_{\mathbf{y}_n|\mathbf{x}_n}(\boldsymbol{y}_n\,|\,\boldsymbol{x}_n) f_{\mathbf{x}_n}(\boldsymbol{x}_n) d\boldsymbol{x}_n \\ &= \prod_{n=1}^{N} f_{\mathbf{y}_n}(\boldsymbol{y}_n) \qquad (4.47) \\ &= \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{y}_n; \boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}} + \boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}}), \qquad (4.48) \end{aligned}$$

where (4.34) was used in the last step.

We now insert (4.42), (4.44), and (4.47) into expression (4.41) and obtain for the joint posterior pdf

$$f_{\mathbf{x}_{1:N}|\mathbf{y}_{1:N}}(\boldsymbol{x}_{1:N}\,|\,\boldsymbol{y}_{1:N}) = \prod_{n=1}^{N} \frac{f_{\mathbf{y}_n|\mathbf{x}_n}(\boldsymbol{y}_n\,|\,\boldsymbol{x}_n) f_{\mathbf{x}_n}(\boldsymbol{x}_n)}{f_{\mathbf{y}_n}(\boldsymbol{y}_n)} = \prod_{n=1}^{N} f_{\mathbf{x}_n|\mathbf{y}_n}(\boldsymbol{x}_n\,|\,\boldsymbol{y}_n). \tag{4.49}$$

Using (4.49), we can now perform the marginalization in (4.39). We obtain

$$\begin{aligned} f_{\mathbf{x}_n|\mathbf{y}_{1:N}}(\boldsymbol{x}_n\,|\,\boldsymbol{y}_{1:N}) &= \int_{\boldsymbol{x}_{\neg n}} \left(\prod_{n'=1}^{N} f_{\mathbf{x}_{n'}|\mathbf{y}_{n'}}(\boldsymbol{x}_{n'}\,|\,\boldsymbol{y}_{n'}) \right) d\boldsymbol{x}_{\neg n} \\ &= f_{\mathbf{x}_n|\mathbf{y}_n}(\boldsymbol{x}_n\,|\,\boldsymbol{y}_n) \int_{\boldsymbol{x}_{\neg n}} \left(\prod_{n' \neq n} f_{\mathbf{x}_{n'}|\mathbf{y}_{n'}}(\boldsymbol{x}_{n'}\,|\,\boldsymbol{y}_{n'}) \right) d\boldsymbol{x}_{\neg n} \\ &= f_{\mathbf{x}_n|\mathbf{y}_n}(\boldsymbol{x}_n\,|\,\boldsymbol{y}_n) \prod_{n' \neq n} \int_{\boldsymbol{x}_{n'}} f_{\mathbf{x}_{n'}|\mathbf{y}_{n'}}(\boldsymbol{x}_{n'}\,|\,\boldsymbol{y}_{n'}) d\boldsymbol{x}_{n'} \end{aligned}$$

$$= f_{\mathbf{x}_n \,|\, \mathbf{y}_n}(\boldsymbol{x}_n \,|\, \boldsymbol{y}_n), \tag{4.50}$$

where we used $\int_{\boldsymbol{x}_{n'}} f_{\mathbf{x}_{n'} \,|\, \mathbf{y}_{n'}}(\boldsymbol{x}_{n'} \,|\, \boldsymbol{y}_{n'}) d\boldsymbol{x}_{n'} = 1$. Note that the marginal posterior in (4.50) only depends on $\boldsymbol{y}_n$ even though we have all measurements $\boldsymbol{y}_{1:N}$ available. Finally, we apply Bayes' theorem to (4.50) and use (4.6), (4.30), and (4.34), to obtain

$$f_{\mathbf{x}_n \,|\, \mathbf{y}_n}(\boldsymbol{x}_n \,|\, \boldsymbol{y}_n) = \frac{f_{\mathbf{y}_n \,|\, \mathbf{x}_n}(\boldsymbol{y}_n \,|\, \boldsymbol{x}_n) f_{\mathbf{x}_n}(\boldsymbol{x}_n)}{f_{\mathbf{y}_n}(\boldsymbol{y}_n)} = \frac{\mathcal{N}(\boldsymbol{y}_n; \boldsymbol{x}_n, \boldsymbol{\Sigma_v}) \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu_\theta}, \boldsymbol{\Sigma_\theta} + \boldsymbol{\Sigma_u})}{\mathcal{N}(\boldsymbol{y}_n; \boldsymbol{\mu_\theta}, \boldsymbol{\Sigma_\theta} + \boldsymbol{\Sigma_u} + \boldsymbol{\Sigma_v})}. \tag{4.51}$$

This must be a Gaussian pdf, since $\mathbf{x}_n$ and $\mathbf{y}_n$ are jointly Gaussian, which follows from (4.8) along with the fact that $\boldsymbol{\theta}_n$, $\mathbf{u}_n$, and $\mathbf{v}_n$ are statistically independent Gaussian random vectors. Thus, we have

$$f_{\mathbf{x}_n \,|\, \mathbf{y}_n}(\boldsymbol{x}_n \,|\, \boldsymbol{y}_n) = \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}_{\boldsymbol{x} \,|\, \boldsymbol{y}_n}, \boldsymbol{\Sigma}_{\boldsymbol{x} \,|\, \boldsymbol{y}}), \tag{4.52}$$

with some posterior mean $\boldsymbol{\mu}_{\boldsymbol{x} \,|\, \boldsymbol{y}_n}$ and posterior covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x} \,|\, \boldsymbol{y}}$. Using (2.40) and (2.41), we obtain

$$\boldsymbol{\Sigma}_{\boldsymbol{x} \,|\, \boldsymbol{y}} = \boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{\Sigma}_{\boldsymbol{xy}} \boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{yx}} \tag{4.53}$$

and

$$\boldsymbol{\mu}_{\boldsymbol{x} \,|\, \boldsymbol{y}_n} = \boldsymbol{\mu}_{\boldsymbol{x}} + \boldsymbol{\Sigma}_{\boldsymbol{xy}} \boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1} (\boldsymbol{y}_n - \boldsymbol{\mu}_{\boldsymbol{y}}). \tag{4.54}$$

We have $\boldsymbol{\Sigma}_{\boldsymbol{y}} = \boldsymbol{\Sigma}_{\boldsymbol{\theta}} + \boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}}$ (see (4.37)) and $\boldsymbol{\Sigma}_{\boldsymbol{xy}} = \operatorname{cov}(\mathbf{x}_n, \mathbf{y}_n) = \operatorname{cov}(\mathbf{x}_n, \mathbf{x}_n + \mathbf{v}_n) = \operatorname{cov}(\mathbf{x}_n) + \operatorname{cov}(\mathbf{x}_n, \mathbf{v}_n) = \operatorname{cov}(\mathbf{x}_n) = \boldsymbol{\Sigma}_{\boldsymbol{x}} = \boldsymbol{\Sigma}_{\boldsymbol{\theta}} + \boldsymbol{\Sigma}_{\boldsymbol{u}}$ (see (4.33)) as well as $\boldsymbol{\mu}_{\boldsymbol{y}} = \boldsymbol{\mu}_{\boldsymbol{x}} = \boldsymbol{\mu}_{\boldsymbol{\theta}}$ (see (4.32) and (4.36)), so that the posterior covariance (4.53) becomes

$$\boldsymbol{\Sigma}_{\boldsymbol{x} \,|\, \boldsymbol{y}} = \boldsymbol{\Sigma}_{\boldsymbol{\theta}} + \boldsymbol{\Sigma}_{\boldsymbol{u}} - (\boldsymbol{\Sigma}_{\boldsymbol{\theta}} + \boldsymbol{\Sigma}_{\boldsymbol{u}}) (\boldsymbol{\Sigma}_{\boldsymbol{\theta}} + \boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1} (\boldsymbol{\Sigma}_{\boldsymbol{\theta}} + \boldsymbol{\Sigma}_{\boldsymbol{u}}). \tag{4.55}$$

Using the matrix identity (B.4) with $\boldsymbol{\Sigma}_{\boldsymbol{A}} = \boldsymbol{\Sigma}_{\boldsymbol{\theta}} + \boldsymbol{\Sigma}_{\boldsymbol{u}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{B}} = \boldsymbol{\Sigma}_{\boldsymbol{v}}$, this becomes further

$$\boldsymbol{\Sigma}_{\boldsymbol{x} \,|\, \boldsymbol{y}} = \boldsymbol{\Sigma}_{\boldsymbol{v}} (\boldsymbol{\Sigma}_{\boldsymbol{\theta}} + \boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1} (\boldsymbol{\Sigma}_{\boldsymbol{\theta}} + \boldsymbol{\Sigma}_{\boldsymbol{u}}). \tag{4.56}$$

Similarly, (4.54) becomes

$$\boldsymbol{\mu}_{\boldsymbol{x} \,|\, \boldsymbol{y}_n} = \boldsymbol{\mu}_{\boldsymbol{\theta}} + (\boldsymbol{\Sigma}_{\boldsymbol{\theta}} + \boldsymbol{\Sigma}_{\boldsymbol{u}}) (\boldsymbol{\Sigma}_{\boldsymbol{\theta}} + \boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1} (\boldsymbol{y}_n - \boldsymbol{\mu}_{\boldsymbol{\theta}}). \tag{4.57}$$

Using (4.50) and (4.57) we can find a closed form expression for the MMSE estimator $\hat{\boldsymbol{x}}_n^{(1)}(\boldsymbol{y}_{1:N})$ in (4.38). According to (4.38), $\hat{\boldsymbol{x}}_n^{(1)}(\boldsymbol{y}_{1:N})$ is equal to the posterior mean $\mathbb{E}[\mathbf{x}_n \,|\, \mathbf{y}_{1:N} = \boldsymbol{y}_{1:N}]$, which, due to (4.50), is equal to $\mathbb{E}[\mathbf{x}_n \,|\, \mathbf{y}_n = \boldsymbol{y}_n] = \boldsymbol{\mu}_{\boldsymbol{x}\,|\,\boldsymbol{y}_n}$, as given by (4.57). Hence, the MMSE estimator is given by

$$\hat{\boldsymbol{x}}_n^{(1)}(\boldsymbol{y}_{1:N}) = \boldsymbol{\mu}_{\boldsymbol{\theta}} + (\boldsymbol{\Sigma}_{\boldsymbol{\theta}} + \boldsymbol{\Sigma}_{\boldsymbol{u}})(\boldsymbol{\Sigma}_{\boldsymbol{\theta}} + \boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}(\boldsymbol{y}_n - \boldsymbol{\mu}_{\boldsymbol{\theta}}). \tag{4.58}$$

4.2.3 MSE

In order to compare the performance of $\hat{\boldsymbol{x}}_n^{(1)}(\boldsymbol{y}_{1:N})$ with that of other estimators in subsequent sections, we consider the minimum MSE, i.e., the MSE achieved by the MMSE estimator. As shown in (2.12), the minimum MSE is given by

$$\mathrm{MSE}_{\mathrm{min}}^{(1)} = \frac{1}{D}\,\mathbb{E}_{\mathbf{y}_{1:N}}\Big[\mathrm{tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{x}_n\,|\,\boldsymbol{y}_{1:N}}\right]\Big]. \tag{4.59}$$

Due to (4.50), the posterior covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}_n\,|\,\boldsymbol{y}_{1:N}}$ is equal to $\boldsymbol{\Sigma}_{\boldsymbol{x}_n\,|\,\boldsymbol{y}_n} = \boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y}}$, hence we further obtain

$$\mathrm{MSE}_{\mathrm{min}}^{(1)} = \frac{1}{D}\,\mathbb{E}_{\mathbf{y}_{1:N}}\Big[\mathrm{tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y}}\right]\Big] = \frac{1}{D}\,\mathrm{tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y}}\right], \tag{4.60}$$

where we exploited the fact that, according to (4.56), the posterior covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y}}$ is not functionally dependent on the measurements $\mathbf{y}_{1:N}$. Note that this expression does not depend on the object index n. Using (4.56), the minimum MSE for the first scenario is finally obtained as

$$\mathrm{MSE}_{\mathrm{min}}^{(1)} = \frac{1}{D}\,\mathrm{tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{v}}(\boldsymbol{\Sigma}_{\boldsymbol{\theta}} + \boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}(\boldsymbol{\Sigma}_{\boldsymbol{\theta}} + \boldsymbol{\Sigma}_{\boldsymbol{u}})\right]. \tag{4.61}$$

In particular, if $\boldsymbol{\theta}_n$, $\mathbf{u}_n$ and $\mathbf{v}_n$ are random vectors with i.i.d. components, i.e., $\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \sigma_\theta^2\mathbf{I}_D$, $\boldsymbol{\Sigma}_{\boldsymbol{u}} = \sigma_u^2\mathbf{I}_D$, and $\boldsymbol{\Sigma}_{\boldsymbol{v}} = \sigma_v^2\mathbf{I}_D$, then (4.61) simplifies to

$$\mathrm{MSE}_{\mathrm{min}}^{(1)} = \frac{1}{D}\,\mathrm{tr}\left[\frac{\sigma_v^2(\sigma_\theta^2 + \sigma_u^2)}{\sigma_\theta^2 + \sigma_u^2 + \sigma_v^2}\mathbf{I}_D\right] = \frac{\sigma_v^2(\sigma_\theta^2 + \sigma_u^2)}{\sigma_\theta^2 + \sigma_u^2 + \sigma_v^2}. \tag{4.62}$$

4.3 Second Scenario

In this scenario, we observe the hyperparameter $\boldsymbol{\theta}_n = \boldsymbol{\theta}_n$ for all $n \in \{1, \ldots, N\}$.

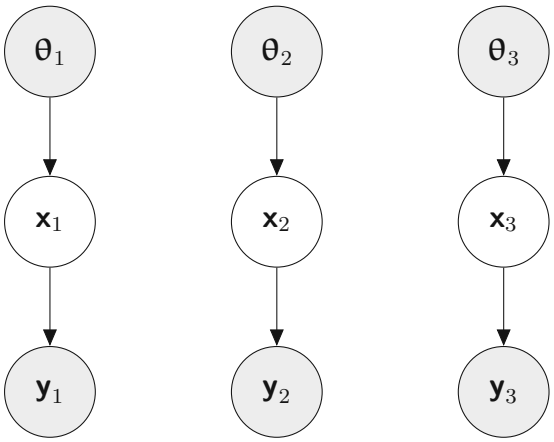


Figure 7: Bayesian network for the second scenario, assuming three objects ($N = 3$). Observed random variables are displayed in shaded disks.

4.3.1 Statistical Model

The hyperparameter Θ_n is still modeled as a random vector; however, in this scenario, we treat it as observed data, just like $\mathbf{y}_n = \boldsymbol{y}_n$. Otherwise, we use the same general model as in the first scenario (see Section 4.1). Therefore, several conditional pdfs are the same and we can use some of the previous results. The conditional dependencies in this scenario are visualized in Figure 7.

4.3.2 MMSE Estimator

As in our first scenario, our goal is to calculate the MMSE estimator of $\mathbf{x}_n$, denoted as $\hat{\boldsymbol{x}}_n^{(2)}(\boldsymbol{y}_{1:N}, \boldsymbol{\theta}_{1:N})$. Since we consider both $\mathbf{y}_{1:N}$ and $\Theta_{1:N}$ as our data, the MMSE estimator is now given by (see (2.10))

$$\hat{\boldsymbol{x}}_n^{(2)}(\boldsymbol{y}_{1:N}, \boldsymbol{\theta}_{1:N}) = \mathbb{E}[\mathbf{x}_n \,|\, \mathbf{y}_{1:N} = \boldsymbol{y}_{1:N}, \Theta_{1:N} = \boldsymbol{\theta}_{1:N}] = \int_{\boldsymbol{x}_n} \boldsymbol{x}_n \, f_{\mathbf{x}_n \,|\, \mathbf{y}_{1:N}, \Theta_{1:N}}(\boldsymbol{x}_n \,|\, \boldsymbol{y}_{1:N}, \boldsymbol{\theta}_{1:N}) \, d\boldsymbol{x}_n, \tag{4.63}$$

with the posterior pdf $f_{\mathbf{x}_n \,|\, \mathbf{y}_{1:N}, \Theta_{1:N}}(\boldsymbol{x}_n \,|\, \boldsymbol{y}_{1:N}, \boldsymbol{\theta}_{1:N})$. Similar to the first scenario, we can obtain the posterior pdf $f_{\mathbf{x}_n \,|\, \mathbf{y}_{1:N}, \Theta_{1:N}}(\boldsymbol{x}_n \,|\, \boldsymbol{y}_{1:N}, \boldsymbol{\theta}_{1:N})$ from the joint posterior pdf $f_{\mathbf{x}_{1:N} \,|\, \mathbf{y}_{1:N}, \Theta_{1:N}}(\boldsymbol{x}_{1:N} \,|\, \boldsymbol{y}_{1:N}, \boldsymbol{\theta}_{1:N})$ according to

$$f_{\mathbf{x}_n \,|\, \mathbf{y}_{1:N}, \Theta_{1:N}}(\boldsymbol{x}_n \,|\, \boldsymbol{y}_{1:N}, \boldsymbol{\theta}_{1:N}) = \int_{\boldsymbol{x}_{\neg n}} f_{\mathbf{x}_{1:N} \,|\, \mathbf{y}_{1:N}, \Theta_{1:N}}(\boldsymbol{x}_{1:N} \,|\, \boldsymbol{y}_{1:N}, \boldsymbol{\theta}_{1:N}) d\boldsymbol{x}_{\neg n}, \tag{4.64}$$

where $\boldsymbol{x}_{\neg n}$ was defined in (4.40). Furthermore, using Bayes' theorem, we obtain for the joint posterior

$$f_{\mathbf{x}_{1:N}\,|\,\mathbf{y}_{1:N},\boldsymbol{\theta}_{1:N}}(\boldsymbol{x}_{1:N}\,|\,\boldsymbol{y}_{1:N},\boldsymbol{\theta}_{1:N}) = \frac{f_{\mathbf{y}_{1:N}\,|\,\mathbf{x}_{1:N},\boldsymbol{\theta}_{1:N}}(\boldsymbol{y}_{1:N}\,|\,\boldsymbol{x}_{1:N},\boldsymbol{\theta}_{1:N}) f_{\mathbf{x}_{1:N}\,|\,\boldsymbol{\theta}_{1:N}}(\boldsymbol{x}_{1:N}\,|\,\boldsymbol{\theta}_{1:N})}{f_{\mathbf{y}_{1:N}\,|\,\boldsymbol{\theta}_{1:N}}(\boldsymbol{y}_{1:N}\,|\,\boldsymbol{\theta}_{1:N})}. \tag{4.65}$$

Next, we will work out the individual factors in this expression.

- The conditional pdf $f_{\mathbf{x}_{1:N}\,|\,\boldsymbol{\theta}_{1:N}}(\boldsymbol{x}_{1:N}\,|\,\boldsymbol{\theta}_{1:N})$ is still given by (4.19) and (4.20), i.e.,

$$f_{\mathbf{x}_{1:N}\,|\,\boldsymbol{\theta}_{1:N}}(\boldsymbol{x}_{1:N}\,|\,\boldsymbol{\theta}_{1:N}) = \prod_{n=1}^{N} f_{\mathbf{x}_n\,|\,\boldsymbol{\theta}_n}(\boldsymbol{x}_n\,|\,\boldsymbol{\theta}_n) \tag{4.66}$$

$$= \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{x}_n;\boldsymbol{\theta}_n,\boldsymbol{\Sigma}_{\boldsymbol{u}}). \tag{4.67}$$

- For the conditional pdf $f_{\mathbf{y}_{1:N}\,|\,\mathbf{x}_{1:N},\boldsymbol{\theta}_{1:N}}(\boldsymbol{y}_{1:N}\,|\,\boldsymbol{x}_{1:N},\boldsymbol{\theta}_{1:N})$, we obtain by (4.16), (4.17), and (4.18)

$$f_{\mathbf{y}_{1:N}\,|\,\mathbf{x}_{1:N},\boldsymbol{\theta}_{1:N}}(\boldsymbol{y}_{1:N}\,|\,\boldsymbol{x}_{1:N},\boldsymbol{\theta}_{1:N}) \overset{(4.16)}{=} \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{y}_n;\boldsymbol{x}_n,\boldsymbol{\Sigma}_{\boldsymbol{v}}) \tag{4.68}$$

$$\overset{(4.17)}{=} \prod_{n=1}^{N} f_{\mathbf{y}_n\,|\,\mathbf{x}_n}(\boldsymbol{y}_n\,|\,\boldsymbol{x}_n) \tag{4.69}$$

$$\overset{(4.18)}{=} f_{\mathbf{y}_{1:N}\,|\,\mathbf{x}_{1:N}}(\boldsymbol{y}_{1:N}\,|\,\boldsymbol{x}_{1:N}). \tag{4.70}$$

- The conditional pdf $f_{\mathbf{y}_{1:N}\,|\,\boldsymbol{\theta}_{1:N}}(\boldsymbol{y}_{1:N}\,|\,\boldsymbol{\theta}_{1:N})$ can be factorized according to (4.23) and (4.27), i.e.,

$$f_{\mathbf{y}_{1:N}\,|\,\boldsymbol{\theta}_{1:N}}(\boldsymbol{y}_{1:N}\,|\,\boldsymbol{\theta}_{1:N}) = \prod_{n=1}^{N} f_{\mathbf{y}_n\,|\,\boldsymbol{\theta}_n}(\boldsymbol{y}_n\,|\,\boldsymbol{\theta}_n) \tag{4.71}$$

$$= \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{y}_n;\boldsymbol{\theta}_n,\boldsymbol{\Sigma}_{\boldsymbol{u}}+\boldsymbol{\Sigma}_{\boldsymbol{v}}). \tag{4.72}$$

Next, we insert (4.66), (4.69), and (4.71) into the expression for the joint posterior pdf (4.65), i.e.,

$$f_{\mathbf{x}_{1:N}\,|\,\mathbf{y}_{1:N},\boldsymbol{\theta}_{1:N}}(\boldsymbol{x}_{1:N}\,|\,\boldsymbol{y}_{1:N},\boldsymbol{\theta}_{1:N}) = \prod_{n=1}^{N} \frac{f_{\mathbf{y}_n\,|\,\mathbf{x}_n}(\boldsymbol{y}_n\,|\,\boldsymbol{x}_n) f_{\mathbf{x}_n\,|\,\boldsymbol{\theta}_n}(\boldsymbol{x}_n\,|\,\boldsymbol{\theta}_n)}{f_{\mathbf{y}_n\,|\,\boldsymbol{\theta}_n}(\boldsymbol{y}_n\,|\,\boldsymbol{\theta}_n)}. \tag{4.73}$$

Using (4.12), we obtain further

$$f_{\mathbf{x}_{1:N}\,|\,\mathbf{y}_{1:N},\boldsymbol{\theta}_{1:N}}(\boldsymbol{x}_{1:N}\,|\,\boldsymbol{y}_{1:N},\boldsymbol{\theta}_{1:N}) = \prod_{n=1}^{N} \frac{f_{\mathbf{y}_n\,|\,\mathbf{x}_n,\boldsymbol{\theta}_n}(\boldsymbol{y}_n\,|\,\boldsymbol{x}_n,\boldsymbol{\theta}_n) f_{\mathbf{x}_n\,|\,\boldsymbol{\theta}_n}(\boldsymbol{x}_n\,|\,\boldsymbol{\theta}_n)}{f_{\mathbf{y}_n\,|\,\boldsymbol{\theta}_n}(\boldsymbol{y}_n\,|\,\boldsymbol{\theta}_n)} \tag{4.74}$$

$$= \prod_{n=1}^{N} f_{\mathbf{x}_n\,|\,\mathbf{y}_n,\boldsymbol{\theta}_n}(\boldsymbol{x}_n\,|\,\boldsymbol{y}_n,\boldsymbol{\theta}_n). \tag{4.75}$$

We now insert the factorization (4.75) into (4.64) and obtain

$$\begin{aligned} f_{\mathbf{x}_n\,|\,\mathbf{y}_{1:N},\boldsymbol{\theta}_{1:N}}(\boldsymbol{x}_n\,|\,\boldsymbol{y}_{1:N},\boldsymbol{\theta}_{1:N}) &= \int_{\boldsymbol{x}_{\neg n}} \left(\prod_{n'=1}^{N} f_{\mathbf{x}_{n'}\,|\,\mathbf{y}_{n'},\boldsymbol{\theta}_{n'}}(\boldsymbol{x}_{n'}\,|\,\boldsymbol{y}_{n'},\boldsymbol{\theta}_{n'}) \right) d\boldsymbol{x}_{\neg n} \\ &= f_{\mathbf{x}_n\,|\,\mathbf{y}_n,\boldsymbol{\theta}_n}(\boldsymbol{x}_n\,|\,\boldsymbol{y}_n,\boldsymbol{\theta}_n) \int_{\boldsymbol{x}_{\neg n}} \left(\prod_{n'\neq n} f_{\mathbf{x}_{n'}\,|\,\mathbf{y}_{n'},\boldsymbol{\theta}_{n'}}(\boldsymbol{x}_{n'}\,|\,\boldsymbol{y}_{n'},\boldsymbol{\theta}_{n'}) \right) d\boldsymbol{x}_{\neg n} \\ &= f_{\mathbf{x}_n\,|\,\mathbf{y}_n,\boldsymbol{\theta}_n}(\boldsymbol{x}_n\,|\,\boldsymbol{y}_n,\boldsymbol{\theta}_n) \prod_{n'\neq n} \int_{\boldsymbol{x}_{n'}} f_{\mathbf{x}_{n'}\,|\,\mathbf{y}_{n'},\boldsymbol{\theta}_{n'}}(\boldsymbol{x}_{n'}\,|\,\boldsymbol{y}_{n'},\boldsymbol{\theta}_{n'}) d\boldsymbol{x}_{n'} \\ &= f_{\mathbf{x}_n\,|\,\mathbf{y}_n,\boldsymbol{\theta}_n}(\boldsymbol{x}_n\,|\,\boldsymbol{y}_n,\boldsymbol{\theta}_n). \end{aligned} \tag{4.76}$$

Finally, we apply Bayes' theorem to (4.76) and use (4.12) together with (4.3), (4.6), and (4.26), to obtain

$$f_{\mathbf{x}_n\,|\,\mathbf{y}_n,\boldsymbol{\theta}_n}(\boldsymbol{x}_n\,|\,\boldsymbol{y}_n,\boldsymbol{\theta}_n) = \frac{f_{\mathbf{y}_n\,|\,\mathbf{x}_n,\boldsymbol{\theta}_n}(\boldsymbol{y}_n\,|\,\boldsymbol{x}_n,\boldsymbol{\theta}_n) f_{\mathbf{x}_n\,|\,\boldsymbol{\theta}_n}(\boldsymbol{x}_n\,|\,\boldsymbol{\theta}_n)}{f_{\mathbf{y}_n\,|\,\boldsymbol{\theta}_n}(\boldsymbol{y}_n\,|\,\boldsymbol{\theta}_n)} \tag{4.77}$$

$$= \frac{f_{\mathbf{y}_n\,|\,\mathbf{x}_n}(\boldsymbol{y}_n\,|\,\boldsymbol{x}_n) f_{\mathbf{x}_n\,|\,\boldsymbol{\theta}_n}(\boldsymbol{x}_n\,|\,\boldsymbol{\theta}_n)}{f_{\mathbf{y}_n\,|\,\boldsymbol{\theta}_n}(\boldsymbol{y}_n\,|\,\boldsymbol{\theta}_n)}$$

$$= \frac{\mathcal{N}(\boldsymbol{y}_n;\boldsymbol{x}_n,\boldsymbol{\Sigma_v})\mathcal{N}(\boldsymbol{x_n};\boldsymbol{\theta}_n,\boldsymbol{\Sigma_u})}{\mathcal{N}(\boldsymbol{y}_n;\boldsymbol{\theta}_n,\boldsymbol{\Sigma_u}+\boldsymbol{\Sigma_v})}, \tag{4.78}$$

which again is a Gaussian pdf, as $\mathbf{y}_n$ and $\mathbf{x}_n$ are still jointly Gaussian. Similar to the first scenario, we obtain

$$f_{\mathbf{x}_n\,|\,\mathbf{y}_n,\boldsymbol{\theta}_n}(\boldsymbol{x}_n\,|\,\boldsymbol{y}_n,\boldsymbol{\theta}_n) = \mathcal{N}(\boldsymbol{x}_n;\boldsymbol{\mu}_{\boldsymbol{x}\,|\,\boldsymbol{y}_n,\boldsymbol{\theta}_n},\boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y},\boldsymbol{\theta}}) \tag{4.79}$$

with (see (2.40))

$$\boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y},\boldsymbol{\theta}} = \boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{\theta}} - \boldsymbol{\Sigma}_{\boldsymbol{xy}\,|\,\boldsymbol{\theta}} \boldsymbol{\Sigma}^{-1}_{\boldsymbol{y}\,|\,\boldsymbol{\theta}} \boldsymbol{\Sigma}_{\boldsymbol{yx}\,|\,\boldsymbol{\theta}} \tag{4.80}$$

and (see (2.41))

$$\boldsymbol{\mu}_{\boldsymbol{x}\,|\,\boldsymbol{y}_n,\boldsymbol{\theta}_n} = \boldsymbol{\mu}_{\boldsymbol{x}\,|\,\boldsymbol{\theta}_n} + \boldsymbol{\Sigma}_{\boldsymbol{xy}\,|\,\boldsymbol{\theta}}\boldsymbol{\Sigma}_{\boldsymbol{y}\,|\,\boldsymbol{\theta}}^{-1}(\boldsymbol{y}_n - \boldsymbol{\mu}_{\boldsymbol{y}_n\,|\,\boldsymbol{\theta}_n}). \tag{4.81}$$

Using $\boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{\theta}} = \boldsymbol{\Sigma}_{\boldsymbol{u}}$ (see (4.3)), $\boldsymbol{\Sigma}_{\boldsymbol{y}\,|\,\boldsymbol{\theta}} = \boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}}$ (see (4.26)), and $\boldsymbol{\Sigma}_{\boldsymbol{xy}\,|\,\boldsymbol{\theta}} = \boldsymbol{\Sigma}_{\boldsymbol{yx}\,|\,\boldsymbol{\theta}} = \mathrm{cov}(\mathbf{x}_n, \mathbf{y}_n \,|\, \boldsymbol{\theta}_n) = \mathrm{cov}(\boldsymbol{\theta}_n + \mathbf{u}_n, \boldsymbol{\theta}_n + \mathbf{u}_n + \mathbf{v}_n \,|\, \boldsymbol{\theta}_n) = \mathrm{cov}(\mathbf{u}_n, \mathbf{u}_n + \mathbf{v}_n) = \mathrm{cov}(\mathbf{u}_n) = \boldsymbol{\Sigma}_{\boldsymbol{u}}$. We can write (4.80) as

$$\boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y},\boldsymbol{\theta}} = \boldsymbol{\Sigma}_{\boldsymbol{u}} - \boldsymbol{\Sigma}_{\boldsymbol{u}}(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}\boldsymbol{\Sigma}_{\boldsymbol{u}} \tag{4.82}$$

$$\overset{(B.4)}{=} \boldsymbol{\Sigma}_{\boldsymbol{v}}(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}\boldsymbol{\Sigma}_{\boldsymbol{u}}. \tag{4.83}$$

Similarly, using $\boldsymbol{\mu}_{\boldsymbol{x}\,|\,\boldsymbol{\theta}_n} = \boldsymbol{\theta}_n$ (see (4.3)) and $\boldsymbol{\mu}_{\boldsymbol{y}_n\,|\,\boldsymbol{\theta}_n} = \boldsymbol{\theta}_n$ (see (4.26)). we can write (4.81) as

$$\boldsymbol{\mu}_{\boldsymbol{x}\,|\,\boldsymbol{y}_n,\boldsymbol{\theta}_n} = \boldsymbol{\theta}_n + \boldsymbol{\Sigma}_{\boldsymbol{u}}(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}(\boldsymbol{y}_n - \boldsymbol{\theta}_n). \tag{4.84}$$

According to (4.63), the MMSE estimator $\hat{\boldsymbol{x}}_n^{(2)}(\boldsymbol{y}_{1:N}, \boldsymbol{\theta}_{1:N})$ is equal to the posterior mean $\mathbb{E}[\mathbf{x}_n \,|\, \mathbf{y}_{1:N} = \boldsymbol{y}_{1:N}, \boldsymbol{\theta}_{1:N} = \boldsymbol{\theta}_{1:N}]$, which due to (4.76) is equal to $\mathbb{E}[\mathbf{x}_n \,|\, \mathbf{y}_n = \boldsymbol{y}_n, \boldsymbol{\theta}_n = \boldsymbol{\theta}_n] = \boldsymbol{\mu}_{\boldsymbol{x}\,|\,\boldsymbol{y}_n,\boldsymbol{\theta}_n}$. Using (4.84), we then obtain

$$\hat{\boldsymbol{x}}_n^{(2)}(\boldsymbol{y}_{1:N}, \boldsymbol{\theta}_{1:N}) = \boldsymbol{\theta}_n + \boldsymbol{\Sigma}_{\boldsymbol{u}}(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}(\boldsymbol{y}_n - \boldsymbol{\theta}_n). \tag{4.85}$$

4.3.3 MSE

According to (2.12), the minimum MSE is given by

$$\mathrm{MSE}_{\mathrm{min}}^{(2)} = \frac{1}{D}\,\mathbb{E}_{\mathbf{y}_{1:N},\boldsymbol{\theta}_{1:N}}\left[\mathrm{tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{x}_n\,|\,\boldsymbol{y}_{1:N},\boldsymbol{\theta}_{1:N}}\right]\right]. \tag{4.86}$$

Due to (4.76), the posterior covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}_n\,|\,\boldsymbol{y}_{1:N},\boldsymbol{\theta}_{1:N}}$ is equal to $\boldsymbol{\Sigma}_{\boldsymbol{x}_n\,|\,\boldsymbol{y}_n,\boldsymbol{\theta}_n} = \boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y},\boldsymbol{\theta}}$, hence we further obtain

$$\mathrm{MSE}_{\mathrm{min}}^{(2)} = \frac{1}{D}\,\mathbb{E}_{\mathbf{y}_{1:N},\boldsymbol{\theta}_{1:N}}\left[\mathrm{tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y},\boldsymbol{\theta}}\right]\right] = \frac{1}{D}\,\mathrm{tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y},\boldsymbol{\theta}}\right], \tag{4.87}$$

where we exploited the fact that the posterior covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y},\boldsymbol{\theta}}$ is not functionally dependent on $\mathbf{y}_{1:N}$ or on $\boldsymbol{\theta}_{1:N}$, as evidenced by (4.83). Using (4.83), the minimum MSE for

the second scenario is finally obtained as

$$\mathrm{MSE}_{\min}^{(2)} = \frac{1}{D}\,\mathrm{tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{v}}(\boldsymbol{\Sigma}_{\boldsymbol{u}}+\boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}\boldsymbol{\Sigma}_{\boldsymbol{u}}\right]. \tag{4.88}$$

Again, this expression does not depend on the object index n.

In the special case where $\mathbf{u}_n$ and $\mathbf{v}_n$ are random vectors with i.i.d. components, i.e., $\boldsymbol{\Sigma}_{\boldsymbol{u}} = \sigma_u^2\mathbf{I}_D$ and $\boldsymbol{\Sigma}_{\boldsymbol{v}} = \sigma_v^2\mathbf{I}_D$, expression (4.88) reduces to

$$\mathrm{MSE}_{\min}^{(2)} = \frac{1}{D}\,\mathrm{tr}\left[\frac{\sigma_u^2\sigma_v^2}{\sigma_u^2+\sigma_v^2}\mathbf{I}_D\right] = \frac{\sigma_u^2\sigma_v^2}{\sigma_u^2+\sigma_v^2}. \tag{4.89}$$

4.4 Comparison of $\mathrm{MSE}_{\min}^{(1)}$ and $\mathrm{MSE}_{\min}^{(2)}$

It is interesting to compare $\mathrm{MSE}_{\min}^{(1)}$, i.e., the minimum MSE for Scenario 1, and $\mathrm{MSE}_{\min}^{(2)}$, i.e., the minimum MSE for Scenario 2. For simplicity, we restrict our discussion to the case where $\mathbf{u}_n$, $\mathbf{v}_n$, and in Scenario 1 also $\boldsymbol{\theta}_n$ are random vectors with i.i.d. components, i.e., $\boldsymbol{\Sigma}_{\boldsymbol{u}} = \sigma_u^2\mathbf{I}_D$, $\boldsymbol{\Sigma}_{\boldsymbol{v}} = \sigma_v^2\mathbf{I}_D$, and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \sigma_\theta^2\mathbf{I}_D$. For a quantitative comparison of $\mathrm{MSE}_{\min}^{(1)}$ and $\mathrm{MSE}_{\min}^{(2)}$, we develop our expression for $\mathrm{MSE}_{\min}^{(1)}$ in (4.62) to obtain

$$\begin{aligned}\mathrm{MSE}_{\min}^{(1)} &= \sigma_v^2\frac{\sigma_\theta^2+\sigma_u^2}{\sigma_\theta^2+\sigma_u^2+\sigma_v^2}\\ &= \sigma_v^2\frac{\sigma_\theta^2+\sigma_u^2+\sigma_v^2-\sigma_v^2}{\sigma_\theta^2+\sigma_u^2+\sigma_v^2}\\ &= \sigma_v^2\left(1-\frac{\sigma_v^2}{\sigma_\theta^2+\sigma_u^2+\sigma_v^2}\right).\end{aligned} \tag{4.90}$$

Similarly, we manipulate our expression for $\mathrm{MSE}_{\min}^{(2)}$ in (4.89) to obtain

$$\begin{aligned}\mathrm{MSE}_{\min}^{(2)} &= \sigma_v^2\frac{\sigma_u^2}{\sigma_u^2+\sigma_v^2}\\ &= \sigma_v^2\frac{\sigma_u^2+\sigma_v^2-\sigma_v^2}{\sigma_u^2+\sigma_v^2}\\ &= \sigma_v^2\left(1-\frac{\sigma_v^2}{\sigma_u^2+\sigma_v^2}\right).\end{aligned} \tag{4.91}$$

Now

$$\frac{\sigma_v^2}{\sigma_\theta^2+\sigma_u^2+\sigma_v^2} \le \frac{\sigma_v^2}{\sigma_u^2+\sigma_v^2} \tag{4.92}$$

and further

$$1 - \frac{\sigma_v^2}{\sigma_\theta^2 + \sigma_u^2 + \sigma_v^2} \geq 1 - \frac{\sigma_v^2}{\sigma_u^2 + \sigma_v^2}. \tag{4.93}$$

Together with expressions (4.90) and (4.91), this implies

$$\mathrm{MSE}_{\min}^{(1)} \geq \mathrm{MSE}_{\min}^{(2)}, \tag{4.94}$$

with equality, i.e., $\mathrm{MSE}_{\min}^{(1)} = \mathrm{MSE}_{\min}^{(2)}$, if $\sigma_\theta^2 = 0$. Note that if $\sigma_\theta^2 \to 0$ in Scenario 1, then $\boldsymbol{\theta}_n \to \boldsymbol{\mu}_{\boldsymbol{\theta}}$, which means that $\boldsymbol{\theta}_n$ is known and thus Scenario 1 reduces to Scenario 2.

5 Inherent Clustering Scenarios and Estimators

In the previous chapter, we considered Scenario 1 with unknown i.i.d. hyperparameters $\boldsymbol{\theta}_n$, and Scenario 2, where the hyperparameters $\boldsymbol{\theta}_n$ are assumed to be observed and thus known. In this chapter, we consider two further scenarios, referred to as Scenarios 3 and 4. Their difference from the previously discussed Scenarios 1 and 2 is that the underlying Bayesian models involve a DP prior on the hyperparameters $\boldsymbol{\theta}_n$, which allows the estimators to exploit the associated cluster structure to improve the estimation performance. In Scenario 4, we assume that we know which objects belong to each cluster, whereas in Scenario 3, the cluster assignment needs to be inferred and the estimator is only provided with the measurements.

This chapter is organized as follows. In Section 5.1, we present the statistical model for this section, which still relies on the general model presented in Section 4.1. This means that all independence assumptions and factorizations of pdfs described in Section 4.1 are also valid for Scenarios 3 and 4.

In Section 5.2, we derive the MMSE estimator for Scenario 3. This estimator takes into account the DP prior on the hyperparameters $\boldsymbol{\theta}_n$. Because the estimator cannot be calculated in closed form, we provide a Monte Carlo (MC) approximation.

In Section 5.3, we derive the MMSE estimator for Scenario 4. While we still impose the DP prior, we now assume that we know which objects belong to each cluster, i.e., the estimator is provided with the cluster assignment variables $\mathbf{C}_{1:N} = (\mathsf{C}_1, \ldots, \mathsf{C}_N)^{\mathrm{T}}$ as defined in (3.39). We obtain a closed-form expression of this estimator.

5.1 Statistical Model for Scenarios 3 and 4

We assume that $\boldsymbol{\theta}_n \in \{1, \ldots, N\}$ is distributed according to a Dirichlet process (DP) as introduced in Chapter 3. We have (see (3.10))

$$\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N \,|\, (\mathsf{f}_{\mathrm{DP}} = f_{\mathrm{DP}}) \sim_{\text{i.i.d.}} f_{\mathrm{DP}}, \tag{5.1}$$

where $\mathsf{f}_{\mathrm{DP}} \sim \mathrm{DP}(\alpha, f_{\mathrm{H}})$ (see (3.7)) with concentration parameter $\alpha > 0$. The base distribution of the DP is assumed to be Gaussian, i.e.,

$$f_{\mathrm{H}}(\boldsymbol{\theta}^*) = \mathcal{N}(\boldsymbol{\theta}^*; \boldsymbol{\mu}_{\boldsymbol{\theta}^*}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}), \tag{5.2}$$

with some mean $\boldsymbol{\mu}_{\boldsymbol{\theta}^*}$ and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}$. The random cluster hyperparameters $\boldsymbol{\theta}_l^*$ are i.i.d. and each is distributed according to the base pdf (see (3.1)), i.e.,

$$\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \ldots \sim_{\text{i.i.d.}} f_{\text{H}}. \tag{5.3}$$

As opposed to the first scenario, discussed in Section 4.2, the hyperparameters $\boldsymbol{\theta}_n$ are not independent across the object index n; however, they are conditionally i.i.d. given $\mathsf{f}_{\text{DP}} = f_{\text{DP}}$ as stated in (5.1).

The random parameters of interest $\mathbf{x}_n$ are parametrized by the respective hyperparameters $\boldsymbol{\theta}_n$ according to (3.78), i.e.,

$$f_{\mathbf{x}_n \,|\, \boldsymbol{\theta}_n}(\boldsymbol{x}_n \,|\, \boldsymbol{\theta}_n) = \phi(\boldsymbol{x}_n \,|\, \boldsymbol{\theta}_n), \quad n \in \{1, \ldots, N\} \tag{5.4}$$

with some continuous pdf $\phi(\boldsymbol{x}_n \,|\, \boldsymbol{\theta}_n)$. Hence, the $\mathbf{x}_n$ are distributed according to a DPM as discussed in Section 3.4 and $\mathbf{x}_n$ is conditionally i.i.d. given $\boldsymbol{\theta}_{1:N}$, i.e.,

$$\mathbf{x}_n \,|\, (\boldsymbol{\theta}_{1:N} = \boldsymbol{\theta}_{1:N}) \sim_{\text{i.i.d.}} f_{\mathbf{x}_n \,|\, \boldsymbol{\theta}_{1:N}}(\boldsymbol{x}_n \,|\, \boldsymbol{\theta}_{1:N}). \tag{5.5}$$

We recall that given $\boldsymbol{\theta}_n$, $\mathbf{x}_n$ is conditionally independent of all other hyperparameters $\boldsymbol{\theta}_{n'}$ with $n' \neq n$, i.e.,

$$f_{\mathbf{x}_n \,|\, \boldsymbol{\theta}_{1:N}}(\boldsymbol{x}_n \,|\, \boldsymbol{\theta}_{1:N}) = f_{\mathbf{x}_n \,|\, \boldsymbol{\theta}_n}(\boldsymbol{x}_n \,|\, \boldsymbol{\theta}_n) \tag{5.6}$$

and

$$f_{\mathbf{x}_{1:N} \,|\, \boldsymbol{\theta}_{1:N}}(\boldsymbol{x}_{1:N} \,|\, \boldsymbol{\theta}_{1:N}) = \prod_{n=1}^{N} f_{\mathbf{x}_n \,|\, \boldsymbol{\theta}_{1:N}}(\boldsymbol{x}_n \,|\, \boldsymbol{\theta}_{1:N}) = \prod_{n=1}^{N} f_{\mathbf{x}_n \,|\, \boldsymbol{\theta}_n}(\boldsymbol{x}_n \,|\, \boldsymbol{\theta}_n). \tag{5.7}$$

Furthermore, still assuming the Gaussian additive-noise model in (4.1), we have (see (4.3))

$$\phi(\boldsymbol{x}_n \,|\, \boldsymbol{\theta}_n) = f_{\mathbf{x}_n \,|\, \boldsymbol{\theta}_n}(\boldsymbol{x}_n \,|\, \boldsymbol{\theta}_n) = \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\theta}_n, \boldsymbol{\Sigma}_{\boldsymbol{u}}), \tag{5.8}$$

with some covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{u}}$.

The measurements $\mathbf{y}_n$ are generated according to (4.4), i.e.,

$$\mathbf{y}_n = \mathbf{x}_n + \mathbf{v}_n, \tag{5.9}$$

where $\mathbf{v}_n$ is i.i.d. and zero-mean Gaussian. Therefore (see (4.6))

$$f_{\mathbf{y}_n\,|\,\mathbf{x}_n}(\boldsymbol{y}_n\,|\,\boldsymbol{x}_n) = \mathcal{N}(\boldsymbol{y}_n; \boldsymbol{x}_n, \boldsymbol{\Sigma}_{\boldsymbol{v}}), \tag{5.10}$$

with some covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{v}}$.

We will use the cluster assignment variables $\mathbf{C}_{1:N} = (\mathsf{C}_1, \ldots, \mathsf{C}_N)^{\mathrm{T}}$ to express the assignment of objects to clusters, i.e. (see (3.39))

$$\mathsf{C}_n = l \quad \text{if} \quad \boldsymbol{\theta}_n = \boldsymbol{\theta}_l^*. \tag{5.11}$$

We can also write (see (3.40))

$$\boldsymbol{\theta}_n = \boldsymbol{\theta}_{\mathsf{C}_n}^*, \tag{5.12}$$

which means that

$$\boldsymbol{\theta}_{1:N} = (\boldsymbol{\theta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\theta}_N^{\mathrm{T}})^{\mathrm{T}} = (\boldsymbol{\theta}_{\mathsf{C}_1}^{*\mathrm{T}}, \ldots, \boldsymbol{\theta}_{\mathsf{C}_N}^{*\mathrm{T}})^{\mathrm{T}}. \tag{5.13}$$

In other words, for all objects n that share the same cluster hyperparameter $\boldsymbol{\theta}_l^*$ we assign cluster $\mathsf{C}_n = l$. For example, let us consider three different objects with parameters of interest $\mathbf{x}_i$, $\mathbf{x}_j$ and $\mathbf{x}_k$, and let us assume that two parameters of interest belong to the same cluster, i.e., $\mathbf{x}_j = \boldsymbol{\theta}_t^* + \mathbf{u}_j$ and $\mathbf{x}_k = \boldsymbol{\theta}_t^* + \mathbf{u}_k$, whereas $\mathbf{x}_i = \boldsymbol{\theta}_p^* + \mathbf{u}_i$, with $p \neq t$. This means that the cluster assignment variables are given by $\mathsf{C}_j = \mathsf{C}_k = t$ and $\mathsf{C}_i = p$. We use $m_l(N)$ as defined in (3.47) to denote the number of objects that belong to the same cluster l. Finally, we recall from Section 3.3.2 the notation

$$\mathscr{C}(N) = \{\mathbf{C}_{1:N}\} \triangleq \{\mathsf{C}_1, \ldots, \mathsf{C}_N\}, \tag{5.14}$$

which is the set containing all unique cluster assignment variables C_n, $n = 1, \ldots, N$, and we define

$$\boldsymbol{\theta}_{\mathscr{C}(N)}^* \triangleq (\boldsymbol{\theta}_l^*)_{l \in \mathscr{C}(N)} \tag{5.15}$$

as the vector composed of all cluster hyperparameters $\boldsymbol{\theta}_l^*$.

5.2 Third Scenario

In this section, we will develop an MMSE estimator for Scenario 3, using the general model introduced in the previous section. We assume that the cluster assignment variables C_n, $n =$

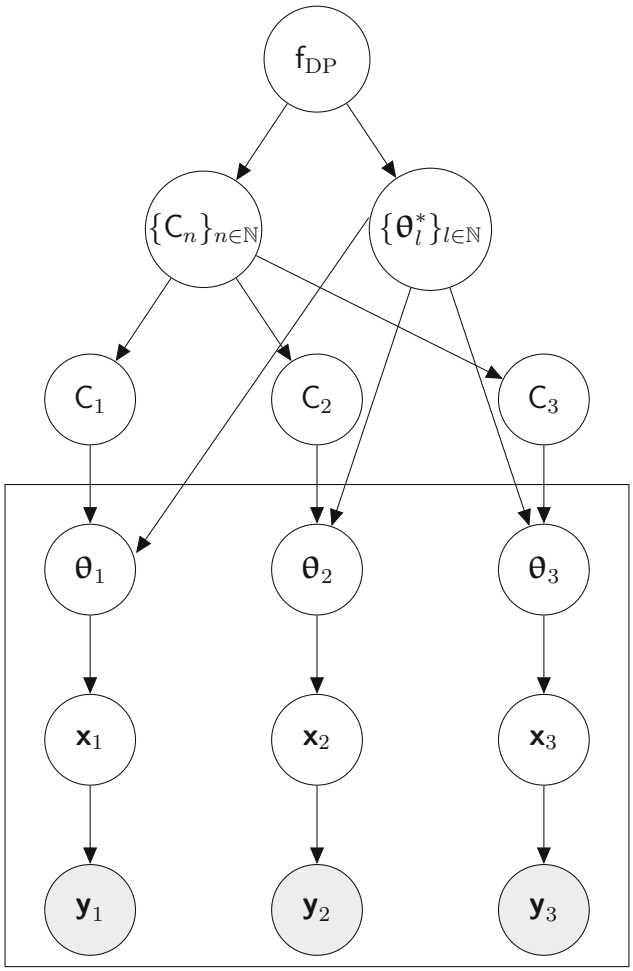


Figure 8: Bayesian network for the third scenario, assuming three objects ($N = 3$). Random variables displayed in shaded disks are observed. The cluster assignment variables $\{\mathsf{C}_n\}_{n\in\mathbb{N}}$ and the cluster hyperparameters $\{\boldsymbol{\theta}^*_l\}_{l\in\mathbb{N}}$ are generated from the DP. Each cluster assignment variable C_n is then related to exactly one hyperparameter $\boldsymbol{\theta}_n$; however, the cluster assignment variables may be equal for two different objects and thereby also relate these objects to the same cluster hyperparameter $\boldsymbol{\theta}^*_l$.

$1, \ldots, N$ are unknown to the estimator (as opposed to Scenario 4 discussed in Section 5.3). The statistical model for this scenario is shown in Figure 8.

5.2.1 MC Approximation of the MMSE Estimator

Similarly to our first and second scenarios, we want to calculate the MMSE estimate

$$\hat{\boldsymbol{x}}_n^{(3)}(\boldsymbol{y}_{1:N}) = \mathbb{E}[\mathbf{x}_n \,|\, \mathbf{y}_{1:N} = \boldsymbol{y}_{1:N}] = \int_{\boldsymbol{x}_n} \boldsymbol{x}_n \, f_{\mathbf{x}_n \,|\, \mathbf{y}_{1:N}}(\boldsymbol{x}_n \,|\, \boldsymbol{y}_{1:N}) \, d\boldsymbol{x}_n. \tag{5.16}$$

Due to the prior (5.1) imposed on the hyperparameters $\boldsymbol{\theta}_n$, we cannot calculate the posterior pdf $f_{\mathbf{x}_n \,|\, \mathbf{y}_{1:N}}(\boldsymbol{x}_n \,|\, \boldsymbol{y}_{1:N})$ in closed form as in the other scenarios. We can, however, generate

a set of samples $\boldsymbol{x}_n^{(q)}$ for $q = 1, \ldots, Q$ from $f_{\mathbf{x}_n \,|\, \mathbf{y}_{1:N}}(\boldsymbol{x}_n \,|\, \boldsymbol{y}_{1:N})$, where $Q \in \mathbb{N}$ denotes the number of samples, and use these samples to obtain the following MC approximation of the MMSE estimator in (5.16):

$$\hat{\boldsymbol{x}}_n^{(3)}(\boldsymbol{y}_{1:N}) \approx \frac{1}{Q} \sum_{q=1}^{Q} \boldsymbol{x}_n^{(q)}. \tag{5.17}$$

It follows from the law of large numbers that this approximation is accurate for sufficiently large Q and is exact for $Q \to \infty$ [2]. However, as it is impossible to sample directly from the posterior distribution $f_{\mathbf{x}_n \,|\, \mathbf{y}_{1:N}}(\boldsymbol{x}_n \,|\, \boldsymbol{y}_{1:N})$, we will adapt the Markov Chain Monte Carlo (MCMC) approach; concretely, we will develop two different variations of the Gibbs sampler [21]. The Gibbs sampler can be viewed as a special case of the Metropolis-Hastings algorithm with kernel cycles, with the acceptance probability for each proposal equal to one [2] [6]. The Gibbs sampler is an iterative algorithm that generates samples from "full conditional pdfs".

The posterior pdf $f_{\mathbf{x}_n \,|\, \mathbf{y}_{1:N}}(\boldsymbol{x}_n \,|\, \boldsymbol{y}_{1:N})$ can be obtained from the joint posterior pdf $f_{\mathbf{x}_{1:N} \,|\, \mathbf{y}_{1:N}}(\boldsymbol{x}_n \,|\, \boldsymbol{y}_{1:N})$ according to

$$f_{\mathbf{x}_n \,|\, \mathbf{y}_{1:N}}(\boldsymbol{x}_n \,|\, \boldsymbol{y}_{1:N}) = \int_{\boldsymbol{x}_{\neg n}} f_{\mathbf{x}_{1:N} \,|\, \mathbf{y}_{1:N}}(\boldsymbol{x}_{1:N} \,|\, \boldsymbol{y}_{1:N}) d\boldsymbol{x}_{\neg n}. \tag{5.18}$$

As a result of our statistical model and, in particular, the DP prior imposed on the hyper-parameters $\boldsymbol{\theta}_n$, the $\mathbf{x}_n$ are dependent across the object index n; indeed, $\mathbf{x}_n$ and $\mathbf{x}_{n'}$ belong to the same cluster if $\boldsymbol{\theta}_n = \boldsymbol{\theta}_{n'}$ for $n \neq n'$. Therefore, it is again impossible to sample from (5.18) directly. We can expand the integrand in (5.18), i.e., the joint posterior pdf $f_{\mathbf{x}_{1:N} \,|\, \mathbf{y}_{1:N}}(\boldsymbol{x}_{1:N} \,|\, \boldsymbol{y}_{1:N})$ according to

$$f_{\mathbf{x}_{1:N} \,|\, \mathbf{y}_{1:N}}(\boldsymbol{x}_{1:N} \,|\, \boldsymbol{y}_{1:N}) = \int_{\boldsymbol{\theta}_{1:N}} f_{\mathbf{x}_{1:N}, \boldsymbol{\theta}_{1:N} \,|\, \mathbf{y}_{1:N}}(\boldsymbol{x}_{1:N}, \boldsymbol{\theta}_{1:N} \,|\, \boldsymbol{y}_{1:N}) d\boldsymbol{\theta}_{1:N}. \tag{5.19}$$

By inserting (5.19) into (5.18) we obtain

$$f_{\mathbf{x}_n \,|\, \mathbf{y}_{1:N}}(\boldsymbol{x}_n \,|\, \boldsymbol{y}_{1:N}) = \int_{\boldsymbol{x}_{\neg n}} \int_{\boldsymbol{\theta}_{1:N}} f_{\mathbf{x}_{1:N}, \boldsymbol{\theta}_{1:N} \,|\, \mathbf{y}_{1:N}}(\boldsymbol{x}_{1:N}, \boldsymbol{\theta}_{1:N} \,|\, \boldsymbol{y}_{1:N}) d\boldsymbol{\theta}_{1:N} d\boldsymbol{x}_{\neg n}. \tag{5.20}$$

We will show that by using the Gibbs sampler, we can obtain samples $\boldsymbol{x}_n^{(q)}$ from the joint pdf $f_{\mathbf{x}_{1:N}, \boldsymbol{\theta}_{1:N} \,|\, \mathbf{y}_{1:N}}(\boldsymbol{x}_{1:N}, \boldsymbol{\theta}_{1:N} \,|\, \boldsymbol{y}_{1:N})$; however, we will also need to obtain samples $\boldsymbol{\theta}_n^{(q)}$, even though we are only interested in $\boldsymbol{x}_n^{(q)}$. The approach of sampling an additional random

variable $\mathsf{\Theta}_n$ allows us to obtain a closed-form expression for both full conditional pdfs of x_n and $\mathsf{\Theta}_n$, which are needed in the Gibbs sampler algorithm. Sampling an additional random variable (in our case $\mathsf{\Theta}_n$) to facilitate sampling of the parameter of interest (in our case x_n) is sometimes referred to as *linchpin variable sampler* [31]. Once we obtain Q pairs of samples $\boldsymbol{x}_n^{(q)}$ and $\boldsymbol{\theta}_n^{(q)}$ for $n = 1, \ldots, N$, the marginalization in (5.20) is done simply by discarding all samples of the linchpin variable, i.e., $\boldsymbol{\theta}_n^{(q)}$ for $n = 1, \ldots, N$, and keeping only the samples $\boldsymbol{x}_n^{(q)}$, which are then used in (5.17). For each random variable (in our case $\boldsymbol{x}_n^{(q)}$ and $\boldsymbol{\theta}_n^{(q)}$), we will derive the full conditional pdf in what follows. During each iteration $q \in \{1, \ldots, Q\}$, the Gibbs sampler "loops" over the object index $n = 1, \ldots, N$ and calculates new samples $\boldsymbol{x}_n^{(q)}$ and $\boldsymbol{\theta}_n^{(q)}$ for each $n = 1, \ldots, N$, using the samples from the previous iteration $q - 1$ or, if already available, it uses the samples from the current iteration q.

We will consider two versions of the Gibbs sampler: a "simple" Gibbs sampler that samples $\boldsymbol{x}_n^{(q)}$ and $\boldsymbol{\theta}_n^{(q)}$ for each object index $n = 1, \ldots, N$ separately, and a more sophisticated Gibbs sampler that uses the cluster assignment variables $\mathsf{C}_{1:N}$.

5.2.2 Simple Gibbs Sampler

In each iteration q of the "simple" Gibbs sampler algorithm, we obtain the samples $\boldsymbol{\theta}_n^{(q)}$ and $\boldsymbol{x}_n^{(q)}$ for each $n = 1, \ldots, N$ by sampling from their respective full conditional pdfs. The full conditional pdf of $\mathsf{\Theta}_n$ is the pdf of $\mathsf{\Theta}_n$ conditioned on all the other random variables, i.e., the remaining hyperparameters, $\mathsf{\Theta}_{\neg n} = (\mathsf{\Theta}_1^{\mathrm{T}}, \ldots, \mathsf{\Theta}_{n-1}^{\mathrm{T}}, \mathsf{\Theta}_{n+1}^{\mathrm{T}}, \ldots, \mathsf{\Theta}_N^{\mathrm{T}})^{\mathrm{T}}$, the parameters of interest, $\mathsf{x}_{1:N}$, and the measurements, $\mathsf{y}_{1:N}$, i.e., $f_{\mathsf{\Theta}_n \,|\, \mathsf{\Theta}_{\neg n}, \mathsf{x}_{1:N}, \mathsf{y}_{1:N}}(\boldsymbol{\theta}_n \,|\, \boldsymbol{\theta}_{\neg n}, \boldsymbol{x}_{1:N}, \boldsymbol{y}_{1:N})$. Similarly, the full conditional pdf of x_n is the pdf of x_n, conditioned on the remaining parameters of interest $\mathsf{x}_{\neg n} = (\mathsf{x}_1^{\mathrm{T}}, \ldots, \mathsf{x}_{n-1}^{\mathrm{T}}, \mathsf{x}_{n+1}^{\mathrm{T}}, \ldots, \mathsf{x}_N^{\mathrm{T}})^{\mathrm{T}}$, the hyperparameters $\mathsf{\Theta}_{1:N}$, and the measurements $\mathsf{y}_{1:N}$, i.e., $f_{\mathsf{x}_n \,|\, \mathsf{x}_{\neg n}, \mathsf{\Theta}_{1:N}, \mathsf{y}_{1:N}}(\boldsymbol{x}_n \,|\, \boldsymbol{x}_{\neg n}, \boldsymbol{\theta}_{1:N}, \boldsymbol{y}_{1:N})$. These full conditional pdf are evaluated using samples from the previous iteration $q - 1$ or, if already available, from the current iteration q. We accordingly define

$$\boldsymbol{\theta}_{\neg n}^{(q,q-1)} \triangleq \left(\boldsymbol{\theta}_1^{(q)\mathrm{T}}, \ldots, \boldsymbol{\theta}_{n-1}^{(q)\mathrm{T}}, \boldsymbol{\theta}_{n+1}^{(q-1)\mathrm{T}}, \ldots, \boldsymbol{\theta}_N^{(q-1)\mathrm{T}}\right)^{\mathrm{T}}, \tag{5.21}$$

as the vector containing the hyperparameter samples $\boldsymbol{\theta}_{n'}^{(q)}$ from the current iteration q for all $n' = 1, \ldots, n-1$ and the hyperparameter samples $\boldsymbol{\theta}_{n'}^{(q-1)}$ from the previous iteration for

$n' = n+1, \ldots, N$. Similarly, we define

$$\boldsymbol{x}_{\neg n}^{(q,q-1)} \triangleq \left(\boldsymbol{x}_1^{(q)\mathrm{T}}, \ldots, \boldsymbol{x}_{n-1}^{(q)\mathrm{T}}, \boldsymbol{x}_{n+1}^{(q-1)\mathrm{T}}, \ldots, \boldsymbol{x}_N^{(q-1)\mathrm{T}}\right)^{\mathrm{T}}. \tag{5.22}$$

In each iteration q of the simple Gibbs sampler algorithm, we will sample the random variables in the following order:

1. Obtain samples $\boldsymbol{\theta}_n^{(q)}$ of the hyperparameter θ_n for all n according to

$$\theta_n^{(q)} \sim f_{\theta_n \mid \theta_{\neg n}, \mathbf{x}_{1:N}, \mathbf{y}_{1:N}}(\boldsymbol{\theta}_n^{(q)} \mid \boldsymbol{\theta}_{\neg n}^{(q,q-1)}, \boldsymbol{x}_{1:N}^{(q-1)}, \boldsymbol{y}_{1:N}), \tag{5.23}$$

   using $\boldsymbol{\theta}_{\neg n}^{(q,q-1)}$ (see (5.21)) and the samples $\boldsymbol{x}_{1:N}^{(q-1)}$ from the previous iteration.

2. Obtain samples $\boldsymbol{x}_n^{(q)}$ of the parameter of interest $\mathbf{x}_n$ for all n according to

$$\mathbf{x}_n^{(q)} \sim f_{\mathbf{x}_n \mid \mathbf{x}_{\neg n}, \theta_{1:N}, \mathbf{y}_{1:N}}(\boldsymbol{x}_n^{(q)} \mid \boldsymbol{x}_{\neg n}^{(q,q-1)}, \boldsymbol{\theta}_{1:N}^{(q)}, \boldsymbol{y}_{1:N}), \tag{5.24}$$

   using $\boldsymbol{x}_{\neg n}^{(q,q-1)}$ (see (5.22)) and samples the $\boldsymbol{\theta}_{1:N}^{(q)}$ from the current iteration q.

Full Conditional pdf of θ_n

We will now derive the full conditional pdf $f_{\theta_n \mid \theta_{\neg n}, \mathbf{x}_{1:N}, \mathbf{y}_{1:N}}(\boldsymbol{\theta}_n \mid \boldsymbol{\theta}_{\neg n}, \boldsymbol{x}_{1:N}, \boldsymbol{y}_{1:N})$ (see (5.23)). By Bayes' theorem, we have

$$f_{\theta_n \mid \theta_{\neg n}, \mathbf{x}_{1:N}, \mathbf{y}_{1:N}}(\boldsymbol{\theta}_n \mid \boldsymbol{\theta}_{\neg n}, \boldsymbol{x}_{1:N}, \boldsymbol{y}_{1:N}) \propto f_{\mathbf{x}_{1:N}, \mathbf{y}_{1:N} \mid \theta_n, \theta_{\neg n}}(\boldsymbol{x}_{1:N}, \boldsymbol{y}_{1:N} \mid \boldsymbol{\theta}_n, \boldsymbol{\theta}_{\neg n}) f_{\theta_n \mid \theta_{\neg n}}(\boldsymbol{\theta}_n \mid \boldsymbol{\theta}_{\neg n}). \tag{5.25}$$

Here, the first factor in (5.25) can be factorized as

$$\begin{aligned} &f_{\mathbf{x}_{1:N}, \mathbf{y}_{1:N} \mid \theta_n, \theta_{\neg n}}(\boldsymbol{x}_{1:N}, \boldsymbol{y}_{1:N} \mid \boldsymbol{\theta}_n, \boldsymbol{\theta}_{\neg n}) \\ &= f_{\mathbf{y}_{1:N} \mid \mathbf{x}_{1:N}, \theta_n, \theta_{\neg n}}(\boldsymbol{y}_{1:N} \mid \boldsymbol{x}_{1:N}, \boldsymbol{\theta}_n, \boldsymbol{\theta}_{\neg n}) f_{\mathbf{x}_{1:N} \mid \theta_n, \theta_{\neg n}}(\boldsymbol{x}_{1:N} \mid \boldsymbol{\theta}_n, \boldsymbol{\theta}_{\neg n}). \end{aligned} \tag{5.26}$$

Noting that $(\theta_n^{\mathrm{T}}, \theta_{\neg n}^{\mathrm{T}})^{\mathrm{T}} = \theta_{1:N}$ and using (4.18) and (4.20), we obtain further

$$\begin{aligned} &f_{\mathbf{x}_{1:N}, \mathbf{y}_{1:N} \mid \theta_n, \theta_{\neg n}}(\boldsymbol{x}_{1:N}, \boldsymbol{y}_{1:N} \mid \boldsymbol{\theta}_n, \boldsymbol{\theta}_{\neg n}) \\ &\overset{(4.18)}{=} f_{\mathbf{y}_{1:N} \mid \mathbf{x}_{1:N}}(\boldsymbol{y}_{1:N} \mid \boldsymbol{x}_{1:N}) f_{\mathbf{x}_{1:N} \mid \theta_n, \theta_{\neg n}}(\boldsymbol{x}_{1:N} \mid \boldsymbol{\theta}_n, \boldsymbol{\theta}_{\neg n}) \end{aligned}$$

$$\overset{(4.20)}{=} f_{\mathbf{y}_{1:N}\,|\,\mathbf{x}_{1:N}}(\boldsymbol{y}_{1:N}\,|\,\boldsymbol{x}_{1:N}) f_{\mathbf{x}_n\,|\,\boldsymbol{\theta}_n}(\boldsymbol{x}_n\,|\,\boldsymbol{\theta}_n) \prod_{n'\in\{1,\ldots,N\}\setminus\{n\}} f_{\mathbf{x}_{n'}\,|\,\boldsymbol{\theta}_{n'}}(\boldsymbol{x}_{n'}\,|\,\boldsymbol{\theta}_{n'}). \tag{5.27}$$

Next, we note that only the second factor $f_{\mathbf{x}_n\,|\,\boldsymbol{\theta}_n}(\boldsymbol{x}_n\,|\,\boldsymbol{\theta}_n)$ in (5.27) is functionally dependent on $\boldsymbol{\theta}_n$, therefore we have

$$f_{\mathbf{x}_{1:N},\mathbf{y}_{1:N}\,|\,\boldsymbol{\theta}_n,\boldsymbol{\theta}_{\neg n}}(\boldsymbol{x}_{1:N},\boldsymbol{y}_{1:N}\,|\,\boldsymbol{\theta}_n,\boldsymbol{\theta}_{\neg n}) \propto f_{\mathbf{x}_n\,|\,\boldsymbol{\theta}_n}(\boldsymbol{x}_n\,|\,\boldsymbol{\theta}_n). \tag{5.28}$$

Inserting (5.28) into (5.25), we obtain

$$f_{\boldsymbol{\theta}_n\,|\,\boldsymbol{\theta}_{\neg n},\mathbf{x}_{1:N},\mathbf{y}_{1:N}}(\boldsymbol{\theta}_n\,|\,\boldsymbol{\theta}_{\neg n},\boldsymbol{x},\boldsymbol{y}_{1:N}) \propto f_{\mathbf{x}_n\,|\,\boldsymbol{\theta}_n}(\boldsymbol{x}_n\,|\,\boldsymbol{\theta}_n) f_{\boldsymbol{\theta}_n\,|\,\boldsymbol{\theta}_{\neg n}}(\boldsymbol{\theta}_n\,|\,\boldsymbol{\theta}_{\neg n}). \tag{5.29}$$

We now consider the second factor in (5.29). By adapting expression (3.23), we obtain in a straightforward manner

$$f_{\boldsymbol{\theta}_n\,|\,\boldsymbol{\theta}_{\neg n}}(\boldsymbol{\theta}_n\,|\,\boldsymbol{\theta}_{\neg n}) = \frac{\alpha}{\alpha+N-1} f_{\mathrm{H}}(\boldsymbol{\theta}_n) + \frac{1}{\alpha+N-1} \sum_{n'\in\{1,\ldots,N\}\setminus\{n\}} \delta_{\boldsymbol{\theta}_{n'}}(\boldsymbol{\theta}_n). \tag{5.30}$$

Finally, using (5.30) in (5.29), the full conditional distribution can be given by

$$\begin{aligned}
&f_{\boldsymbol{\theta}_n\,|\,\boldsymbol{\theta}_{\neg n},\mathbf{x}_{1:N},\mathbf{y}_{1:N}}(\boldsymbol{\theta}_n\,|\,\boldsymbol{\theta}_{\neg n},\boldsymbol{x}_{1:N},\boldsymbol{y}_{1:N}) \\
&\propto \frac{\alpha}{\alpha+N-1} f_{\mathbf{x}_n\,|\,\boldsymbol{\theta}_n}(\boldsymbol{x}_n\,|\,\boldsymbol{\theta}_n) f_{\mathrm{H}}(\boldsymbol{\theta}_n) + \frac{1}{\alpha+N-1} \sum_{n'\in\{1,\ldots,N\}\setminus\{n\}} f_{\mathbf{x}_n\,|\,\boldsymbol{\theta}_n}(\boldsymbol{x}_n\,|\,\boldsymbol{\theta}_n)\delta_{\boldsymbol{\theta}_{n'}}(\boldsymbol{\theta}_n) \\
&\propto \alpha f_{\mathbf{x}_n\,|\,\boldsymbol{\theta}_n}(\boldsymbol{x}_n\,|\,\boldsymbol{\theta}_n) f_{\mathrm{H}}(\boldsymbol{\theta}_n) + \sum_{n'\in\{1,\ldots,N\}\setminus\{n\}} f_{\mathbf{x}_n\,|\,\boldsymbol{\theta}_n}(\boldsymbol{x}_n\,|\,\boldsymbol{\theta}_{n'})\delta_{\boldsymbol{\theta}_{n'}}(\boldsymbol{\theta}_n).
\end{aligned} \tag{5.31}$$

This expression shows that the full conditional pdf of $\boldsymbol{\theta}_n$ does not functionally depend on the measurements $\mathbf{y}_{1:N}$. This follows from the hierarchical Markov chain generation $\boldsymbol{\theta}_n \to \mathbf{x}_n \to \mathbf{y}_n$ and also can be seen in our statistical model shown in Figure 8. Furthermore, the full conditional pdf (5.31) is seen to be a discrete-continuous mixture distribution; more specifically, $\boldsymbol{\theta}_n$ is either drawn from the continuous mixture component $f_{\mathbf{x}_n\,|\,\boldsymbol{\theta}_n}(\boldsymbol{x}_n\,|\,\boldsymbol{\theta}_n) f_{\mathrm{H}}(\boldsymbol{\theta}_n)$, which means $\boldsymbol{\theta}_n$ belongs to a newly created cluster, or $\boldsymbol{\theta}_n$ equals to one of the $N-1$ previously drawn $\boldsymbol{\theta}_{n'}$ constituted by discrete mixture components $\delta_{\boldsymbol{\theta}_{n'}}(\boldsymbol{\theta}_n)$, weighted by $\propto \frac{1}{\alpha+N-1} f_{\mathbf{x}_n\,|\,\boldsymbol{\theta}_n}(\boldsymbol{x}_n\,|\,\boldsymbol{\theta}_{n'})$.

Using (5.2) and (5.8), the continuous component of the mixture in (5.31) is given by

$$f_{\mathbf{x}_n \,|\, \boldsymbol{\theta}_n}(\boldsymbol{x}_n \,|\, \boldsymbol{\theta}_n) f_{\mathrm{H}}(\boldsymbol{\theta}_n) = \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\theta}_n; \boldsymbol{\Sigma}_{\boldsymbol{u}}) \mathcal{N}(\boldsymbol{\theta}_n; \boldsymbol{\mu}_{\boldsymbol{\theta}^*}; \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}). \tag{5.32}$$

Using the identity established by (2.69) with the substitutions $\boldsymbol{x} \to \boldsymbol{\theta}_n$, $\boldsymbol{\mu}_{\boldsymbol{x}} \to \boldsymbol{\mu}_{\boldsymbol{\theta}^*}$, $\boldsymbol{\Sigma}_{\boldsymbol{x}} \to \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}$, $\boldsymbol{y} \to \boldsymbol{x}_n$ and $\boldsymbol{\Sigma}_{\boldsymbol{y}} \to \boldsymbol{\Sigma}_{\boldsymbol{u}}$, we see that (5.32) is proportional to a Gaussian pdf in $\boldsymbol{\theta}_n$, i.e.,

$$f_{\mathbf{x}_n \,|\, \boldsymbol{\theta}_n}(\boldsymbol{x}_n \,|\, \boldsymbol{\theta}_n) f_{\mathrm{H}}(\boldsymbol{\theta}_n) = \tilde{\gamma}(\boldsymbol{x}_n) \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}_{\boldsymbol{\theta} \,|\, \boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta} \,|\, \boldsymbol{x}}) \tag{5.33}$$

with covariance matrix (see (2.67))

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta} \,|\, \boldsymbol{x}} = \boldsymbol{\Sigma}_{\boldsymbol{u}} \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*} \right)^{-1} \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*} \tag{5.34}$$

and mean (see (2.68))

$$\boldsymbol{\mu}_{\boldsymbol{\theta} \,|\, \boldsymbol{x}} = \boldsymbol{\Sigma}_{\boldsymbol{u}} \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*} \right)^{-1} \boldsymbol{\mu}_{\boldsymbol{\theta}^*} + \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*} \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*} \right)^{-1} \boldsymbol{x}_n. \tag{5.35}$$

Furthermore, we obtain

$$\tilde{\gamma}(\boldsymbol{x}_n) = \mathcal{N}\left(\boldsymbol{x}_n; \boldsymbol{\mu}_{\boldsymbol{\theta}^*}, \boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*} \right), \tag{5.36}$$

which is independent of $\boldsymbol{\theta}_n$. Inserting (5.33) and (5.8) into (5.31), the full conditional pdf of $\boldsymbol{\theta}_n$ is thus finally obtained as

$$\begin{aligned} &f_{\boldsymbol{\theta}_n \,|\, \boldsymbol{\theta}_{\neg n}, \mathbf{x}_{1:N}, \mathbf{y}_{1:N}}(\boldsymbol{\theta}_n \,|\, \boldsymbol{\theta}_{\neg n}, \boldsymbol{x}_{1:N}, \boldsymbol{y}_{1:N}) \\ &\propto \alpha \tilde{\gamma}(\boldsymbol{x}_n) \mathcal{N}(\boldsymbol{\theta}_n; \boldsymbol{\mu}_{\boldsymbol{\theta} \,|\, \boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta} \,|\, \boldsymbol{x}}) + \sum_{n' \in \{1,\ldots,N\} \setminus \{n\}} \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\theta}_{n'}, \boldsymbol{\Sigma}_{\boldsymbol{u}}) \delta_{\boldsymbol{\theta}_{n'}}(\boldsymbol{\theta}_n), \end{aligned} \tag{5.37}$$

with $\boldsymbol{\Sigma}_{\boldsymbol{\theta} \,|\, \boldsymbol{x}}$ and $\boldsymbol{\mu}_{\boldsymbol{\theta} \,|\, \boldsymbol{x}}$ as given by (5.34) and (5.35).

For use in the Gibbs sampler, we need to evaluate expression (5.37) at the already available samples (see (5.23)). We obtain

$$f_{\boldsymbol{\theta}_n \,|\, \boldsymbol{\theta}_{\neg n}, \mathbf{x}_{1:N}, \mathbf{y}_{1:N}}(\boldsymbol{\theta}_n^{(q)} \,|\, \boldsymbol{\theta}_{\neg n}^{(q,q-1)}, \boldsymbol{x}_{1:N}^{(q-1)}, \boldsymbol{y}_{1:N})$$

$$\propto \alpha\tilde{\gamma}(\boldsymbol{x}_n^{(q-1)})\mathcal{N}(\boldsymbol{\theta}_n^{(q)}; \boldsymbol{\mu}_{\boldsymbol{\theta}\,|\,\boldsymbol{x}^{(q-1)}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}\,|\,\boldsymbol{x}}) + \sum_{n'\in\{1,\ldots,N\}\backslash\{n\}} \mathcal{N}(\boldsymbol{x}_n^{(q-1)}; \boldsymbol{\theta}_{n',n}^{(q,q-1)}, \boldsymbol{\Sigma}_{\boldsymbol{u}})\delta_{\boldsymbol{\theta}_{n',n}^{(q,q-1)}}(\boldsymbol{\theta}_n^{(q)}), \tag{5.38}$$

with $\boldsymbol{\theta}_{n',n}^{(q,q-1)}$ defined as

$$\boldsymbol{\theta}_{n',n}^{(q,q-1)} \triangleq \begin{cases} \boldsymbol{\theta}_{n'}^{(q)} & \text{for} \quad n' = 1,\ldots,n-1 \\ \boldsymbol{\theta}_{n'}^{(q-1)} & \text{for} \quad n' = n+1,\ldots,N. \end{cases} \tag{5.39}$$

and

$$\boldsymbol{\mu}_{\boldsymbol{\theta}\,|\,\boldsymbol{x}}^{(q-1)} = \boldsymbol{\Sigma}_{\boldsymbol{u}}\left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}\right)^{-1}\boldsymbol{\mu}_{\boldsymbol{\theta}^*} + \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}\left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}\right)^{-1}\boldsymbol{x}_n^{(q-1)}. \tag{5.40}$$

Lastly, we normalize (5.38), i.e.,

$$\begin{aligned} &f_{\theta_n\,|\,\boldsymbol{\theta}_{\neg n},\mathbf{x}_{1:N},\mathbf{y}_{1:N}}(\boldsymbol{\theta}_n^{(q)}\,|\,\boldsymbol{\theta}_{\neg n}^{(q,q-1)}, \boldsymbol{x}_{1:N}^{(q-1)}, \boldsymbol{y}_{1:N}) \\ &= \pi_n\,\mathcal{N}(\boldsymbol{\theta}_n^{(q)}; \boldsymbol{\mu}_{\boldsymbol{\theta}\,|\,\boldsymbol{x}^{(q-1)}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}\,|\,\boldsymbol{x}}) + \sum_{n'\in\{1,\ldots,N\}\backslash\{n\}} \pi_{n'}\,\delta_{\boldsymbol{\theta}_{n',n}^{(q,q-1)}}(\boldsymbol{\theta}_n^{(q-1)}), \end{aligned} \tag{5.41}$$

with

$$\pi_n = \frac{\alpha\tilde{\gamma}(\boldsymbol{x}_n^{(q-1)})}{\alpha\tilde{\gamma}(\boldsymbol{x}_n^{(q-1)}) + \sum_{n'\in\{1,\ldots,N\}\backslash\{n\}}\mathcal{N}(\boldsymbol{x}_n^{(q-1)}; \boldsymbol{\theta}_{n',n}^{(q,q-1)}, \boldsymbol{\Sigma}_{\boldsymbol{u}})} \tag{5.42}$$

and

$$\pi_{n'} = \frac{\mathcal{N}(\boldsymbol{x}_n^{(q-1)}; \boldsymbol{\theta}_{n',n}^{(q,q-1)}, \boldsymbol{\Sigma}_{\boldsymbol{u}})}{\alpha\tilde{\gamma}(\boldsymbol{x}_n^{(q-1)}) + \sum_{n'\in\{1,\ldots,N\}\backslash\{n\}}\mathcal{N}(\boldsymbol{x}_n^{(q-1)}; \boldsymbol{\theta}_{n',n}^{(q,q-1)}, \boldsymbol{\Sigma}_{\boldsymbol{u}})}, \quad \text{for} \quad n' \in \{1,\ldots,N\}\backslash\{n\}. \tag{5.43}$$

The pseudocode for sampling from a discrete-continuous mixture distribution is given in Algorithm 1. For a detailed discussion of mixture distributions, we refer to [32].

Full conditional pdf of $\mathbf{x}_n$

Next, we will derive the full conditional pdf of $\mathbf{x}_n$, given all the other random variables, i.e. $f_{\mathbf{x}_n\,|\,\mathbf{x}_{\neg n},\boldsymbol{\theta}_{1:N},\mathbf{y}_{1:N}}(\boldsymbol{x}_n\,|\,\boldsymbol{x}_{\neg n}, \boldsymbol{\theta}_{1:N}, \boldsymbol{y}_{1:N})$ (see (5.24)). Using Bayes' theorem, we obtain

$$\begin{aligned} &f_{\mathbf{x}_n\,|\,\mathbf{x}_{\neg n},\boldsymbol{\theta}_{1:N},\mathbf{y}_{1:N}}(\boldsymbol{x}_n\,|\,\boldsymbol{x}_{\neg n}, \boldsymbol{\theta}_{1:N}, \boldsymbol{y}_{1:N}) \\ &= \frac{f_{\mathbf{y}_{1:N}\,|\,\mathbf{x}_n,\mathbf{x}_{\neg n},\boldsymbol{\theta}_{1:N}}(\boldsymbol{y}_{1:N}\,|\,\boldsymbol{x}_n, \boldsymbol{x}_{\neg n}, \boldsymbol{\theta}_{1:N}) f_{\mathbf{x}_n\,|\,\mathbf{x}_{\neg n},\boldsymbol{\theta}_{1:N}}(\boldsymbol{x}_n\,|\,\boldsymbol{x}_{\neg n}, \boldsymbol{\theta}_{1:N})}{f_{\mathbf{y}_{1:N}\,|\,\mathbf{x}_{\neg n},\boldsymbol{\theta}_{1:N}}(\boldsymbol{y}_{1:N}\,|\,\boldsymbol{x}_{\neg n}, \boldsymbol{\theta}_{1:N})}. \end{aligned} \tag{5.44}$$

Algorithm 1 Discrete-continuous mixture sampling

Input: $\boldsymbol{\theta}_{1:N}^{(q-1)}, \boldsymbol{x}_{1:N}^{(q-1)}, \boldsymbol{y}_{1:N}$
  for all $n = 1, \dots, N$ **do**
    sample $b_n^{(q)}$ from $\mathcal{U}(b_n^{(q)}; 0, 1)$
    if $b_n^{(q)} \in \left[\sum_{i=1}^{n'-1} \pi_i, \sum_{i=1}^{n'} \pi_i\right)$ **then**
      set $\boldsymbol{\theta}_n^{(q)}$ equal to $\boldsymbol{\theta}_{n',n}^{(q,q-1)}$ as defined in (5.39)
    else
      sample $\boldsymbol{\theta}_n^{(q)}$ from $\mathcal{N}(\boldsymbol{\theta}_n^{(q)}; \boldsymbol{\mu}_{\boldsymbol{\theta} \mid \boldsymbol{x}^{(q-1)}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta} \mid \boldsymbol{x}})$
    end if
  end for
Output: $\boldsymbol{\theta}_{1:N}^{(q)}$

Noting that $(\mathbf{x}_n^{\mathrm{T}}, \mathbf{x}_{\neg n}^{\mathrm{T}})^{\mathrm{T}} = \mathbf{x}_{1:N}$ as well as using (4.17) and the conditional independence of $\mathbf{x}_n$ (see (5.6)), the denominator of (5.44) simplifies to

$$\begin{aligned} f_{\mathbf{y}_{1:N} \mid \mathbf{x}_{\neg n}, \boldsymbol{\theta}_{1:N}}(\boldsymbol{y}_{1:N} \mid \boldsymbol{x}_{\neg n}, \boldsymbol{\theta}_{1:N}) &= \int_{\boldsymbol{x}_n} f_{\mathbf{y}_{1:N} \mid \mathbf{x}_{1:N}, \boldsymbol{\theta}_{1:N}}(\boldsymbol{y}_{1:N} \mid \boldsymbol{x}_{1:N}, \boldsymbol{\theta}_{1:N}) f_{\mathbf{x}_n \mid \boldsymbol{\theta}_{1:N}}(\boldsymbol{x}_n \mid \boldsymbol{\theta}_{1:N}) d\boldsymbol{x}_n \\ &= \int_{\boldsymbol{x}_n} f_{\mathbf{y}_n \mid \mathbf{x}_n}(\boldsymbol{y}_n \mid \boldsymbol{x}_n) f_{\mathbf{x}_n \mid \boldsymbol{\theta}_n}(\boldsymbol{x}_n \mid \boldsymbol{\theta}_n) d\boldsymbol{x}_n \prod_{n' \neq n} f_{\mathbf{y}_{n'} \mid \mathbf{x}_{n'}}(\boldsymbol{y}_{n'} \mid \boldsymbol{x}_{n'}) \\ &= f_{\mathbf{y}_n \mid \boldsymbol{\theta}_n}(\boldsymbol{y}_n \mid \boldsymbol{\theta}_n) \prod_{n' \neq n} f_{\mathbf{y}_{n'} \mid \mathbf{x}_{n'}}(\boldsymbol{y}_{n'} \mid \boldsymbol{x}_{n'}). \end{aligned} \tag{5.45}$$

Using (4.17) to factorize the nominator of (5.44) and inserting (5.45) into (5.44), we obtain

$$\begin{aligned} &f_{\mathbf{x}_n \mid \mathbf{x}_{\neg n}, \boldsymbol{\theta}_{1:N}, \mathbf{y}_{1:N}}(\boldsymbol{x}_n \mid \boldsymbol{x}_{\neg n}, \boldsymbol{\theta}_{1:N}, \boldsymbol{y}_{1:N}) \\ &= \frac{f_{\mathbf{y}_n \mid \mathbf{x}_n}(\boldsymbol{y}_n \mid \boldsymbol{x}_n) \left(\prod_{n' \neq n} f_{\mathbf{y}_{n'} \mid \mathbf{x}_{n'}}(\boldsymbol{y}_{n'} \mid \boldsymbol{x}_{n'})\right) f_{\mathbf{x}_n \mid \mathbf{x}_{\neg n}, \boldsymbol{\theta}_{1:N}}(\boldsymbol{x}_n \mid \boldsymbol{x}_{\neg n}, \boldsymbol{\theta}_{1:N})}{f_{\mathbf{y}_n \mid \boldsymbol{\theta}_n}(\boldsymbol{y}_n \mid \boldsymbol{\theta}_n) \prod_{n' \neq n} f_{\mathbf{y}_{n'} \mid \mathbf{x}_{n'}}(\boldsymbol{y}_{n'} \mid \boldsymbol{x}_{n'})} \\ &= \frac{f_{\mathbf{y}_n \mid \mathbf{x}_n}(\boldsymbol{y}_n \mid \boldsymbol{x}_n) f_{\mathbf{x}_n \mid \boldsymbol{\theta}_{1:N}}(\boldsymbol{x}_n \mid \boldsymbol{\theta}_{1:N})}{f_{\mathbf{y}_n \mid \boldsymbol{\theta}_n}(\boldsymbol{y}_n \mid \boldsymbol{\theta}_n)}. \end{aligned} \tag{5.46}$$

where in the last step (5.5) was used. Finally, using (4.12) and (5.6) we obtain

$$\begin{aligned} f_{\mathbf{x}_n \mid \mathbf{x}_{\neg n}, \boldsymbol{\theta}_{1:N}, \mathbf{y}_{1:N}}(\boldsymbol{x}_n \mid \boldsymbol{x}_{\neg n}, \boldsymbol{\theta}_{1:N}, \boldsymbol{y}_{1:N}) &= \frac{f_{\mathbf{y}_n \mid \mathbf{x}_n, \boldsymbol{\theta}_n}(\boldsymbol{y}_n \mid \boldsymbol{x}_n, \boldsymbol{\theta}_n) f_{\mathbf{x}_n \mid \boldsymbol{\theta}_n}(\boldsymbol{x} \mid \boldsymbol{\theta}_n)}{f_{\mathbf{y}_n \mid \boldsymbol{\theta}_n}(\boldsymbol{y}_n \mid \boldsymbol{\theta}_n)} \\ &= f_{\mathbf{x}_n \mid \mathbf{y}_n, \boldsymbol{\theta}_n}(\boldsymbol{x}_n \mid \boldsymbol{y}_n, \boldsymbol{\theta}_n) \\ &= \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}_{\boldsymbol{x} \mid \boldsymbol{y}_n, \boldsymbol{\theta}_n}, \boldsymbol{\Sigma}_{\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\theta}}), \end{aligned} \tag{5.47}$$

where Bayes' theorem and (4.79) were used. By (5.47) the full conditional pdf of $\mathbf{x}_n$ is a Gaussian pdf with covariance matrix (see (4.83))

$$\begin{aligned}\boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y},\boldsymbol{\theta}} &= \boldsymbol{\Sigma}_{\boldsymbol{u}} - \boldsymbol{\Sigma}_{\boldsymbol{u}}(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}\boldsymbol{\Sigma}_{\boldsymbol{u}} \\ &\overset{\text{(B.4)}}{=} \boldsymbol{\Sigma}_{\boldsymbol{v}}(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}\boldsymbol{\Sigma}_{\boldsymbol{u}},\end{aligned} \tag{5.48}$$

and mean (see (4.84))

$$\boldsymbol{\mu}_{\boldsymbol{x}\,|\,\boldsymbol{y}_n,\boldsymbol{\theta}_n} = \boldsymbol{\theta}_n + \boldsymbol{\Sigma}_{\boldsymbol{u}}(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}(\boldsymbol{y}_n - \boldsymbol{\theta}_n). \tag{5.49}$$

By evaluating (5.47) at the already available samples (see (5.24)), we finally obtain

$$f_{\mathbf{x}_n\,|\,\mathbf{x}_{\neg n},\boldsymbol{\theta}_{1:N},\mathbf{y}_{1:N}}(\boldsymbol{x}_n^{(q)}\,|\,\boldsymbol{x}_{\neg n}^{(q,q-1)},\boldsymbol{\theta}_{1:N}^{(q)},\boldsymbol{y}_{1:N}) = \mathcal{N}(\boldsymbol{x}_n^{(q)};\boldsymbol{\mu}_{\boldsymbol{x}\,|\,\boldsymbol{y}_n,\boldsymbol{\theta}_n}^{(q)},\boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y},\boldsymbol{\theta}}), \tag{5.50}$$

with

$$\boldsymbol{\mu}_{\boldsymbol{x}\,|\,\boldsymbol{y}_n,\boldsymbol{\theta}_n}^{(q)} = \boldsymbol{\theta}_n^{(q)} + \boldsymbol{\Sigma}_{\boldsymbol{u}}(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}(\boldsymbol{y}_n - \boldsymbol{\theta}_n^{(q)}). \tag{5.51}$$

Pseudocode for the Simple Gibbs Sampler

Finally, based on (5.23) and (5.24) we can formulate the Gibbs sampler for our scenario, using (5.37) to generate samples $\boldsymbol{\theta}_n^{(q)}$ and (5.50) to generate samples $\boldsymbol{x}_n^{(q)}$ for all $n = 1,\ldots,N$. The pseudocode for the qth iteration of the Gibbs sampler is given in Algorithm 2.

Algorithm 2 Naive Gibbs sampler

Input: $\boldsymbol{\theta}_{1:N}^{(q-1)}, \boldsymbol{x}_{1:N}^{(q-1)}, \boldsymbol{y}_{1:N}$
  for all $n = 1,\ldots,N$ **do**
    sample $\boldsymbol{\theta}_n^{(q)}$ from $f_{\theta_n\,|\,\boldsymbol{\theta}_{\neg n},\mathbf{x}_{1:N},\mathbf{y}_{1:N}}(\boldsymbol{\theta}_n^{(q)}\,|\,\boldsymbol{\theta}_{\neg n}^{(q,q-1)},\boldsymbol{x}_{1:N}^{(q-1)},\boldsymbol{y}_{1:N})$ as given by (5.37)
  end for
  for all $n = 1,\ldots,N$ **do**
    sample $\boldsymbol{x}_n^{(q)}$ from $f_{\mathbf{x}_n\,|\,\mathbf{x}_{\neg n},\boldsymbol{\theta}_{1:N},\mathbf{y}_{1:N}}(\boldsymbol{x}_n^{(q)}\,|\,\boldsymbol{x}_{\neg n}^{(q,q-1)},\boldsymbol{\theta}_{1:N}^{(q)},\boldsymbol{y}_{1:N})$ as given by (5.50)
  end for
Output: $\boldsymbol{\theta}_{1:N}^{(q)}, \boldsymbol{x}_{1:N}^{(q)}$

The algorithm is initialized for $q = 0$ by sampling the hyperparameters $\boldsymbol{\theta}_n^{(0)}$ from the Gaussian base distribution (5.2), i.e.,

$$\boldsymbol{\theta}_n^{(0)} \sim \mathcal{N}(\boldsymbol{\theta}_n^{(0)};\boldsymbol{\mu}_{\boldsymbol{\theta}^*},\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}), \quad n = 1,\ldots,N, \tag{5.52}$$

and by calculating the parameter samples $\boldsymbol{x}_n^{(0)}$ based on the measurements $\mathbf{y}_n$. More specifically, inspired by (4.58), we set

$$\boldsymbol{x}_n^{(0)} = \boldsymbol{\mu}_{\boldsymbol{\theta}^*} + (\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*} + \boldsymbol{\Sigma}_{\boldsymbol{u}})(\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*} + \boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}(\boldsymbol{y}_n - \boldsymbol{\mu}_{\boldsymbol{\theta}^*}), \quad n = 1, \ldots, N. \tag{5.53}$$

Unfortunately, this Gibbs sampler algorithm exhibits a rather slow convergence towards the posterior distribution. This is because the algorithm does not update the cluster hyperparameters $\boldsymbol{\theta}_l^*$, which are independent of each other, but rather the individual hyperparameters $\boldsymbol{\theta}_n$. Therefore, to improve the convergence of the Gibbs sampler algorithm, we will next modify it using the cluster assignment variables $\mathbf{C}_{1:N}$ and cluster hyperparameters $\boldsymbol{\theta}^*_{\mathscr{C}(N)}$ (see (5.15)).

5.2.3 Gibbs Sampler Using Cluster Assignment Variables

The Gibbs sampling algorithm using the cluster assignment variables $\mathbf{C}_{1:N}$ was presented by MacEachern in [33] and [34] and also by Escobar and West in [35]; in the literature, it is sometimes referred to as MacEachern's algorithm. This algorithm updates samples of the cluster hyperparameters $\boldsymbol{\theta}_l^*$, as opposed to the simple Gibbs sampler presented in Section 5.2.2 that updates samples of the hyperparameter $\boldsymbol{\theta}_n$. In each iteration q of this algorithm, we will need to sample three random variables in the following order:

1. Obtain samples $C_n^{(q)}$ of the cluster assignment variable C_n for all n

   1.A. If $C_n^{(q)}$ is distinct from other previously obtained samples, i.e., $C_n^{(q)} = l_{\text{new}}$, obtain sample $\boldsymbol{\theta}_{l_{\text{new}}}^{*(q-1)}$ of the new cluster hyperparameter $\boldsymbol{\theta}^*_{l_{\text{new}}}$

2. Obtain samples $\boldsymbol{\theta}_l^{*(q)}$ of the cluster hyperparameters $\boldsymbol{\theta}_l^*$ for all l (including l_{new})

3. Obtain samples $\boldsymbol{x}_n^{(q)}$ of the parameters of interest $\mathbf{x}_n$ for all n

In what follows, we will derive the full conditional pdfs and pmfs necessary for this Gibbs sampler.

Full Conditional pmf of C_n

First, we calculate samples $C_n^{(q)}$ of the cluster assignment variables C_n for all $n = 1, \ldots, N$, using the full conditional pmf of C_n given the cluster assignment variables $\mathsf{C}_{n'} = C_{n'}$ of the

other objects ($n' \neq n$), i.e.,

$$\boldsymbol{C}_{\neg n} \triangleq (C_1, \ldots, C_{n-1}, C_{n+1}, \ldots, C_N)^{\mathrm{T}}, \tag{5.54}$$

as well as all cluster hyperparameters $\boldsymbol{\theta}^*_{\mathscr{C}(\neg n)} = \boldsymbol{\theta}^*_{\mathcal{C}(\neg n)}$ of the other objects ($n' \neq n$), where we define

$$\mathcal{C}(\neg n) = \{\boldsymbol{C}_{\neg n}\} \triangleq \{C_1, \ldots, C_{n-1}, C_{n+1}, \ldots, C_N\} \tag{5.55}$$

as the set comprising all unique elements of the vector $\boldsymbol{C}_{\neg n}$ (cf. (5.14)), as well as $\mathbf{x}_{1:N}$ and $\mathbf{y}_{1:N}$, i.e.,$\mathsf{C}_n \sim p_{\mathsf{C}_n \,|\, \mathbf{C}_{\neg n}, \boldsymbol{\theta}^*_{\mathscr{C}(\neg n)}, \mathbf{x}_{1:N}, \mathbf{y}_{1:N}}(l \,|\, \boldsymbol{C}_{\neg n}, \boldsymbol{\theta}^*_{\mathcal{C}(\neg n)}, \boldsymbol{x}_{1:N}, \boldsymbol{y}_{1:N})$.

We will consider two distinct cases:

- C_n is equal to one of the cluster assignment variables of the other objects, i.e., $C_n \in \mathcal{C}(\neg n)$.
- C_n is different from all cluster assignment variables of the other objects, i.e., $C_n \notin \mathcal{C}(\neg n)$.

For $\mathsf{C}_n = l$, the full conditional pmf is thus given by

$$p_{\mathsf{C}_n \,|\, \mathbf{C}_{\neg n}, \boldsymbol{\theta}^*_{\mathscr{C}(\neg n)}, \mathbf{x}_{1:N}, \mathbf{y}_{1:N}}(l \,|\, \boldsymbol{C}_{\neg n}, \boldsymbol{\theta}^*_{\mathcal{C}(\neg n)}, \boldsymbol{x}_{1:N}, \boldsymbol{y}_{1:N}) = \begin{cases} b_{n,l} & \text{for } \; l \in \mathcal{C}(\neg n) \\ b_{n,l_{\text{new}}} & \text{for } \; l \notin \mathcal{C}(\neg n), \end{cases} \tag{5.56}$$

with

$$b_{n,l} \triangleq \mathbb{P}\left(\mathsf{C}_n = l \,|\, \mathbf{C}_{\neg n} = \boldsymbol{C}_{\neg n}, \boldsymbol{\theta}^*_{\mathscr{C}(\neg n)} = \boldsymbol{\theta}^*_{\mathcal{C}(\neg n)}, \mathbf{x}_{1:N} = \boldsymbol{x}_{1:N}, \mathbf{y}_{1:N} = \boldsymbol{y}_{1:N}\right), \quad l \in \mathcal{C}(\neg n) \tag{5.57}$$

as defined in [21, Eq. 2.34] and

$$b_{n,l_{\text{new}}} \triangleq \mathbb{P}\left(\mathsf{C}_n = l_{\text{new}} \,|\, \mathbf{C}_{\neg n} = \boldsymbol{C}_{\neg n}, \boldsymbol{\theta}^*_{\mathscr{C}(\neg n)} = \boldsymbol{\theta}^*_{\mathcal{C}(\neg n)}, \mathbf{x}_{1:N} = \boldsymbol{x}_{1:N}, \mathbf{y}_{1:N} = \boldsymbol{y}_{1:N}\right), \quad l_{\text{new}} \notin \mathcal{C}(\neg n) \tag{5.58}$$

as defined in [21, Eq. 2.35]. We will now derive the conditional probabilities $b_{n,l}$ and $b_{n,l_{\text{new}}}$.

First, to calculate $b_{n,l}$ we assume that $l \in \mathcal{C}(\neg n)$, i.e., l is equal to one of the cluster assignment variables of the other objects. Using (5.11), we have that $\mathsf{C}_n = l$ implies $\boldsymbol{\theta}_n = \boldsymbol{\theta}^*_l$, and thus we can equivalently write (5.57) as

$$b_{n,l} = \mathbb{P}\left(\boldsymbol{\theta}_n = \boldsymbol{\theta}^*_l \,|\, \mathbf{C}_{\neg n} = \boldsymbol{C}_{\neg n}, \boldsymbol{\theta}^*_{\mathscr{C}(\neg n)} = \boldsymbol{\theta}^*_{\mathcal{C}(\neg n)}, \mathbf{x}_{1:N} = \boldsymbol{x}_{1:N}, \mathbf{y}_{1:N} = \boldsymbol{y}_{1:N}\right)$$

$$= \int_{\{\boldsymbol{\theta}_l^*\}} f_{\boldsymbol{\theta}_n \,|\, \mathbf{C}_{\neg n}, \boldsymbol{\theta}^*_{\mathscr{C}(\neg n)}, \mathbf{x}_{1:N}, \mathbf{y}_{1:N}}(\boldsymbol{\theta}_n \,|\, \boldsymbol{C}_{\neg n}, \boldsymbol{\theta}^*_{\mathcal{C}(\neg n)}, \boldsymbol{x}_{1:N}, \boldsymbol{y}_{1:N}) d\boldsymbol{\theta}_n. \tag{5.59}$$

The conditional pdf $f_{\boldsymbol{\theta}_n \,|\, \mathbf{C}_{\neg n}, \boldsymbol{\theta}^*_{\mathscr{C}(\neg n)}, \mathbf{x}_{1:N}, \mathbf{y}_{1:N}}(\boldsymbol{\theta}_n \,|\, \boldsymbol{C}_{\neg n}, \boldsymbol{\theta}^*_{\mathcal{C}(\neg n)}, \boldsymbol{x}_{1:N}, \boldsymbol{y}_{1:N})$ can be obtained by invoking (3.83), i.e., conditioning $f_{\boldsymbol{\theta}_n \,|\, \boldsymbol{\theta}_{\neg n}, \mathbf{x}_{1:N}, \mathbf{y}_{1:N}}(\boldsymbol{\theta}_n \,|\, \boldsymbol{\theta}_{\neg n}, \boldsymbol{x}_{1:N}, \boldsymbol{y}_{1:N})$ on $\mathbf{C}_{\neg n}$ and $\boldsymbol{\theta}^*_{\mathscr{C}(\neg n)}$ instead of $\boldsymbol{\theta}_{\neg n}$. We then obtain from (5.31)

$$\begin{aligned} &f_{\boldsymbol{\theta}_n \,|\, \mathbf{C}_{\neg n}, \boldsymbol{\theta}^*_{\mathscr{C}(\neg n)}, \mathbf{x}_{1:N}, \mathbf{y}_{1:N}}(\boldsymbol{\theta}_n \,|\, \boldsymbol{C}_{\neg n}, \boldsymbol{\theta}^*_{\mathcal{C}(\neg n)}, \boldsymbol{x}_{1:N}, \boldsymbol{y}_{1:N}) \\ &\propto \alpha f_{\mathbf{x}_n \,|\, \boldsymbol{\theta}_n}(\boldsymbol{x}_n \,|\, \boldsymbol{\theta}_n) f_{\mathrm{H}}(\boldsymbol{\theta}_n) + \sum_{n' \in \{1,\ldots,N\} \setminus \{n\}} f_{\mathbf{x}_n \,|\, \boldsymbol{\theta}_n}(\boldsymbol{x}_n \,|\, \boldsymbol{\theta}^*_{C_{n'}}) \delta_{\boldsymbol{\theta}^*_{C_{n'}}}(\boldsymbol{\theta}_n). \end{aligned} \tag{5.60}$$

Furthermore, we define analogously to (3.47)

$$m_l(\neg n) \triangleq \sum_{n'=1}^{n-1} \mathbb{1}(C_{n'} = l) + \sum_{n'=n+1}^{N} \mathbb{1}(C_{n'} = l), \quad l \in \mathcal{C}(\neg n), \tag{5.61}$$

which denotes the number of cluster assignment variables $C_{n'}$ of the other objects, $n' \neq n$, $\mathcal{C}(\neg n)$ that are equal to l. Using (5.61) in (5.60), we obtain

$$\begin{aligned} &f_{\boldsymbol{\theta}_n \,|\, \mathbf{C}_{\neg n}, \boldsymbol{\theta}^*_{\mathscr{C}(\neg n)}, \mathbf{x}_{1:N}, \mathbf{y}_{1:N}}(\boldsymbol{\theta}_n \,|\, \boldsymbol{C}_{\neg n}, \boldsymbol{\theta}^*_{\mathcal{C}(\neg n)}, \boldsymbol{x}_{1:N}, \boldsymbol{y}_{1:N}) \\ &\propto \alpha f_{\mathbf{x}_n \,|\, \boldsymbol{\theta}_n}(\boldsymbol{x}_n \,|\, \boldsymbol{\theta}_n) f_{\mathrm{H}}(\boldsymbol{\theta}_n) + \sum_{l \in \mathcal{C}(\neg n)} m_l(\neg n) f_{\mathbf{x}_n \,|\, \boldsymbol{\theta}_n}(\boldsymbol{x}_n \,|\, \boldsymbol{\theta}^*_l) \delta_{\boldsymbol{\theta}^*_l}(\boldsymbol{\theta}_n). \end{aligned} \tag{5.62}$$

Finally, inserting (5.62) into (5.59) gives

$$\begin{aligned} b_{n,l} &\propto \int_{\{\boldsymbol{\theta}_l^*\}} \alpha f_{\mathbf{x}_n \,|\, \boldsymbol{\theta}_n}(\boldsymbol{x}_n \,|\, \boldsymbol{\theta}_n) f_{\mathrm{H}}(\boldsymbol{\theta}_n) + \sum_{l' \in \mathcal{C}(\neg n)} m_{l'}(\neg n) f_{\mathbf{x}_n \,|\, \boldsymbol{\theta}_n}(\boldsymbol{x}_n \,|\, \boldsymbol{\theta}^*_{l'}) \delta_{\boldsymbol{\theta}^*_{l'}}(\boldsymbol{\theta}_n) d\boldsymbol{\theta}_n \\ &= \alpha \int_{\{\boldsymbol{\theta}_l^*\}} f_{\mathbf{x}_n \,|\, \boldsymbol{\theta}_n}(\boldsymbol{x}_n \,|\, \boldsymbol{\theta}_n) f_{\mathrm{H}}(\boldsymbol{\theta}_n) d\boldsymbol{\theta}_n + \sum_{l' \in \mathcal{C}(\neg n)} m_{l'}(\neg n) f_{\mathbf{x}_n \,|\, \boldsymbol{\theta}_n}(\boldsymbol{x}_n \,|\, \boldsymbol{\theta}^*_{l'}) \int_{\{\boldsymbol{\theta}_l^*\}} \delta_{\boldsymbol{\theta}^*_{l'}}(\boldsymbol{\theta}_n) d\boldsymbol{\theta}_n \end{aligned} \tag{5.63}$$

Assuming that $f_{\mathbf{x}_n \,|\, \boldsymbol{\theta}_n}(\boldsymbol{x}_n \,|\, \boldsymbol{\theta}_n)$ and $f_{\mathrm{H}}(\boldsymbol{\theta}_n)$ do not contain any discrete (Dirac) components, the first integral in (5.63) is equal to 0. Furthermore, using our earlier assumption that $l \in \mathcal{C}(\neg n)$, the second integral in (5.63) is 1 for $l' = l$ and 0 for $l' \neq l$, i.e., the sum becomes

$$\sum_{l' \in \mathcal{C}(\neg n)} m_{l'}(\neg n) f_{\mathbf{x}_n \,|\, \boldsymbol{\theta}_n}(\boldsymbol{x}_n \,|\, \boldsymbol{\theta}^*_{l'}) \delta_{l',l} = m_l(\neg n) f_{\mathbf{x}_n \,|\, \boldsymbol{\theta}_n}(\boldsymbol{x}_n \,|\, \boldsymbol{\theta}^*_l) \tag{5.64}$$

Thus the expression (5.63) becomes

$$b_{n,l} \propto m_l(\neg n) f_{\mathbf{x}_n \mid \boldsymbol{\theta}_n}(\boldsymbol{x}_n \mid \boldsymbol{\theta}_l^*). \tag{5.65}$$

Finally, using (5.8) and (5.61), we obtain

$$b_{n,l} \propto \sum_{n'=1}^{n-1} \mathbb{1}(C_{n'} = l) + \sum_{n'=n+1}^{N} \mathbb{1}(C_{n'} = l)\mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\theta}_l^*, \boldsymbol{\Sigma}_{\boldsymbol{u}}), \quad l \in \mathcal{C}(\neg n). \tag{5.66}$$

We now consider the second case, $l_{\text{new}} \notin \mathcal{C}(\neg n)$, and provide an expression for $b_{n,l_{\text{new}}}$ in (5.58). In this case, C_n is different from all the other cluster assignment variables, which implies that a new cluster l_{new} is created. According to (5.11), $\mathsf{C}_n = l_{\text{new}}$ implies $\boldsymbol{\Theta}_n = \boldsymbol{\theta}_{l_{\text{new}}}^*$, therefore the new cluster hyperparameter $\boldsymbol{\theta}_{l_{\text{new}}}^*$ is not equal to any other cluster hyperparameters $\boldsymbol{\theta}_l^*$ with $l \in \mathcal{C}(\neg n)$ i.e., it needs to be sampled from the subset $\mathbb{R}^D \backslash \{\boldsymbol{\theta}_{\mathcal{C}(\neg n)}^*\}_D$. Here, $\{\boldsymbol{\theta}_{\mathcal{C}(\neg n)}^*\}_D$ is the set[4] that contains all cluster hyperparameters $\boldsymbol{\Theta}_l^* \in \mathbb{R}^D$ for $l \in \mathcal{C}(\neg n)$. Similar to (5.59), we can rewrite (5.58) equivalently as

$$\begin{aligned} b_{n,l_{\text{new}}} &= \mathbb{P}\left(\boldsymbol{\Theta}_n = \boldsymbol{\theta}_{l_{\text{new}}}^* \mid \mathbf{C}_{\neg n} = \boldsymbol{C}_{\neg n}, \boldsymbol{\Theta}_{\mathscr{C}(\neg n)}^* = \boldsymbol{\theta}_{\mathcal{C}(\neg n)}^*, \mathbf{x}_{1:N} = \boldsymbol{x}_{1:N}, \mathbf{y}_{1:N} = \boldsymbol{y}_{1:N}\right) \\ &= \int_{\mathbb{R}^D \backslash \{\boldsymbol{\theta}_{\mathcal{C}(\neg n)}^*\}_D} f_{\boldsymbol{\Theta}_n \mid \mathbf{C}_{\neg n}, \boldsymbol{\Theta}_{\mathscr{C}(\neg n)}^*, \mathbf{x}_{1:N}, \mathbf{y}_{1:N}}(\boldsymbol{\theta}_n \mid \boldsymbol{C}_{\neg n}, \boldsymbol{\theta}_{\mathcal{C}(\neg n)}^*, \boldsymbol{x}_{1:N}, \boldsymbol{y}_{1:N}) d\boldsymbol{\theta}_n. \end{aligned} \tag{5.67}$$

Using (5.62) in (5.67), we obtain further

$$\begin{aligned} b_{n,l_{\text{new}}} &\propto \int_{\mathbb{R}^D \backslash \{\boldsymbol{\theta}_{\mathcal{C}(\neg n)}^*\}_D} \alpha f_{\mathbf{x}_n \mid \boldsymbol{\theta}_n}(\boldsymbol{x}_n \mid \boldsymbol{\theta}_n) f_{\mathrm{H}}(\boldsymbol{\theta}_n) + \sum_{l \in \mathcal{C}(\neg n)} m_l(\neg n) f_{\mathbf{x}_n \mid \boldsymbol{\theta}_n}(\boldsymbol{x}_n \mid \boldsymbol{\theta}_l^*) \delta_{\boldsymbol{\theta}_l^*}(\boldsymbol{\theta}_n) d\boldsymbol{\theta}_n \\ &= \int_{\mathbb{R}^D \backslash \{\boldsymbol{\theta}_{\mathcal{C}(\neg n)}^*\}_D} \alpha f_{\mathbf{x}_n \mid \boldsymbol{\theta}_n}(\boldsymbol{x}_n \mid \boldsymbol{\theta}_n) f_{\mathrm{H}}(\boldsymbol{\theta}_n) d\boldsymbol{\theta}_n \qquad (5.68) \\ &\quad + \sum_{l \in \mathcal{C}(\neg n)} m_l(\neg n) \int_{\mathbb{R}^D \backslash \{\boldsymbol{\theta}_{\mathcal{C}(\neg n)}^*\}_D} f_{\mathbf{x}_n \mid \boldsymbol{\theta}_n}(\boldsymbol{x}_n \mid \boldsymbol{\theta}_l^*) \delta_{\boldsymbol{\theta}_l^*}(\boldsymbol{\theta}_n) d\boldsymbol{\theta}_n \qquad (5.69) \end{aligned}$$

In the second integral (5.69), the points excluded in the integration domain $\mathbb{R}^D \backslash \{\boldsymbol{\theta}_{\mathcal{C}(\neg n)}^*\}_D$ are exactly the locations of the Dirac components $\delta_{\boldsymbol{\theta}_l^*}(\boldsymbol{\theta}_n)$ in the integrand, and therefore the second integral (5.69) is equal to 0. Furthermore, we assume that $f_{\mathbf{x}_n \mid \boldsymbol{\theta}_n}(\boldsymbol{x}_n \mid \boldsymbol{\theta}_n)$ and $f_{\mathrm{H}}(\boldsymbol{\theta}_n)$ do not contain any discrete (Dirac) components, we can replace the integration domain of

[4] Here we use $\{\cdot\}_D$ to note that the set contains vectors of size D and not the individual components as used in Chapter 3.

the integral (5.68) by $\mathbb{R}^D$. Thus, we obtain

$$b_{n,l_{\text{new}}} \propto \alpha \int_{\mathbb{R}^D} f_{\mathbf{x}_n \,|\, \boldsymbol{\theta}_n}(\boldsymbol{x}_n \,|\, \boldsymbol{\theta}_n) f_{\text{H}}(\boldsymbol{\theta}_n) d\boldsymbol{\theta}_n. \tag{5.70}$$

Recalling (5.8) and (5.2), this becomes

$$\begin{aligned} b_{n,l_{\text{new}}} &\propto \alpha \int_{\mathbb{R}^D} \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\theta}_n, \boldsymbol{\Sigma}_u) \mathcal{N}(\boldsymbol{\theta}_n; \boldsymbol{\mu}_{\boldsymbol{\theta}^*}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}) d\boldsymbol{\theta}_n \\ &= \alpha \int_{\mathbb{R}^D} \mathcal{N}(\boldsymbol{x}_n - \boldsymbol{\theta}_n; \mathbf{0}, \boldsymbol{\Sigma}_u) \mathcal{N}(\boldsymbol{\theta}_n; \boldsymbol{\mu}_{\boldsymbol{\theta}^*}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}) d\boldsymbol{\theta}_n \\ &= \alpha \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}_{\boldsymbol{\theta}^*}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*} + \boldsymbol{\Sigma}_u), \end{aligned} \tag{5.71}$$

where we used the identity (2.19). Lastly, we need to normalize $b_{n,l}$ for all $l \in \mathcal{C}(\neg n)$ and $b_{n,l_{\text{new}}}$ so that

$$\sum_{l' \in \mathcal{C}(\neg n) \cup \{l_{\text{new}}\}} b_{n,l'} = 1. \tag{5.72}$$

Finally, to obtain samples $C_n^{(q)}$ of the cluster assignment variable, we evaluate the full conditional pmf $p_{\mathsf{C}_n \,|\, \mathsf{C}_{\neg n}, \boldsymbol{\theta}^*_{\mathcal{C}(\neg n)}, \mathbf{x}_{1:N}, \mathbf{y}_{1:N}}(l \,|\, \boldsymbol{C}_{\neg n}, \boldsymbol{\theta}^*_{\mathcal{C}(\neg n)}, \boldsymbol{x}_{1:N}, \boldsymbol{y}_{1:N})$ at the samples already available in the q^{th} iteration of the Gibbs sampler algorithm. Similarly to (5.21) and (5.22), we define a vector of already available cluster assignment variables samples of other objects as (cf. (5.54))

$$\boldsymbol{C}_{\neg n}^{(q,q-1)} \triangleq \left(C_1^{(q)}, \ldots, C_{n-1}^{(q)}, C_{n+1}^{(q-1)}, \ldots, C_N^{(q-1)}\right)^{\text{T}} \tag{5.73}$$

and a corresponding set, comprising all unique elements of the vector $\boldsymbol{C}^{(q,q-1)}(\neg n)$, as (cf. (5.55))

$$\mathcal{C}^{(q,q-1)}(\neg n) = \{\boldsymbol{C}_{\neg n}^{(q,q-1)}\} \triangleq \{C_1^{(q)}, \ldots, C_{n-1}^{(q)}, C_{n+1}^{(q-1)}, \ldots, C_N^{(q-1)}\}. \tag{5.74}$$

At each iteration q, the cluster assignment variables C_n are sampled first, therefore we will use the cluster hyperparameter samples $\boldsymbol{\theta}_l^{*(q-1)}$ from the previous iteration; however, we must use the already available cluster assignment variables $\boldsymbol{C}_{\neg n}^{(q,q-1)}$. Similar to (5.15), we define

$$\boldsymbol{\theta}^{*(q-1)}_{\mathcal{C}^{(q,q-1)}(\neg n)} \triangleq (\boldsymbol{\theta}_l^{*(q-1)})_{l \in \mathcal{C}^{(q,q-1)}(\neg n)}, \tag{5.75}$$

as well as the vector of the samples of all the parameters of interest from the previous iteration, i.e.,

$$\boldsymbol{x}_{1:N}^{(q-1)} \triangleq \left(\boldsymbol{x}_1^{(q-1)\text{T}}, \ldots, \boldsymbol{x}_N^{(q-1)\text{T}}\right)^{\text{T}}. \tag{5.76}$$

By evaluating (5.56) at the available samples given in (5.73-5.76) we obtain

$$p_{\mathsf{C}_n \,|\, \mathsf{C}_{\neg n}, \boldsymbol{\theta}^*_{\mathscr{C}(\neg n)}, \mathbf{x}_{1:N}, \mathbf{y}_{1:N}}(l \,|\, \boldsymbol{C}^{(q,q-1)}_{\neg n}, \boldsymbol{\theta}^{*(q-1)}_{\mathcal{C}^{(q,q-1)}(\neg n)}, \boldsymbol{x}^{(q-1)}_{1:N}, \boldsymbol{y}_{1:N}) = \begin{cases} b^{(q)}_{n,l} & \text{for } \; l \in \mathcal{C}^{(q,q-1)}(\neg n) \\ b^{(q)}_{n,l_{\text{new}}} & \text{for } \; l \notin \mathcal{C}^{(q,q-1)}(\neg n), \end{cases} \tag{5.77}$$

with (see (5.66))

$$b^{(q)}_{n,l} \propto \left(\sum_{n'=1}^{n-1} \mathbb{1}(C^{(q)}_{n'} = l) + \sum_{n'=n+1}^{N} \mathbb{1}(C^{(q-1)}_{n'} = l) \right) \mathcal{N}(\boldsymbol{x}^{(q-1)}_n; \boldsymbol{\theta}^{*(q-1)}_l, \boldsymbol{\Sigma}_{\boldsymbol{u}}), \quad l \in \mathcal{C}^{(q,q-1)}(\neg n), \tag{5.78}$$

and (see (5.71))

$$b^{(q)}_{n,l_{\text{new}}} \propto \alpha \mathcal{N}(\boldsymbol{x}^{(q-1)}_n; \boldsymbol{\mu}_{\boldsymbol{\theta}^*}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*} + \boldsymbol{\Sigma}_u). \tag{5.79}$$

For later use, we define

$$\mathcal{C}^{(q)}(N) = \{\boldsymbol{C}^{(q)}_{1:N}\} \triangleq \{C^{(q)}_1, \ldots, C^{(q)}_N\} \tag{5.80}$$

as the set of all unique cluster assignment variables samples $C^{(q)}_n$ obtained in the q^{th} iteration of the algorithm.

Lastly, if $C^{(q)}_n = l_{\text{new}}$ was obtained, we also need to sample a new cluster hyperparameter $\boldsymbol{\theta}^{*(q-1)}_{l_{\text{new}}}$. We note that if $C^{(q)}_n = l_{\text{new}}$, we set

$$l_{\text{new}} = \max\{\max_{n' \in \{1,\ldots,n-1\}} C^{(q)}_{n'}, l^{(q-1)}_{\max}\} + 1, \tag{5.81}$$

and we include l_{new} in $\mathcal{C}^{(q)}(N)$, i.e., $l_{\text{new}} \in \mathcal{C}^{(q)}(N)$.

Full Conditional pdf of Θ^*_l

The samples $\boldsymbol{\theta}^{*(q)}_l$ of the cluster hyperparameters Θ^*_l are obtained from the full conditional pdf of Θ^*_l given the other cluster hyperparameters $\Theta^*_{\mathscr{C}(N)\setminus\{l\}}$, the cluster assignment variables $\mathbf{C}_{1:N}$, the parameters of interest $\mathbf{x}_{1:N}$, and the measurements $\mathbf{y}_{1:N}$, i.e., $f_{\Theta^*_l \,|\, \Theta^*_{\mathcal{C}(N)\setminus\{l\}}, \mathbf{C}_{1:N}, \mathbf{x}_{1:N}, \mathbf{y}_{1:N}}(\boldsymbol{\theta}^*_l \,|\, \boldsymbol{\theta}^*_{\mathcal{C}(N)\setminus\{l\}}, \boldsymbol{C}_{1:N}, \boldsymbol{x}_{1:N}, \boldsymbol{y}_{1:N})$.

We first derive the full conditional pdf and then evaluate it at the samples available at

the q^{th} iteration. Using Bayes' theorem, we obtain

$$\begin{aligned}
&f_{\theta_l^* \mid \theta^*_{\mathscr{C}(N)\setminus\{l\}}, \mathbf{C}_{1:N}, \mathbf{x}_{1:N}, \mathbf{y}_{1:N}}(\boldsymbol{\theta}_l^* \mid \boldsymbol{\theta}^*_{\mathcal{C}(N)\setminus\{l\}}, \boldsymbol{C}_{1:N}, \boldsymbol{x}_{1:N}, \boldsymbol{y}_{1:N}) \\
&\propto f_{\mathbf{x}_{1:N}, \mathbf{y}_{1:N} \mid \theta_l^*, \theta^*_{\mathscr{C}(N)\setminus\{l\}}, \mathbf{C}_{1:N}}(\boldsymbol{x}_{1:N}, \boldsymbol{y}_{1:N} \mid \boldsymbol{\theta}_l^*, \boldsymbol{\theta}^*_{\mathcal{C}(N)\setminus\{l\}}, \boldsymbol{C}_{1:N}) f_{\theta_l^* \mid \theta^*_{\mathscr{C}(N)\setminus\{l\}}, \mathbf{C}_{1:N}}(\boldsymbol{\theta}_l^* \mid \boldsymbol{\theta}^*_{\mathcal{C}(N)\setminus\{l\}}, \boldsymbol{C}_{1:N})
\end{aligned} \tag{5.82}$$

We recall from Section 5.1 that we can express the hyperparameter θ_n equivalently using the cluster assignment variable C_n and the corresponding cluster hyperparameter θ_l^* as $\theta_n = \theta^*_{\mathsf{C}_n}$ (see (3.40)). Accordingly, we can express the hyperparameter vector $\theta_{1:N}$ as

$$\theta_{1:N} = (\theta^{*\mathrm{T}}_{\mathsf{C}_1}, \dots, \theta^{*\mathrm{T}}_{\mathsf{C}_N})^{\mathrm{T}}. \tag{5.83}$$

Restricting to the set of distinct random variables $\theta^*_{\mathsf{C}_n}$, this is equivalent to the random vector $\theta^*_{\mathscr{C}(N)} = \big(\theta_l^*\big)_{l \in \mathscr{C}(N)}$. Therefore, we obtain for (5.82)

$$\begin{aligned}
&f_{\theta_l^* \mid \theta^*_{\mathscr{C}(N)\setminus\{l\}}, \mathbf{C}_{1:N}, \mathbf{x}_{1:N}, \mathbf{y}_{1:N}}(\boldsymbol{\theta}_l^* \mid \boldsymbol{\theta}^*_{\mathcal{C}(N)\setminus\{l\}}, \boldsymbol{C}_{1:N}, \boldsymbol{x}_{1:N}, \boldsymbol{y}_{1:N}) \\
&\propto f_{\mathbf{x}_{1:N}, \mathbf{y}_{1:N} \mid \theta^*_{\mathscr{C}(N)}, \mathbf{C}_{1:N}}(\boldsymbol{x}_{1:N}, \boldsymbol{y}_{1:N} \mid \boldsymbol{\theta}^*_{\mathcal{C}(N)}, \boldsymbol{C}_{1:N}) f_{\theta_l^* \mid \theta^*_{\mathscr{C}(N)\setminus\{l\}}, \mathbf{C}_{1:N}}(\boldsymbol{\theta}_l^* \mid \boldsymbol{\theta}^*_{\mathcal{C}(N)\setminus\{l\}}, \boldsymbol{C}_{1:N}).
\end{aligned} \tag{5.84}$$

The first factor in (5.84) can be factorized as

$$\begin{aligned}
&f_{\mathbf{x}_{1:N}, \mathbf{y}_{1:N} \mid \theta^*_{\mathscr{C}(N)}, \mathbf{C}_{1:N}}(\boldsymbol{x}_{1:N}, \boldsymbol{y}_{1:N} \mid \boldsymbol{\theta}^*_{\mathcal{C}(N)}, \boldsymbol{C}_{1:N}) \\
&= f_{\mathbf{y}_{1:N} \mid \mathbf{x}_{1:N}, \theta^*_{\mathscr{C}(N)}, \mathbf{C}_{1:N}}(\boldsymbol{y}_{1:N} \mid \boldsymbol{x}_{1:N}, \boldsymbol{\theta}^*_{\mathcal{C}(N)}, \boldsymbol{C}_{1:N}) f_{\mathbf{x}_{1:N} \mid \theta^*_{\mathscr{C}(N)}, \mathbf{C}_{1:N}}(\boldsymbol{x}_{1:N} \mid \boldsymbol{\theta}^*_{\mathcal{C}(N)}, \boldsymbol{C}_{1:N}).
\end{aligned} \tag{5.85}$$

By invoking (5.83), i.e., conditioning on $\theta_{1:N}$ instead of $\mathbf{C}_{1:N}$ and $\theta^*_{\mathscr{C}(N)}$ and in turn using (4.17) and (3.86), (5.85) can be simplified to

$$\begin{aligned}
&f_{\mathbf{x}_{1:N}, \mathbf{y}_{1:N} \mid \theta^*_{\mathcal{C}(N)}, \mathbf{C}_{1:N}}(\boldsymbol{x}_{1:N}, \boldsymbol{y}_{1:N} \mid \boldsymbol{\theta}^*_{\mathcal{C}(N)}, \boldsymbol{C}_{1:N}) \\
&\overset{(5.83)}{=} f_{\mathbf{y}_{1:N} \mid \mathbf{x}_{1:N}, \theta_{1:N}}(\boldsymbol{y}_{1:N} \mid \boldsymbol{x}_{1:N}, \boldsymbol{\theta}_{1:N}) f_{\mathbf{x}_{1:N} \mid \theta_{1:N}}(\boldsymbol{x}_{1:N} \mid \boldsymbol{\theta}_{1:N}) \\
&\overset{(4.17)}{=} \Big(\prod_{n=1}^{N} f_{\mathbf{y}_n \mid \mathbf{x}_n}(\boldsymbol{y}_n \mid \boldsymbol{x}_n) \Big) f_{\mathbf{x}_{1:N} \mid \theta_{1:N}}(\boldsymbol{x}_{1:N} \mid \boldsymbol{\theta}_{1:N}) \\
&\overset{(3.86)}{=} \Big(\prod_{n=1}^{N} f_{\mathbf{y}_n \mid \mathbf{x}_n}(\boldsymbol{y}_n \mid \boldsymbol{x}_n) \Big) \prod_{l \in \mathcal{C}(N)} \prod_{n' : C_n = l} f_{\mathbf{x}_{n'} \mid \theta_l^*, \mathsf{C}_{n'}}(\boldsymbol{x}_{n'} \mid \boldsymbol{\theta}_l^*, C_{n'})
\end{aligned} \tag{5.86}$$

We now take up the second factor in (5.84). We recall that the cluster hyperparame-

ters θ_l^* are i.i.d. with pdf $f_{\theta_l^*}(\boldsymbol{\theta}_l^*) = f_{\mathrm{H}}(\boldsymbol{\theta}_l^*)$ (see (5.3)). Using (3.55), the conditional pdf $f_{\theta_l^* \,|\, \theta^*_{\mathscr{C}(N)\setminus\{l\}}, \mathsf{C}_{1:N}}(\boldsymbol{\theta}_l^* \,|\, \boldsymbol{\theta}^*_{\mathcal{C}(N)\setminus\{l\}}, \boldsymbol{C}_{1:N})$ becomes

$$f_{\theta_l^* \,|\, \theta^*_{\mathscr{C}(N)\setminus\{l\}}, \mathsf{C}_{1:N}}(\boldsymbol{\theta}_l^* \,|\, \boldsymbol{\theta}^*_{\mathcal{C}(N)\setminus\{l\}}, \boldsymbol{C}_{1:N}) = f_{\mathrm{H}}(\boldsymbol{\theta}_l^*). \tag{5.87}$$

Inserting (5.86) and (5.87) into (5.82) yields

$$\begin{aligned} &f_{\theta_l^* \,|\, \theta^*_{\mathscr{C}(N)\setminus\{l\}}, \mathsf{C}_{1:N}, \mathsf{x}_{1:N}, \mathsf{y}_{1:N}}(\boldsymbol{\theta}_l^* \,|\, \boldsymbol{\theta}^*_{\mathcal{C}(N)\setminus\{l\}}, \boldsymbol{C}_{1:N}, \boldsymbol{x}_{1:N}, \boldsymbol{y}_{1:N}) \\ &\propto \left(\prod_{n=1}^{N} f_{\mathsf{y}_n \,|\, \mathsf{x}_n}(\boldsymbol{y}_n \,|\, \boldsymbol{x}_n) \right) \left(\prod_{l' \in \mathcal{C}(N)} \prod_{n' : C_n = l'} f_{\mathsf{x}_{n'} \,|\, \theta^*_{l'}, \mathsf{C}_{n'}}(\boldsymbol{x}_{n'} \,|\, \boldsymbol{\theta}^*_{l'}, C_{n'}) \right) f_{\mathrm{H}}(\boldsymbol{\theta}_l^*). \end{aligned} \tag{5.88}$$

Lastly, all factors in (5.88) that do not functionally depend on θ_l^* are considered as constant; therefore, we obtain the final expression for the full conditional pdf of θ_l^* as

$$\begin{aligned} &f_{\theta_l^* \,|\, \theta^*_{\mathcal{C}(N)\setminus\{l\}}, \mathsf{C}_{1:N}, \mathsf{x}_{1:N}, \mathsf{y}_{1:N}}(\boldsymbol{\theta}_l^* \,|\, \boldsymbol{\theta}^*_{\mathcal{C}(N)\setminus\{l\}}, \boldsymbol{C}_{1:N}, \boldsymbol{x}_{1:N}, \boldsymbol{y}_{1:N}) \\ &\propto \left(\prod_{n : C_n = l} f_{\mathsf{x}_n \,|\, \theta_l^*, \mathsf{C}_n}(\boldsymbol{x}_n \,|\, \boldsymbol{\theta}_l^*, C_n) \right) f_{\mathrm{H}}(\boldsymbol{\theta}_l^*) \\ &= \left(\prod_{n : C_n = l} f_{\mathsf{x}_n \,|\, \theta^*_{C_n}, \mathsf{C}_n}(\boldsymbol{x}_n \,|\, \boldsymbol{\theta}^*_{C_n}, C_n) \right) f_{\mathrm{H}}(\boldsymbol{\theta}_l^*) \\ &= \left(\prod_{n : C_n = l} f_{\mathsf{x}_n \,|\, \theta_n}(\boldsymbol{x}_n \,|\, \boldsymbol{\theta}_l^*) \right) f_{\mathrm{H}}(\boldsymbol{\theta}_l^*) \end{aligned} \tag{5.89}$$

where (3.83) was used. Finally, using (5.2) and (4.3), we obtain a product of Gaussian pdfs:

$$\begin{aligned} &f_{\theta_l^* \,|\, \theta^*_{\mathscr{C}(N)\setminus\{l\}}, \mathsf{C}_{1:N}, \mathsf{x}_{1:N}, \mathsf{y}_{1:N}}(\boldsymbol{\theta}_l^* \,|\, \boldsymbol{\theta}^*_{\mathcal{C}(N)\setminus\{l\}}, \boldsymbol{C}_{1:N}, \boldsymbol{x}_{1:N}, \boldsymbol{y}_{1:N}) \\ &\propto \left(\prod_{n : C_n = l} \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\theta}_l^*, \boldsymbol{\Sigma}_{\boldsymbol{u}}) \right) \mathcal{N}(\boldsymbol{\theta}_l^*; \boldsymbol{\mu}_{\boldsymbol{\theta}^*}; \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}). \end{aligned} \tag{5.90}$$

Finally, expression (5.90) is a product of Gaussians, hence using the identity (2.59) with the following substitutions $\boldsymbol{x} \to \boldsymbol{\theta}_l^*$, $\boldsymbol{\mu}_{\boldsymbol{x}} \to \boldsymbol{\mu}_{\boldsymbol{\theta}^*}$, $\boldsymbol{\Sigma}_{\boldsymbol{x}} \to \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}$, $\boldsymbol{y}_n \to \boldsymbol{x}_n$ and $\boldsymbol{\Sigma}_{\boldsymbol{y}} \to \boldsymbol{\Sigma}_{\boldsymbol{u}}$, the expression (5.90) is also Gaussian, i.e.,

$$f_{\theta_l^* \,|\, \theta^*_{\mathscr{C}(N)\setminus\{l\}}, \mathsf{C}_{1:N}, \mathsf{x}_{1:N}, \mathsf{y}_{1:N}}(\boldsymbol{\theta}_l^* \,|\, \boldsymbol{\theta}^*_{\mathcal{C}(N)\setminus\{l\}}, \boldsymbol{C}_{1:N}, \boldsymbol{x}_{1:N}, \boldsymbol{y}_{1:N}) \propto \tilde{\gamma} \mathcal{N}(\boldsymbol{\theta}_l^*; \boldsymbol{\mu}_{\boldsymbol{\theta}_l^* \,|\, \boldsymbol{x}_{1:N}, \boldsymbol{C}_{1:N}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}_l^* \,|\, \boldsymbol{x}_{1:N}, \boldsymbol{C}_{1:N}}). \tag{5.91}$$

Here, the covariance matrix is given by (see (2.53))

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}_l^* \,|\, \boldsymbol{x}_{1:N}, \boldsymbol{C}_{1:N}} = \boldsymbol{\Sigma}_{\boldsymbol{u}} \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + m_l(N) \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*} \right)^{-1} \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*} = \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*} \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + m_l(N) \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*} \right)^{-1} \boldsymbol{\Sigma}_{\boldsymbol{u}}, \tag{5.92}$$

where

$$m_l(N) = \sum_{n=1}^{N} \mathbb{1}(C_n = l) \tag{5.93}$$

denotes the number of objects n for which $C_n^{(q)} = l$. Furthermore, the mean $\boldsymbol{\mu}_{\boldsymbol{\theta}_l^* \,|\, \boldsymbol{x}_{1:N}, \boldsymbol{C}_{1:N}}$ is given by (see (2.56))

$$\boldsymbol{\mu}_{\boldsymbol{\theta}_l^* \,|\, \boldsymbol{x}_{1:N}, \boldsymbol{C}_{1:N}} = \boldsymbol{\Sigma}_{\boldsymbol{u}} \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + m_l(N) \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*} \right)^{-1} \boldsymbol{\mu}_{\boldsymbol{\theta}^*} + m_l(N) \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*} \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + m_l(N) \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*} \right)^{-1} \bar{\boldsymbol{x}}_l \tag{5.94}$$

where $\bar{\boldsymbol{x}}_l$ denotes the sample mean

$$\bar{\boldsymbol{x}}_l = \frac{1}{m_l(N)} \sum_{n: C_n = l} \boldsymbol{x}_n \tag{5.95}$$

of all $\boldsymbol{x}_n$ belonging to cluster l. Here we note that both the mean and the covariance matrix depend on the cluster index l.

We note that similar to (5.31), expression (5.91) does not functionally depend on the measurements $\boldsymbol{y}_{1:N}$. Furthermore, (5.91) does not functionally depend on the other cluster hyperparameters $\boldsymbol{\theta}^*_{\mathcal{C}(N)\backslash\{l\}}$. We also note that only parameters of interest $\boldsymbol{x}_n$ that belong to the same cluster $(\boldsymbol{x}_n)_{n: C_n^{(q)} = l}$ are used.

In order to obtain the cluster hyperparameters samples $\boldsymbol{\theta}_l^{*(q)}$, we have to evaluate the full conditional pdf $f_{\boldsymbol{\theta}_l^* \,|\, \boldsymbol{\theta}^*_{\mathcal{C}(N)\backslash\{l\}}, \boldsymbol{C}_{1:N}, \boldsymbol{x}_{1:N}, \boldsymbol{y}_{1:N}}(\boldsymbol{\theta}_l^* \,|\, \boldsymbol{\theta}^*_{\mathcal{C}(N)\backslash\{l\}}, \boldsymbol{C}_{1:N}, \boldsymbol{x}_{1:N}, \boldsymbol{y}_{1:N})$ at the samples already available at q^{th} iteration, i.e., the samples of the cluster assignment variables $\boldsymbol{C}_{1:N}^{(q)}$, the samples of the parameters of interest $\boldsymbol{x}_{1:N}^{(q-1)}$ (see (5.76)), and the vector of already available samples $\boldsymbol{\theta}^{*(q,q-1)}_{\mathcal{C}^{(q)}(N)\backslash\{l\}}$, whose elements $\boldsymbol{\theta}^{*(q,q-1)}_{l,l'}$, $l' \in \mathcal{C}^{(q)}(N)\backslash\{l\}$ are defined as

$$\boldsymbol{\theta}^{*(q,q-1)}_{l,l'} \triangleq \begin{cases} \boldsymbol{\theta}^{*(q)}_{l,l'} & \text{for } \; l' \in \mathcal{C}^{(q)}(N)\backslash\{l\} \;\; \text{already sampled at the } q^{\text{th}} \text{ iteration} \\ \boldsymbol{\theta}^{*(q-1)}_{l,l'} & \text{for } \; l' \in \mathcal{C}^{(q)}(N)\backslash\{l\} \;\; \text{not yet sampled at the } q^{\text{th}} \text{ iteration.} \end{cases} \tag{5.96}$$

Here, we refer to the values of the cluster hyperparameters $\boldsymbol{\theta}_{l'}^{*(q)}$ that have been sampled in the q^{th} as all the cluster assignment variables $C_n^{(q)}$ have already been sampled previously.

Using (5.91), we obtain

$$f_{\theta_l^* \mid \theta^*_{\mathscr{C}(N)\setminus\{l\}}, \mathbf{C}_{1:N}, \mathbf{x}_{1:N}, \mathbf{y}_{1:N}}(\boldsymbol{\theta}_l^{*(q)} \mid \boldsymbol{\theta}^{*(q,q-1)}_{\mathcal{C}^{(q)}(N)\setminus\{l\}}, \boldsymbol{C}^{(q)}_{1:N}, \boldsymbol{x}^{(q-1)}_{1:N}, \boldsymbol{y}_{1:N}) \propto \mathcal{N}(\boldsymbol{\theta}_l^{*(q)}; \boldsymbol{\mu}^{(q)}_{\boldsymbol{\theta}_l^* \mid \boldsymbol{x}_{1:N}, \boldsymbol{C}_{1:N}}, \boldsymbol{\Sigma}^{(q)}_{\boldsymbol{\theta}_l^* \mid \boldsymbol{x}_{1:N}, \boldsymbol{C}_{1:N}}). \tag{5.97}$$

with

$$\boldsymbol{\Sigma}^{(q)}_{\boldsymbol{\theta}_l^* \mid \boldsymbol{x}_{1:N}, \boldsymbol{C}_{1:N}} = \boldsymbol{\Sigma}_{\boldsymbol{u}} \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + m_l^{(q)}(N) \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*} \right)^{-1} \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*} = \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*} \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + m_l^{(q)}(N) \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*} \right)^{-1} \boldsymbol{\Sigma}_{\boldsymbol{u}}, \tag{5.98}$$

and

$$m_l^{(q)}(N) = \sum_{n=1}^{N} \mathbb{1}(C_n^{(q)} = l). \tag{5.99}$$

Furthermore,

$$\boldsymbol{\mu}^{(q)}_{\boldsymbol{\theta}_l^* \mid \boldsymbol{x}_{1:N}, \boldsymbol{C}_{1:N}} = \boldsymbol{\Sigma}_{\boldsymbol{u}} \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + m_l^{(q)}(N) \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*} \right)^{-1} \boldsymbol{\mu}_{\boldsymbol{\theta}^*} + m_l^{(q)}(N) \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*} \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + m_l^{(q)}(N) \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*} \right)^{-1} \bar{\boldsymbol{x}}_l^{(q)} \tag{5.100}$$

and

$$\bar{\boldsymbol{x}}_l^{(q)} = \frac{1}{m_l^{(q)}(N)} \sum_{n : C_n^{(q)} = l} \boldsymbol{x}_n^{(q-1)}. \tag{5.101}$$

Sampling of $\theta^*_{l_{\text{new}}}$

We recall that if $C_n^{(q)} = l_{\text{new}}$ was obtained, we need to sample a new cluster hyperparameter $\boldsymbol{\theta}_{l_{\text{new}}}^{*(q-1)}$. This is done directly after sampling the cluster assignment variable $C_n^{(q)}$, which means we have available the cluster assignment variable samples $\mathcal{C}^{(q,q-1)}(\neg n)$ (see (5.74)), and from the previous iteration the cluster hyperparameter samples $\boldsymbol{\theta}^{*(q-1)}_{\mathcal{C}^{(q,q-1)}(\neg n)}$ (see (5.75)), and the samples of the parameters of interest $\boldsymbol{x}^{(q-1)}_{1:N}$ (see (5.76). We can easily adapt (5.97), and obtain

$$f_{\theta^*_{l_{\text{new}}} \mid \theta^*_{\mathscr{C}(N)\setminus\{l_{\text{new}}\}}, \mathbf{C}_{1:N}, \mathbf{x}_{1:N}, \mathbf{y}_{1:N}}(\boldsymbol{\theta}^{*(q-1)}_{l_{\text{new}}} \mid \boldsymbol{\theta}^{*(q-1)}_{\mathcal{C}^{(q,q-1)}(N)\setminus\{l_{\text{new}}\}}, \boldsymbol{C}^{(q,q-1)}_{1:N}, \boldsymbol{x}^{(q-1)}_{1:N}, \boldsymbol{y}_{1:N}) \propto \mathcal{N}(\boldsymbol{\theta}^{*(q-1)}_{l_{\text{new}}}; \boldsymbol{\mu}^{(q-1)}_{\boldsymbol{\theta}^*_{l_{\text{new}}} \mid \boldsymbol{x}_n}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*_{l_{\text{new}}} \mid \boldsymbol{x}_n}). \tag{5.102}$$

By definition, the only parameter of interest associated with the cluster l_{new} is $\boldsymbol{x}_n$, which means that $m_{l_{\text{new}}}^{(q)}(N) = 1$. Therefore the covariance matrix is given by (see (5.92))

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*_{l_{\text{new}}} \mid \boldsymbol{x}_n} = \boldsymbol{\Sigma}_{\boldsymbol{u}} \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}\right)^{-1} \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*} = \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*} \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}\right)^{-1} \boldsymbol{\Sigma}_{\boldsymbol{u}}, \tag{5.103}$$

and the mean $\boldsymbol{\mu}^{(q-1)}_{\boldsymbol{\theta}^*_{l_{\text{new}}} \mid \boldsymbol{x}_n}$ is given by (see (5.94))

$$\boldsymbol{\mu}^{(q-1)}_{\boldsymbol{\theta}^*_{l_{\text{new}}} \mid \boldsymbol{x}_n} = \boldsymbol{\Sigma}_{\boldsymbol{u}} \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}\right)^{-1} \boldsymbol{\mu}_{\boldsymbol{\theta}^*} + \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*} \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}\right)^{-1} \boldsymbol{x}_n^{(q-1)}. \tag{5.104}$$

Full Conditional pdf of x_n

Lastly, samples $\boldsymbol{x}_n^{(q)}$ of the parameters of interest x_n are obtained from the full conditional pdf of x_n given the parameters of interest of the other objects, $\mathsf{x}_{\neg n}$, the cluster hyperparameters $\boldsymbol{\theta}^*_{\mathcal{C}(N)}$, the cluster assignment variables $\mathbf{C}_{1:N}$, and the measurements $\mathbf{y}_{1:N}$, i.e., $f_{\mathsf{x}_n \mid \mathsf{x}_{\neg n}, \boldsymbol{\theta}^*_{\mathcal{C}(N)}, \mathbf{C}_{1:N}, \mathbf{y}_{1:N}}(\boldsymbol{x}_n^{(q)} \mid \boldsymbol{x}_{\neg n}^{(q,q-1)}, \boldsymbol{\theta}^{*(q)}_{\mathcal{C}^{(q)}(N)}, \boldsymbol{C}_{1:N}^{(q)}, \boldsymbol{y}_{1:N})$. Using $\boldsymbol{\theta}_n = \boldsymbol{\theta}^*_{\mathsf{C}_n}$ (see (3.45)) and the fact that conditioning on $\boldsymbol{\theta}_{1:N}$ is equivalent to conditioning on $\mathbf{C}_{1:N}$ and $\boldsymbol{\theta}^*_{\mathscr{C}(N)}$ we obtain

$$f_{\mathsf{x}_n \mid \mathsf{x}_{\neg n}, \boldsymbol{\theta}^*_{\mathscr{C}(N)}, \mathbf{C}_{1:N}, \mathbf{y}_{1:N}}(\boldsymbol{x}_n \mid \boldsymbol{x}_{\neg n}, \boldsymbol{\theta}^*_{\mathcal{C}(N)}, \boldsymbol{C}_{1:N}, \boldsymbol{y}_{1:N}) = f_{\mathsf{x}_n \mid \mathsf{x}_{\neg n}, \boldsymbol{\theta}_{1:N}, \mathbf{y}_{1:N}}(\boldsymbol{x}_n \mid \boldsymbol{x}_{\neg n}, \boldsymbol{\theta}_{1:N}, \boldsymbol{y}_{1:N}). \tag{5.105}$$

We can therefore adapt the result of (5.47) so that

$$\begin{aligned} f_{\mathsf{x}_n \mid \mathsf{x}_{\neg n}, \boldsymbol{\theta}^*_{\mathscr{C}(N)}, \mathbf{C}_{1:N}, \mathbf{y}_{1:N}}(\boldsymbol{x}_n \mid \boldsymbol{x}_{\neg n}, \boldsymbol{\theta}^*_{\mathcal{C}(N)}, \boldsymbol{C}_{1:N}, \boldsymbol{y}_{1:N}) &= f_{\mathsf{x}_n \mid \mathsf{y}_n, \boldsymbol{\theta}^*_{\mathsf{C}_n}}(\boldsymbol{x}_n \mid \boldsymbol{y}_n, \boldsymbol{\theta}^*_{C_n}) \\ &= \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}_{\boldsymbol{x} \mid \boldsymbol{y}_n, \boldsymbol{\theta}^*_{C_n}}, \boldsymbol{\Sigma}_{\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\theta}}) \end{aligned} \tag{5.106}$$

with (see (5.48))

$$\begin{aligned} \boldsymbol{\Sigma}_{\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\theta}} &= \boldsymbol{\Sigma}_{\boldsymbol{u}} - \boldsymbol{\Sigma}_{\boldsymbol{u}} (\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1} \boldsymbol{\Sigma}_{\boldsymbol{u}} \\ &\overset{(B.4)}{=} \boldsymbol{\Sigma}_{\boldsymbol{v}} (\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1} \boldsymbol{\Sigma}_{\boldsymbol{u}}, \end{aligned} \tag{5.107}$$

and (see (5.49))

$$\boldsymbol{\mu}_{\boldsymbol{x} \mid \boldsymbol{y}_n, \boldsymbol{\theta}^*_{C_n}} = \boldsymbol{\theta}^*_{C_n} + \boldsymbol{\Sigma}_{\boldsymbol{u}} (\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1} (\boldsymbol{y}_n - \boldsymbol{\theta}^*_{C_n}). \tag{5.108}$$

By evaluating (5.106) at the already available samples, we finally obtain

$$f_{\mathsf{x}_n \mid \mathsf{x}_{\neg n}, \theta^*_{\mathscr{C}(N)}, \mathsf{C}_{1:N}, \mathsf{y}_{1:N}}(\boldsymbol{x}_n^{(q)} \mid \boldsymbol{x}_{\neg n}^{(q,q-1)}, \boldsymbol{\theta}^{*(q)}_{\mathcal{C}^{(q)}(N)}, \boldsymbol{C}_{1:N}^{(q)}, \boldsymbol{y}_{1:N}) = \mathcal{N}(\boldsymbol{x}_n^{(q)}; \boldsymbol{\mu}^{(q)}_{\boldsymbol{x} \mid \boldsymbol{y}_n, \boldsymbol{\theta}^*_{C_n}}, \boldsymbol{\Sigma}_{\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\theta}}) \tag{5.109}$$

with $\boldsymbol{x}_{\neg n}^{(q,q-1)}$ previously defined in (5.22) and

$$\boldsymbol{\mu}^{(q)}_{\boldsymbol{x} \mid \boldsymbol{y}_n, \boldsymbol{\theta}^*_{C_n}} = \boldsymbol{\theta}^{*(q)}_{C_n^{(q)}} + \boldsymbol{\Sigma}_{\boldsymbol{u}}(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}(\boldsymbol{y}_n - \boldsymbol{\theta}^{*(q)}_{C_n^{(q)}}). \tag{5.110}$$

Pseudocode for the Gibbs Sampler Using Cluster Assignment Variables

The pseudocode for the q^{th} iteration of the Gibbs sampling algorithm, using the cluster assignment variables $\mathbf{C}_{1:N}$ is given in Algorithm 3.

Algorithm 3 Gibbs sampler using cluster assignment variables

Input: $\boldsymbol{\theta}_{1:N}^{(q-1)}$, $\boldsymbol{C}_{1:N}^{(q-1)}$, $l_{\max}^{(q-1)}$, $\boldsymbol{x}_{1:N}^{(q-1)}$, $\boldsymbol{y}_{1:N}$
  for all $n = 1, \ldots, N$ **do**
    sample $C_n^{(q)}$ from $p_{\mathsf{C}_n \mid \mathsf{C}_{\neg n}, \theta^*_{\mathscr{C}(\neg n)}, \mathsf{x}_{1:N}, \mathsf{y}_{1:N}}(l \mid \boldsymbol{C}_{\neg n}^{(q,q-1)}, \boldsymbol{\theta}^{*(q-1)}_{\mathcal{C}^{(q,q-1)}(\neg n)}, \boldsymbol{x}_{1:N}^{(q-1)}, \boldsymbol{y}_{1:N})$ as given by (5.77)
    if $C_n^{(q)} = l_{\text{new}}$ **then**
      set $l_{\text{new}} = \max\{\max_{n' \in \{1,\ldots,n-1\}} C_{n'}^{(q)}, l_{\max}^{(q-1)}\} + 1$ (see (5.81)) and sample $\boldsymbol{\theta}^{*(q-1)}_{l_{\text{new}}}$ from $f_{\theta^*_{l_{\text{new}}} \mid \theta^*_{\mathscr{C}(N)\setminus\{l_{\text{new}}\}}, \mathsf{C}_{1:N}, \mathsf{x}_{1:N}, \mathsf{y}_{1:N}}(\boldsymbol{\theta}^{*(q-1)}_{l_{\text{new}}} \mid \boldsymbol{\theta}^{*(q-1)}_{\mathcal{C}^{(q,q-1)}(N)\setminus\{l_{\text{new}}\}}, \boldsymbol{C}_{1:N}^{(q,q-1)}, \boldsymbol{x}_{1:N}^{(q-1)}, \boldsymbol{y}_{1:N})$ as given by (5.102)
    end if
  end for
  set $l_{\max}^{(q)} = \max\{\max_{n \in \{1,\ldots,N\}} C_n^{(q)}, l_{\max}^{(q-1)}\}$
  for all $l \in \mathcal{C}^{(q)}(N)$ **do**
    sample $\boldsymbol{\theta}_l^{*(q)}$ from $f_{\theta^*_l \mid \theta^*_{\mathscr{C}(N)\setminus\{l\}}, \mathbf{C}_{1:N}, \mathbf{x}_{1:N}, \mathbf{y}_{1:N}}(\boldsymbol{\theta}_l^{*(q)} \mid \boldsymbol{\theta}^{*(q,q-1)}_{\mathcal{C}^{(q)}(N)\setminus\{l\}}, \boldsymbol{C}_{1:N}^{(q)}, \boldsymbol{x}_{1:N}^{(q-1)}, \boldsymbol{y}_{1:N})$ as given by (5.97)
  end for
  for all $n = 1, \ldots, N$ **do**
    assign $\boldsymbol{\theta}_n^{(q)} = \boldsymbol{\theta}^{*(q)}_{C_n^{(q)}}$
  end for
  for all $n = 1, \ldots, N$ **do**
    sample $\boldsymbol{x}_n^{(q)}$ from $f_{\mathsf{x}_n \mid \mathsf{x}_{\neg n}, \theta^*_{\mathscr{C}(N)}, \mathsf{C}_{1:N}, \mathsf{y}_{1:N}}(\boldsymbol{x}_n^{(q)} \mid \boldsymbol{x}_{\neg n}^{(q,q-1)}, \boldsymbol{\theta}^{*(q)}_{\mathcal{C}^{(q)}(N)}, \boldsymbol{C}_{1:N}^{(q)}, \boldsymbol{y}_{1:N})$ as given by (5.109))
  end for
Output: $\boldsymbol{\theta}_{1:N}^{(q)}$, $\boldsymbol{C}_{1:N}^{(q)}$, $l_{\max}^{(q)}$, $\boldsymbol{x}_{1:N}^{(q)}$

Similarly to the "simple" Gibbs sampler (see (5.52)), the Gibbs sampling algorithm using the cluster assignment variables $\mathbf{C}_{1:N}$ is initialized for $q = 0$ by sampling the hyperparameters

$\boldsymbol{\theta}_n^{(0)}$ from the Gaussian base distribution, i.e.,

$$\boldsymbol{\theta}_n^{(0)} \sim \mathcal{N}(\boldsymbol{\theta}_n^{(0)}; \boldsymbol{\mu}_{\boldsymbol{\theta}^*}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}), \quad n = 1, \ldots, N. \tag{5.111}$$

Moreover, the cluster assignment variables C_n are initialized according to

$$C_n^{(0)} = n, \quad n = 1, \ldots, N. \tag{5.112}$$

This means that in the zeroth iteration of the algorithm, each object is assigned to an individual exclusive cluster. Finally, the parameter samples $\boldsymbol{x}_n^{(0)}$ are again initialized using the measurements $\mathbf{y}_n$ as in (5.53), i.e.,

$$\boldsymbol{x}_n^{(0)} = \boldsymbol{\mu}_{\boldsymbol{\theta}^*} + (\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*} + \boldsymbol{\Sigma}_{\boldsymbol{u}})(\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*} + \boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}(\boldsymbol{y}_n - \boldsymbol{\mu}_{\boldsymbol{\theta}^*}), \quad n = 1, \ldots, N. \tag{5.113}$$

5.2.4 MSE

No closed-form expression, but can be estimated by empirical MSE, as described in Section 6.2.

5.3 Fourth Scenario

In our final scenario, we assume that we know the cluster assignment variables $\mathbf{C}_{1:N} = (\mathsf{C}_1, \ldots, \mathsf{C}_N)^{\mathrm{T}}$. We recall that the statistical model for this scenario was introduced in Section 5.1. The cluster assignment variables $\mathbf{C}_{1:N}$ together with the measurements $\mathbf{y}_{1:N}$ are now considered as data that are known to the estimator; however, we still model them as random variables. Since $\boldsymbol{\theta}_n = \boldsymbol{\theta}^*_{\mathsf{C}_n}$ (see (3.40)), we know that objects n, for which $\mathsf{C}_n = l$ are associated with cluster hyperparameter $\boldsymbol{\theta}^*_l$; however, we do not know the realization|(value) of $\boldsymbol{\theta}^*_l$. The statistical dependencies for this scenario are visualized in Figure 9. Note that the only difference from Figure 8 is that the cluster assignment variables are considered known to the estimator.

Contrary to Scenario 3, where we had to approximate the MMSE estimator using numerical methods, the present Scenario 4 admits a closed form solution. Since the cluster assignment variables C_n are known, the estimator can now use all the measurements $\mathbf{y}_n$, that are associated with the same cluster, i.e., for all n such that $\mathsf{C}_n = l$, instead of just using one measurement $\mathbf{y}_n$ as in Scenario 1. Therefore, we expect a lower MSE than in the

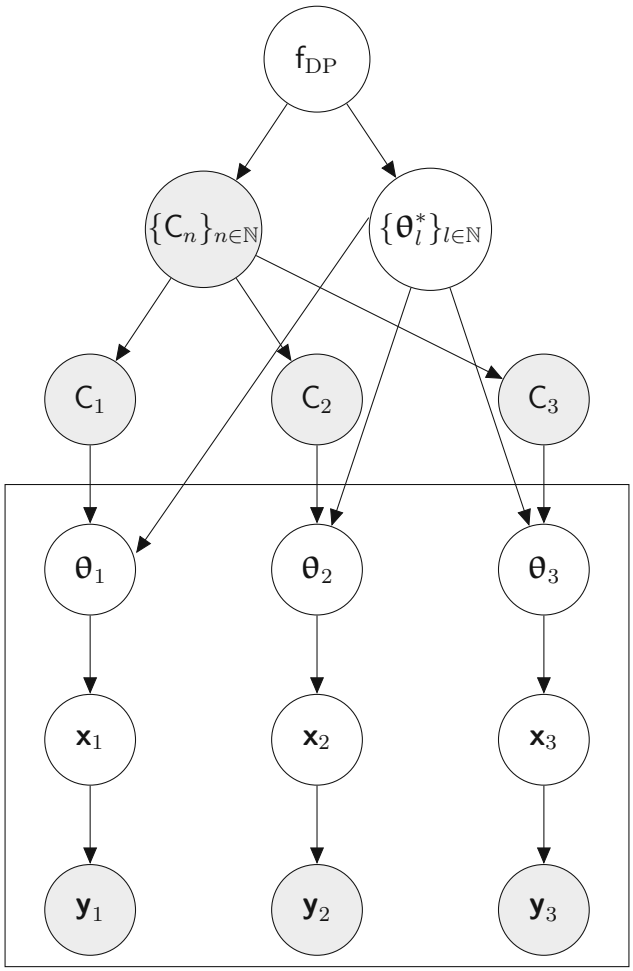


Figure 9: Bayesian network for the fourth scenario, assuming three objects $N = 3$. Random variables displayed in shaded disks are observed. The cluster assignment variables $\{\mathsf{C}_n\}_{n\in\mathbb{N}}$ and the cluster hyperparameters $\{\boldsymbol{\theta}_l^*\}_{l\in\mathbb{N}}$ are generated from the DP. Each cluster assignment variable is then related to exactly one hyperparameter $\boldsymbol{\theta}_n$; however, the cluster assignment variables may be equal for two different objects and thereby relate these objects to the same cluster hyperparameter $\boldsymbol{\theta}_l^*$.

first scenario; however, since the cluster hyperparameters $\boldsymbol{\theta}_l^*$ are unknown, the MSE will be higher than in Scenario 2. In Scenario 3, the estimator also uses clustering to improve the estimate; however, since the cluster assignment variables $\mathbf{C}_{1:N}$ need to be inferred (see (5.77)), the MSE is higher than in the present Scenario 4.

5.3.1 MMSE Estimator

As in our previous scenarios, we are interested in the MMSE estimator, i.e.,

$$\hat{\boldsymbol{x}}_n^{(4)}(\boldsymbol{y}_{1:N}, \boldsymbol{C}_{1:N}) = \mathbb{E}[\mathbf{x}_n \,|\, \mathbf{y}_{1:N} = \boldsymbol{y}_{1:N}, \mathbf{C}_{1:N} = \boldsymbol{C}_{1:N}] = \int_{\boldsymbol{x}_n} \boldsymbol{x}_n \, f_{\mathbf{x}_n \,|\, \mathbf{y}_{1:N}, \mathbf{C}_{1:N}}(\boldsymbol{x}_n \,|\, \boldsymbol{y}_{1:N}, \boldsymbol{C}_{1:N}) \, d\boldsymbol{x}_n. \tag{5.114}$$

Here, the posterior pdf $f_{\mathbf{x}_n \,|\, \mathbf{y}_{1;N}, \mathbf{C}_{1;N}}(\boldsymbol{x}_n \,|\, \boldsymbol{y}_{1:N}, \boldsymbol{C}_{1:N})$ can be obtained from the joint posterior pdf $f_{\mathbf{x}_{1:N} \,|\, \mathbf{y}_{1:N}, \mathbf{C}_{1:N}}(\boldsymbol{x}_{1:N} \,|\, \boldsymbol{y}_{1:N}, \boldsymbol{C}_{1:N})$ as

$$f_{\mathbf{x}_n \,|\, \mathbf{y}_{1;N}, \mathbf{C}_{1;N}}(\boldsymbol{x}_n \,|\, \boldsymbol{y}_{1:N}, \boldsymbol{C}_{1:N}) = \int_{\boldsymbol{x}_{\neg n}} f_{\mathbf{x}_{1:N} \,|\, \mathbf{y}_{1:N}, \mathbf{C}_{1:N}}(\boldsymbol{x}_{1:N} \,|\, \boldsymbol{y}_{1:N}, \boldsymbol{C}_{1:N}) d\boldsymbol{x}_{\neg n}, \tag{5.115}$$

where we recall that $\boldsymbol{x}_{\neg n} = (\boldsymbol{x}_1^{\mathrm{T}}, \ldots, \boldsymbol{x}_{n-1}^{\mathrm{T}}, \boldsymbol{x}_{n+1}^{\mathrm{T}}, \ldots, \boldsymbol{x}_N^{\mathrm{T}})^{\mathrm{T}}$. In order to calculate the joint posterior pdf, we first condition on the cluster hyperparameters $\boldsymbol{\theta}^*_{\mathscr{C}(N)}$, which results in

$$\begin{aligned} &f_{\mathbf{x}_{1:N} \,|\, \mathbf{y}_{1:N}, \mathbf{C}_{1:N}}(\boldsymbol{x}_{1:N} \,|\, \boldsymbol{y}_{1:N}, \boldsymbol{C}_{1:N}) \\ &= \int_{\boldsymbol{\theta}^*_{\mathcal{C}(N)}} f_{\mathbf{x}_{1:N} \,|\, \mathbf{y}_{1:N}, \boldsymbol{\theta}^*_{\mathscr{C}(N)}, \mathbf{C}_{1:N}}(\boldsymbol{x}_{1:N} \,|\, \boldsymbol{y}_{1:N}, \boldsymbol{\theta}^*_{\mathcal{C}(N)}, \boldsymbol{C}_{1:N}) f_{\boldsymbol{\theta}^*_{\mathscr{C}(N)} \,|\, \mathbf{y}_{1:N}, \mathbf{C}_{1:N}}(\boldsymbol{\theta}^*_{\mathcal{C}(N)} \,|\, \boldsymbol{y}_{1:N}, \boldsymbol{C}_{1:N}) d\boldsymbol{\theta}^*_{\mathcal{C}(N)}. \end{aligned} \tag{5.116}$$

First factor of integrand (5.116)

Let us consider the first factor of the integrand in (5.116). As discussed in Section 3.3.2, the cluster assignment variables $\mathbf{C}_{1:N}$ together with the cluster hyperparameters $\boldsymbol{\theta}^*_{\mathscr{C}(N)}$ are statistically equivalent to the hyperparameters $\boldsymbol{\theta}_{1:N}$, since $\boldsymbol{\theta}_n = \boldsymbol{\theta}^*_{\mathbf{C}_n}$ (see (3.45)). This also means

$$f_{\mathbf{x}_{1:N} \,|\, \mathbf{y}_{1:N}, \boldsymbol{\theta}^*_{\mathscr{C}(N)}, \mathbf{C}_{1:N}}(\boldsymbol{x}_{1:N} \,|\, \boldsymbol{y}_{1:N}, \boldsymbol{\theta}^*_{\mathcal{C}(N)}, \boldsymbol{C}_{1:N}) = f_{\mathbf{x}_{1:N} \,|\, \mathbf{y}_{1:N}, \boldsymbol{\theta}_{1:N}}(\boldsymbol{x}_{1:N} \,|\, \boldsymbol{y}_{1:N}, \boldsymbol{\theta}_{1:N}). \tag{5.117}$$

This posterior distribution was already derived in Section 4.3.2, and since our assumptions about the noise vectors $\mathbf{u}_{1:N}$ and $\mathbf{v}_{1:N}$ are the same as in the previous cases, the result in (4.75) is still valid, yielding

$$f_{\mathbf{x}_{1:N} \,|\, \mathbf{y}_{1:N}, \boldsymbol{\theta}_{1:N}}(\boldsymbol{x}_{1:N} \,|\, \boldsymbol{y}_{1:N}, \boldsymbol{\theta}_{1:N}) = \prod_{n=1}^{N} f_{\mathbf{x}_n \,|\, \mathbf{y}_n, \boldsymbol{\theta}_n}(\boldsymbol{x}_n \,|\, \boldsymbol{y}_n, \boldsymbol{\theta}_n). \tag{5.118}$$

Inserting this expression into (5.117) and rewriting the product $\prod_{n=1}^{N}$ using $\mathbf{C}_{1:N}$ and $\boldsymbol{\theta}^*_{\mathscr{C}(N)}$, we obtain

$$f_{\mathbf{x}_{1:N} \,|\, \mathbf{y}_{1:N}, \boldsymbol{\theta}^*_{\mathscr{C}(N)}, \mathbf{C}_{1:N}}(\boldsymbol{x}_{1:N} \,|\, \boldsymbol{y}_{1:N}, \boldsymbol{\theta}^*_{\mathcal{C}(N)}, \boldsymbol{C}_{1:N}) = \prod_{l \in \mathcal{C}(N)} \prod_{n: C_n = l} f_{\mathbf{x}_n \,|\, \mathbf{y}_n, \boldsymbol{\theta}_n}(\boldsymbol{x}_n \,|\, \boldsymbol{y}_n, \boldsymbol{\theta}^*_l). \tag{5.119}$$

Using (4.79) we obtain

$$f_{\mathsf{x}_n \,|\, \mathsf{y}_n, \uptheta_l^*, \mathsf{C}_n}(\boldsymbol{x}_n \,|\, \boldsymbol{y}_n, \boldsymbol{\theta}_l^*, C_n) = \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}_{\boldsymbol{x}\,|\,\boldsymbol{y}_n, \boldsymbol{\theta}_l^*}, \boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y},\boldsymbol{\theta}}), \tag{5.120}$$

with (see (4.83))

$$\boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y},\boldsymbol{\theta}} = \boldsymbol{\Sigma}_{\boldsymbol{v}} \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}}\right)^{-1} \boldsymbol{\Sigma}_{\boldsymbol{u}} \tag{5.121}$$

and (see (4.84))

$$\boldsymbol{\mu}_{\boldsymbol{x}_n\,|\,\boldsymbol{y}_n,\boldsymbol{\theta}_l^*} = \boldsymbol{\theta}_l^* + \boldsymbol{\Sigma}_{\boldsymbol{u}}(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}(\boldsymbol{y}_n - \boldsymbol{\theta}_l^*), \tag{5.122}$$

or equivalently

$$\boldsymbol{\mu}_{\boldsymbol{x}_n\,|\,\boldsymbol{y}_n,\boldsymbol{\theta}_{C_n}^*} = \boldsymbol{\theta}_{C_n}^* + \boldsymbol{\Sigma}_{\boldsymbol{u}}(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}(\boldsymbol{y}_n - \boldsymbol{\theta}_{C_n}^*). \tag{5.123}$$

Finally, inserting (5.120) into (5.119), we obtain

$$f_{\mathsf{x}_{1:N}\,|\,\mathsf{y}_{1:N},\uptheta^*_{\mathscr{C}(N)},\mathsf{C}_{1:N},}(\boldsymbol{x}_{1:N}\,|\,\boldsymbol{y}_{1:N},\boldsymbol{\theta}^*_{\mathcal{C}(N)},\boldsymbol{C}_{1:N}) = \prod_{l\in\mathcal{C}(N)}\prod_{n:C_n=l}\mathcal{N}(\boldsymbol{x}_n;\boldsymbol{\mu}_{\boldsymbol{x}\,|\,\boldsymbol{y}_n,\boldsymbol{\theta}_l^*},\boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y},\boldsymbol{\theta}}). \tag{5.124}$$

Second factor of integrand (5.116)

We now take up the second factor of the integrand in (5.116). Using Bayes' theorem, we have

$$f_{\uptheta^*_{\mathcal{C}(N)}\,|\,\mathsf{y}_{1:N},\mathsf{C}_{1:N}}(\boldsymbol{\theta}^*_{\mathcal{C}(N)}\,|\,\boldsymbol{y}_{1:N},\boldsymbol{C}_{1:N}) \propto f_{\mathsf{y}_{1:N}\,|\,\uptheta^*_{\mathcal{C}(N)},\mathsf{C}_{1:N}}(\boldsymbol{y}_{1:N}\,|\,\boldsymbol{\theta}^*_{\mathcal{C}(N)},\boldsymbol{C}_{1:N}) f_{\uptheta^*_{\mathcal{C}(N)}\,|\,\mathsf{C}_{1:N}}(\boldsymbol{\theta}^*_{\mathcal{C}(N)}\,|\,\boldsymbol{C}_{1:N}). \tag{5.125}$$

By conditioning on $\uptheta_{1:N}$ instead of $\uptheta^*_{\mathscr{C}(N)}$ and $\mathsf{C}_{1:N}$, the first factor in (5.125) can equivalently be written as

$$f_{\mathsf{y}_{1:N}\,|\,\uptheta^*_{\mathscr{C}(N)},\mathsf{C}_{1:N}}(\boldsymbol{y}_{1:N}\,|\,\boldsymbol{\theta}^*_{\mathcal{C}(N)},\boldsymbol{C}_{1:N}) = f_{\mathsf{y}_{1:N}\,|\,\uptheta_{1:N}}(\boldsymbol{y}_{1:N}\,|\,\boldsymbol{\theta}_{1:N}), \tag{5.126}$$

which, using the result in (4.27), yields

$$f_{\mathbf{y}_{1:N}\,|\,\theta^*_{\mathscr{C}(N)},\mathbf{C}_{1:N}}(\boldsymbol{y}_{1:N}\,|\,\boldsymbol{\theta}^*_{\mathcal{C}(N)},\boldsymbol{C}_{1:N}) = \prod_{n=1}^{N}\mathcal{N}(\boldsymbol{y}_n;\boldsymbol{\theta}^*_{C_n},\boldsymbol{\Sigma}_{\boldsymbol{u}}+\boldsymbol{\Sigma}_{\boldsymbol{v}}). \tag{5.127}$$

Similar to (5.119), we rewrite the product $\prod_{n=1}^{N}$ in terms of $\mathbf{C}_{1:N}$ and $\boldsymbol{\theta}^*_{\mathcal{C}(N)}$ and obtain

$$f_{\mathbf{y}_{1:N}\,|\,\theta^*_{\mathscr{C}(N)},\mathbf{C}_{1:N}}(\boldsymbol{y}_{1:N}\,|\,\boldsymbol{\theta}^*_{\mathcal{C}(N)},\boldsymbol{C}_{1:N}) = \prod_{l\in\mathcal{C}(N)}\prod_{n:C_n=l}\mathcal{N}(\boldsymbol{y}_n;\boldsymbol{\theta}^*_l,\boldsymbol{\Sigma}_{\boldsymbol{u}}+\boldsymbol{\Sigma}_{\boldsymbol{v}}). \tag{5.128}$$

The second factor in (5.125) is given by (3.55) as

$$f_{\theta^*_{\mathscr{C}(N)}\,|\,\mathbf{C}_{1:N}}(\boldsymbol{\theta}^*_{\mathcal{C}(N)}\,|\,\boldsymbol{C}_{1:N}) = \prod_{l\in\mathcal{C}(N)} f_{\mathrm{H}}(\boldsymbol{\theta}^*_l) = \prod_{l\in\mathcal{C}(N)}\mathcal{N}(\boldsymbol{\theta}^*_l;\boldsymbol{\mu}_{\boldsymbol{\theta}^*},\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}), \tag{5.129}$$

where in the last step (5.2) was used. Inserting (5.128) and (5.129) into (5.125), we obtain

$$f_{\theta^*_{\mathscr{C}(N)}\,|\,\mathbf{y}_{1:N},\mathbf{C}_{1:N}}(\boldsymbol{\theta}^*_{\mathcal{C}(N)}\,|\,\boldsymbol{y}_{1:N},\boldsymbol{C}_{1:N}) \propto \prod_{l\in\mathcal{C}(N)}\left(\prod_{n:C_n=l}\mathcal{N}(\boldsymbol{y}_n;\boldsymbol{\theta}^*_l,\boldsymbol{\Sigma}_{\boldsymbol{u}}+\boldsymbol{\Sigma}_{\boldsymbol{v}})\right)\mathcal{N}(\boldsymbol{\theta}^*_l;\boldsymbol{\mu}_{\boldsymbol{\theta}^*},\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}). \tag{5.130}$$

We note that the product $\left(\prod_{n:C_n=l}\mathcal{N}(\boldsymbol{y}_n;\boldsymbol{\theta}^*_l,\boldsymbol{\Sigma}_{\boldsymbol{u}}+\boldsymbol{\Sigma}_{\boldsymbol{v}})\right)\mathcal{N}(\boldsymbol{\theta}^*_l;\boldsymbol{\mu}_{\boldsymbol{\theta}^*},\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*})$ in (5.130) is of the same format as left-hand side of the identity (2.59), i.e., with the following substitutions $\boldsymbol{x}\to\boldsymbol{\theta}^*_l$, $\boldsymbol{\mu}_{\boldsymbol{x}}\to\boldsymbol{\mu}_{\boldsymbol{\theta}^*}$, $\boldsymbol{\Sigma}_{\boldsymbol{x}}\to\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}$, and $\boldsymbol{\Sigma}_{\boldsymbol{y}}\to\boldsymbol{\Sigma}_{\boldsymbol{u}}+\boldsymbol{\Sigma}_{\boldsymbol{v}}$, we obtain

$$\left(\prod_{n:C_n=l}\mathcal{N}(\boldsymbol{y}_n;\boldsymbol{\theta}^*_l,\boldsymbol{\Sigma}_{\boldsymbol{u}}+\boldsymbol{\Sigma}_{\boldsymbol{v}})\right)\mathcal{N}(\boldsymbol{\theta}^*_l;\boldsymbol{\mu}_{\boldsymbol{\theta}^*},\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}) = \tilde{\gamma}\mathcal{N}(\boldsymbol{\theta}^*_l;\boldsymbol{\mu}_{\boldsymbol{\theta}^*_l\,|\,\boldsymbol{y}_{1:N},\boldsymbol{C}_{1:N}},\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*_l\,|\,\boldsymbol{y}_{1:N},\boldsymbol{C}_{1:N}}), \tag{5.131}$$

with the covariance matrix given by (see (2.53))

$$\begin{aligned}\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*_l\,|\,\boldsymbol{y}_{1:N},\boldsymbol{C}_{1:N}} &= (\boldsymbol{\Sigma}_{\boldsymbol{u}}+\boldsymbol{\Sigma}_{\boldsymbol{v}})\left(\boldsymbol{\Sigma}_{\boldsymbol{u}}+\boldsymbol{\Sigma}_{\boldsymbol{v}}+m_l(N)\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}\right)^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*} &(5.132)\\ &= \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}\left(\boldsymbol{\Sigma}_{\boldsymbol{u}}+\boldsymbol{\Sigma}_{\boldsymbol{v}}+m_l(N)\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}\right)^{-1}(\boldsymbol{\Sigma}_{\boldsymbol{u}}+\boldsymbol{\Sigma}_{\boldsymbol{v}}) &(5.133)\end{aligned}$$

and mean (see (2.56))

$$\begin{aligned}\boldsymbol{\mu}_{\boldsymbol{\theta}^*_l\,|\,\boldsymbol{y}_{1:N},\boldsymbol{C}_{1:N}} &= (\boldsymbol{\Sigma}_{\boldsymbol{u}}+\boldsymbol{\Sigma}_{\boldsymbol{v}})\left(\boldsymbol{\Sigma}_{\boldsymbol{u}}+\boldsymbol{\Sigma}_{\boldsymbol{v}}+m_l(N)\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}\right)^{-1}\boldsymbol{\mu}_{\boldsymbol{\theta}^*}\\ &\quad+m_l(N)\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}\left(\boldsymbol{\Sigma}_{\boldsymbol{u}}+\boldsymbol{\Sigma}_{\boldsymbol{v}}+m_l(N)\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}\right)^{-1}\bar{\boldsymbol{y}}_l. &(5.134)\end{aligned}$$

Here, $m_l(N)$ denotes the number of objects for which $\mathsf{C}_n = l$, as defined in (3.47), and $\bar{\boldsymbol{y}}_l$ denotes the sample mean of all measurements $(\boldsymbol{y}_n)_{n:C_n=l}$, i.e.,

$$\bar{\boldsymbol{y}}_l \triangleq \frac{1}{m_l(N)} \sum_{n:C_n=l} \boldsymbol{y}_n. \tag{5.135}$$

We note that both the mean and the covariance matrix depend on the cluster index l. Lastly, with the aforementioned substitutions, according to (2.61) the multiplicative factor $\tilde{\gamma}$ in (5.131) does not depend on $\boldsymbol{\theta}_l^*$, therefore inserting (5.131) into (5.130) yields

$$f_{\Theta^*_{\mathscr{C}(N)} \,|\, \mathsf{y}_{1:N}, \mathsf{C}_{1:N}}(\boldsymbol{\theta}^*_{\mathcal{C}(N)} \,|\, \boldsymbol{y}_{1:N}, \boldsymbol{C}_{1:N}) \propto \prod_{l \in \mathcal{C}(N)} \mathcal{N}(\boldsymbol{\theta}_l^*; \boldsymbol{\mu}_{\boldsymbol{\theta}_l^* \,|\, \boldsymbol{y}_{1:N}, \boldsymbol{C}_{1:N}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}_l^* \,|\, \boldsymbol{y}_{1:N}, \boldsymbol{C}_{1:N}}). \tag{5.136}$$

Evaluation of (5.116)

Inserting (5.124) and (5.136) into (5.116) yields

$$\begin{aligned}
&f_{\mathsf{x}_{1:N} \,|\, \mathsf{y}_{1:N}, \mathsf{C}_{1:N}}(\boldsymbol{x}_{1:N} \,|\, \boldsymbol{y}_{1:N}, \boldsymbol{C}_{1:N}) \\
&\propto \int_{\boldsymbol{\theta}^*_{\mathcal{C}(N)}} \prod_{l \in \mathcal{C}(N)} \left(\prod_{n:C_n=l} \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}_{\boldsymbol{x} \,|\, \boldsymbol{y}_n, \boldsymbol{\theta}_l^*}, \boldsymbol{\Sigma}_{\boldsymbol{x} \,|\, \boldsymbol{y}, \boldsymbol{\theta}}) \right) \mathcal{N}(\boldsymbol{\theta}_l^*; \boldsymbol{\mu}_{\boldsymbol{\theta}_l^* \,|\, \boldsymbol{y}_{1:N}, \boldsymbol{C}_{1:N}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}_l^* \,|\, \boldsymbol{y}_{1:N}, \boldsymbol{C}_{1:N}}) d\boldsymbol{\theta}^*_{\mathcal{C}(N)} \\
&= \prod_{l \in \mathcal{C}(N)} \int_{\boldsymbol{\theta}_l^*} \left(\prod_{n:C_n=l} \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}_{\boldsymbol{x} \,|\, \boldsymbol{y}_n, \boldsymbol{\theta}_l^*}, \boldsymbol{\Sigma}_{\boldsymbol{x} \,|\, \boldsymbol{y}, \boldsymbol{\theta}}) \right) \mathcal{N}(\boldsymbol{\theta}_l^*; \boldsymbol{\mu}_{\boldsymbol{\theta}_l^* \,|\, \boldsymbol{y}_{1:N}, \boldsymbol{C}_{1:N}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}_l^* \,|\, \boldsymbol{y}_{1:N}, \boldsymbol{C}_{1:N}}) d\boldsymbol{\theta}_l^*.
\end{aligned} \tag{5.137}$$

To simplify the notation, we write the mean $\boldsymbol{\mu}_{\boldsymbol{x} \,|\, \boldsymbol{y}_n, \boldsymbol{\theta}_l^*}$ (see (5.123)) as

$$\begin{aligned}
\boldsymbol{\mu}_{\boldsymbol{x} \,|\, \boldsymbol{y}_n, \boldsymbol{\theta}_l^*} &= \boldsymbol{\theta}_l^* + \boldsymbol{\Sigma}_{\boldsymbol{u}} \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}} \right)^{-1} \boldsymbol{y}_n - \boldsymbol{\Sigma}_{\boldsymbol{u}} \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}} \right)^{-1} \boldsymbol{\theta}_l^* \\
&= \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}} \right) \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}} \right)^{-1} \boldsymbol{\theta}_l^* + \boldsymbol{\Sigma}_{\boldsymbol{u}} \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}} \right)^{-1} \boldsymbol{y}_n - \boldsymbol{\Sigma}_{\boldsymbol{u}} \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}} \right)^{-1} \boldsymbol{\theta}_l^* \\
&= \boldsymbol{\Sigma}_{\boldsymbol{u}} \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}} \right)^{-1} \boldsymbol{y}_n + \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}} - \boldsymbol{\Sigma}_{\boldsymbol{u}} \right) \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}} \right)^{-1} \boldsymbol{\theta}_l^* \\
&= \boldsymbol{M}_1 \boldsymbol{y}_n + \boldsymbol{M}_2 \boldsymbol{\theta}_l^*,
\end{aligned} \tag{5.138}$$

with

$$\boldsymbol{M}_1 \triangleq \boldsymbol{\Sigma}_{\boldsymbol{u}} \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}} \right)^{-1} \tag{5.139}$$

and

$$\boldsymbol{M}_2 \triangleq \boldsymbol{\Sigma}_{\boldsymbol{v}} \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}} \right)^{-1}, \tag{5.140}$$

where the size of both matrices is $D \times D$. Furthermore, we group all the parameters of interest $\boldsymbol{x}_n$ that belong to the same cluster $C_n = l$, and thus define

$$\boldsymbol{x}_{\Psi_l} \triangleq (\boldsymbol{x}_n)_{n:C_n=l} = (\boldsymbol{x}_{\Psi_l(1)}^{\mathrm{T}}, \ldots, \boldsymbol{x}_{\Psi_l(m_l(N))}^{\mathrm{T}})^{\mathrm{T}}, \tag{5.141}$$

where the function Ψ_l maps each index $p \in \{1, \ldots, m_l(N)\}$ to the corresponding index $n \in \{1, \ldots, N\}$. Similarly, we define

$$\boldsymbol{y}_{\Psi_l} \triangleq (\boldsymbol{y}_n)_{n:C_n=l} = (\boldsymbol{y}_{\Psi_l(1)}^{\mathrm{T}}, \ldots, \boldsymbol{y}_{\Psi_l(m_l(N))}^{\mathrm{T}})^{\mathrm{T}}. \tag{5.142}$$

Using (5.139 - 5.142) we can rewrite (5.137) as

$$\begin{aligned}
&f_{\mathbf{x}_{1:N} \,|\, \mathbf{y}_{1:N}, \mathbf{C}_{1:N}}(\boldsymbol{x}_{1:N} \,|\, \boldsymbol{y}_{1:N}, \boldsymbol{C}_{1:N}) \\
&\propto \prod_{l \in \mathcal{C}(N)} \int_{\boldsymbol{\theta}_l^*} \left(\prod_{p=1}^{m_l(N)} \mathcal{N}(\boldsymbol{x}_{\Psi_l(p)}; \boldsymbol{\mu}_{\boldsymbol{x} \,|\, \boldsymbol{y}_{\Psi_l(p)}, \boldsymbol{\theta}_l^*}, \boldsymbol{\Sigma}_{\boldsymbol{x} \,|\, \boldsymbol{y}, \boldsymbol{\theta}}) \right) \mathcal{N}(\boldsymbol{\theta}_l^*; \boldsymbol{\mu}_{\boldsymbol{\theta}_l^* \,|\, \boldsymbol{y}_{1:N}, \boldsymbol{C}_{1:N}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}_l^* \,|\, \boldsymbol{y}_{1:N}, \boldsymbol{C}_{1:N}}) d\boldsymbol{\theta}_l^* \\
&= \prod_{l \in \mathcal{C}(N)} \int_{\boldsymbol{\theta}_l^*} \left(\prod_{p=1}^{m_l(N)} \mathcal{N}(\boldsymbol{x}_{\Psi_l(p)}; \boldsymbol{M}_1 \boldsymbol{y}_{\Psi_l(p)} + \boldsymbol{M}_2 \boldsymbol{\theta}_l^*, \boldsymbol{\Sigma}_{\boldsymbol{x} \,|\, \boldsymbol{y}, \boldsymbol{\theta}}) \right) \mathcal{N}(\boldsymbol{\theta}_l^*; \boldsymbol{\mu}_{\boldsymbol{\theta}_l^* \,|\, \boldsymbol{y}_{1:N}, \boldsymbol{C}_{1:N}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}_l^* \,|\, \boldsymbol{y}_{1:N}, \boldsymbol{C}_{1:N}}) d\boldsymbol{\theta}_l^*.
\end{aligned} \tag{5.143}$$

Furthermore, with the following substitutions in (C.35) $\boldsymbol{A} \to \boldsymbol{M}_2$, $\boldsymbol{B} \to \boldsymbol{M}_1$, $\boldsymbol{x} \to \boldsymbol{\theta}_l^*$, $\boldsymbol{y}_n \to \boldsymbol{x}_{\Psi_l(p)}$, $\boldsymbol{\mu}_{\boldsymbol{x}} \to \boldsymbol{\mu}_{\boldsymbol{\theta}_l^* \,|\, \boldsymbol{y}_{1:N}, \boldsymbol{C}_{1:N}}$, $\boldsymbol{\Sigma}_{\boldsymbol{x}} \to \boldsymbol{\Sigma}_{\boldsymbol{\theta}_l^* \,|\, \boldsymbol{y}_{1:N}, \boldsymbol{C}_{1:N}}$, $\boldsymbol{\Sigma}_{\boldsymbol{y}} \to \boldsymbol{\Sigma}_{\boldsymbol{x} \,|\, \boldsymbol{y}, \boldsymbol{\theta}}$, $N \to m_l(N)$, the expression (5.143) for the joint posterior is given by

$$f_{\mathbf{x}_{1:N} \,|\, \mathbf{y}_{1:N}, \mathbf{C}_{1:N}}(\boldsymbol{x}_{1:N} \,|\, \boldsymbol{y}_{1:N}, \boldsymbol{C}_{1:N}) \propto \prod_{l \in \mathcal{C}(N)} \mathcal{N}\left(\boldsymbol{x}_{\Psi_l}; \tilde{\boldsymbol{\mu}}_{\boldsymbol{x}_{\Psi_l} \,|\, \boldsymbol{y}_{1:N}, \boldsymbol{C}_{1:N}}, \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}_{\Psi_l} \,|\, \boldsymbol{y}_{1:N}, \boldsymbol{C}_{1:N}}\right), \tag{5.144}$$

with the mean $\tilde{\boldsymbol{\mu}}_{\boldsymbol{x}_{\Psi_l} \,|\, \boldsymbol{y}_{1:N}, \boldsymbol{C}_{1:N}}$ (see (C.37))

$$\tilde{\boldsymbol{\mu}}_{\boldsymbol{x}_{\Psi_l} \,|\, \boldsymbol{y}_{1:N}, \boldsymbol{C}_{1:N}} = \begin{pmatrix} \boldsymbol{M}_1 \boldsymbol{y}_{\Psi_l(1)} + \boldsymbol{M}_2 \boldsymbol{\mu}_{\boldsymbol{\theta}_l^* \,|\, \boldsymbol{y}_{1:N}, \boldsymbol{C}_{1:N}} \\ \vdots \\ \boldsymbol{M}_1 \boldsymbol{y}_{\Psi_l(m_l(N))} + \boldsymbol{M}_2 \boldsymbol{\mu}_{\boldsymbol{\theta}_l^* \,|\, \boldsymbol{y}_{1:N}, \boldsymbol{C}_{1:N}} \end{pmatrix} \tag{5.145}$$

and covariance matrix $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}_{\Psi_l} \,|\, \boldsymbol{y}_{1:N}, \boldsymbol{C}_{1:N}}$ (see (C.30) and (C.38))

$$\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}_{\Psi_l} \,|\, \boldsymbol{y}_{1:N}, \boldsymbol{C}_{1:N}} = \mathbf{I}_{m_l(N)} \otimes \boldsymbol{\Sigma}_{\boldsymbol{x} \,|\, \boldsymbol{y}, \boldsymbol{\theta}} + \mathbf{1}_{m_l(N)} \mathbf{1}_{m_l(N)}^{\mathrm{T}} \otimes \boldsymbol{M}_2 \boldsymbol{\Sigma}_{\boldsymbol{\theta}_l^* \,|\, \boldsymbol{y}_{1:N}, \boldsymbol{C}_{1:N}} \boldsymbol{M}_2^{\mathrm{T}}$$

$$
\begin{aligned}
&= \mathbf{I}_{m_l(N)} \otimes \boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y},\boldsymbol{\theta}} + \mathbf{1}_{m_l(N)}\mathbf{1}_{m_l(N)}^{\mathrm{T}} \otimes \boldsymbol{\Sigma}_{\boldsymbol{M}_2} \\
&= \begin{pmatrix}
\boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y},\boldsymbol{\theta}} + \boldsymbol{\Sigma}_{\boldsymbol{M}_2} & \boldsymbol{\Sigma}_{\boldsymbol{M}_2} & \dots & \boldsymbol{\Sigma}_{\boldsymbol{M}_2} \\
\boldsymbol{\Sigma}_{\boldsymbol{M}_2} & \boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y},\boldsymbol{\theta}} + \boldsymbol{\Sigma}_{\boldsymbol{M}_2} & \ddots & \boldsymbol{\Sigma}_{\boldsymbol{M}_2} \\
\vdots & \ddots & \ddots & \vdots \\
\boldsymbol{\Sigma}_{\boldsymbol{M}_2} & \dots & \dots & \boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y},\boldsymbol{\theta}} + \boldsymbol{\Sigma}_{\boldsymbol{M}_2}
\end{pmatrix},
\end{aligned} \tag{5.146}
$$

where $\mathbf{I}_{m_l(N)}$ denotes the identity matrix of size $m_l(N) \times m_l(N)$, $\mathbf{1}_{m_l(N)}$ denotes the all-one vector of size $m_l(N) \times 1$ and

$$
\boldsymbol{\Sigma}_{\boldsymbol{M}_2} \triangleq \boldsymbol{M}_2 \boldsymbol{\Sigma}_{\boldsymbol{\theta}_l^* \,|\, \boldsymbol{y}_{1:N}, \boldsymbol{C}_{1:N}} \boldsymbol{M}_2^{\mathrm{T}}. \tag{5.147}
$$

Finally, we obtain the posterior of $\mathbf{x}_n$ by inserting (5.144) into (5.115) and perform the marginalization, using the results [20, Eq.2.92 and Eq.2.93], i.e.,

$$
\begin{aligned}
f_{\mathbf{x}_n \,|\, \mathbf{y}_{1;N}, \mathbf{C}_{1;N}}(\boldsymbol{x}_n \,|\, \boldsymbol{y}_{1:N}, \boldsymbol{C}_{1:N}) &\propto \int_{\boldsymbol{x}_{\neg n}} \prod_{l \in \mathcal{C}(N)} \mathcal{N}(\boldsymbol{x}_{\Psi_l}; \tilde{\boldsymbol{\mu}}_{\boldsymbol{x}_{\Psi_l} \,|\, \boldsymbol{y}_{1:N}, \boldsymbol{C}_{1:N}}, \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}_{\Psi_l} \,|\, \boldsymbol{y}_{1:N}, \boldsymbol{C}_{1:N}}) d\boldsymbol{x}_{\neg n} \\
&= \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}_{\boldsymbol{x}\,|\,\boldsymbol{y}_{1:N}, \boldsymbol{C}_{1:N}}, \boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y}_{1:N}, \boldsymbol{C}_{1:N}}),
\end{aligned} \tag{5.148}
$$

where the posterior mean is obtained from (5.145) using the index $\Psi_l(p)$ equal to n, and using $C_n = l$, i.e.,

$$
\boldsymbol{\mu}_{\boldsymbol{x}\,|\,\boldsymbol{y}_{1:N}, \boldsymbol{C}_{1:N}} = \boldsymbol{M}_1 \boldsymbol{y}_n + \boldsymbol{M}_2 \boldsymbol{\mu}_{\boldsymbol{\theta}^*_{C_n} \,|\, \boldsymbol{y}_{1:N}, \boldsymbol{C}_{1:N}}, \tag{5.149}
$$

and the posterior covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y}_{1:N}, \boldsymbol{C}_{1:N}}$ is obtained from (5.146) as

$$
\boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y}_{1:N}, \boldsymbol{C}_{1:N}} = \boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y},\boldsymbol{\theta}} + \boldsymbol{\Sigma}_{\boldsymbol{M}_2} = \boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y},\boldsymbol{\theta}} + \boldsymbol{M}_2 \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*_{C_n} \,|\, \boldsymbol{y}_{1:N}, \boldsymbol{C}_{1:N}} \boldsymbol{M}_2^{\mathrm{T}}. \tag{5.150}
$$

Inserting (5.139), (5.140) and the expression (5.134) (with $C_n = l$) into (5.149), we obtain for the posterior mean

$$
\begin{aligned}
&\boldsymbol{\mu}_{\boldsymbol{x}\,|\,\boldsymbol{y}_{1:N}, \boldsymbol{C}_{1:N}} \\
&= \boldsymbol{\Sigma}_{\boldsymbol{u}} \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}}\right)^{-1} \boldsymbol{y}_n + \boldsymbol{\Sigma}_{\boldsymbol{v}} \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}}\right)^{-1} \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}}\right) \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}} + m_{C_n}(N)\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}\right)^{-1} \boldsymbol{\mu}_{\boldsymbol{\theta}^*} \\
&\quad + m_{C_n}(N)\boldsymbol{\Sigma}_{\boldsymbol{v}} \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}}\right)^{-1} \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*} \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}} + m_{C_n}(N)\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}\right)^{-1} \bar{\boldsymbol{y}}_{C_n} \\
&= \boldsymbol{\Sigma}_{\boldsymbol{u}} \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}}\right)^{-1} \boldsymbol{y}_n + \boldsymbol{\Sigma}_{\boldsymbol{v}} \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}} + m_{C_n}(N)\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}\right)^{-1} \boldsymbol{\mu}_{\boldsymbol{\theta}^*}
\end{aligned}
$$

$$+ m_{C_n}(N)\boldsymbol{\Sigma}_{\boldsymbol{v}} \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}}\right)^{-1} \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*} \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}} + m_{C_n}(N)\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}\right)^{-1} \bar{\boldsymbol{y}}_{C_n}, \tag{5.151}$$

with

$$m_{C_n(N)} \triangleq \sum_{n'=1}^{N} \mathbb{1}(C_{n'} = C_n), \tag{5.152}$$

and $\bar{\boldsymbol{y}}_{C_n}$ (see (5.95))

$$\bar{\boldsymbol{y}}_{C_n} \triangleq \frac{1}{m_{C_n}(N)} \sum_{n'=1}^{N} \mathbb{1}(C_{n'} = C_n)\boldsymbol{y}_{n'}. \tag{5.153}$$

Similarly, inserting (5.140), (5.121), and (5.132) into (5.150), the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y}_{1:N},\boldsymbol{C}_{1:N}}$ is obtained as

$$\begin{aligned}\boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y}_{1:N},\boldsymbol{C}_{1:N}} =& \boldsymbol{\Sigma}_{\boldsymbol{v}} \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}}\right)^{-1} \boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}} \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}}\right)^{-1} \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}}\right) \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}} + m_{C_n}(N)\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}\right)^{-1} \\ & \times \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*} \left(\boldsymbol{\Sigma}_{\boldsymbol{v}} \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}}\right)^{-1}\right)^{\mathrm{T}} \\ =& \boldsymbol{\Sigma}_{\boldsymbol{v}} \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}}\right)^{-1} \boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}} \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}} + m_{C_n}(N)\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}\right)^{-1} \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*} \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}}\right)^{-1} \boldsymbol{\Sigma}_{\boldsymbol{v}}.\end{aligned} \tag{5.154}$$

Final expression for $\hat{\boldsymbol{x}}_n^{(4)}(\boldsymbol{y}_{1:N}, \boldsymbol{C}_{1:N})$

According to (5.114), the MMSE estimator $\hat{\boldsymbol{x}}_n^{(4)}(\boldsymbol{y}_{1:N}, \boldsymbol{C}_{1:N})$ is equal to the posterior mean $\mathbb{E}[\mathbf{x}_n \,|\, \mathbf{y}_{1:N} = \boldsymbol{y}_{1:N}, \mathbf{C}_{1:N} = \boldsymbol{C}_{1:N}] = \boldsymbol{\mu}_{\boldsymbol{x}\,|\,\boldsymbol{y}_{1:N},\boldsymbol{C}_{1:N}}$. Using (5.151), we thus obtain

$$\begin{aligned}\hat{\boldsymbol{x}}_n^{(4)}(\boldsymbol{y}_{1:N}, \boldsymbol{C}_{1:N}) =& \boldsymbol{\Sigma}_{\boldsymbol{u}} \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}}\right)^{-1} \boldsymbol{y}_n + \boldsymbol{\Sigma}_{\boldsymbol{v}} \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}} + m_{C_n}(N)\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}\right)^{-1} \boldsymbol{\mu}_{\boldsymbol{\theta}^*} \\ & + m_{C_n}(N)\boldsymbol{\Sigma}_{\boldsymbol{v}} \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}}\right)^{-1} \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*} \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}} + m_{C_n}(N)\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}\right)^{-1} \bar{\boldsymbol{y}}_{C_n},\end{aligned} \tag{5.155}$$

with the sample mean $\bar{\boldsymbol{y}}_{C_n}$ (see (5.153)) and $m_{C_n}(N)$ (see (5.152)). We can see that (5.155) consists of three additive terms, the measurement $\boldsymbol{y}_n$, multiplied by $\boldsymbol{\Sigma}_{\boldsymbol{u}} \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}}\right)^{-1}$, the prior knowledge $\boldsymbol{\mu}_{\boldsymbol{\theta}^*}$, multiplied by $\boldsymbol{\Sigma}_{\boldsymbol{v}} \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}} + m_{C_n}(N)\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}\right)^{-1}$ and the sample mean $\bar{\boldsymbol{y}}_{C_n}$ of all measurements that are in the same cluster C_n as the object $\mathbf{x}_n$, weighted by $m_{C_n}(N)\boldsymbol{\Sigma}_{\boldsymbol{v}} \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}}\right)^{-1} \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*} \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}} + m_{C_n}(N)\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}\right)^{-1}$.

5.3.2 MSE

According to (2.12), the minimum MSE is given by

$$\mathrm{MSE}_{\min}^{(4)} = \frac{1}{D}\mathbb{E}_{\mathbf{y}_{1:N},\mathbf{C}_{1:N}}\Big[\mathrm{tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y}_{1:N},\boldsymbol{C}_{1:N}}\right]\Big], \tag{5.156}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y}_{1:N},\boldsymbol{C}_{1:N}}$ is given by (5.154). This expression involves $m_{C_n}(N)$ (see (5.152)) and is thus functionally dependent on the cluster assignment variables $\mathbf{C}_{1:N}$. On the other hand, expression (5.154) is not functionally dependent on $\mathbf{y}_{1:N}$. Therefore, the minimum MSE for Scenario 4 is finally obtained as

$$\begin{aligned}\mathrm{MSE}_{\min}^{(4)} &= \frac{1}{D}\,\mathbb{E}_{\mathbf{C}_{1:N}}\Big[\mathrm{tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y}_{1:N},\boldsymbol{C}_{1:N}}\right]\Big] &(5.157)\\ &= \frac{1}{D}\,\mathbb{E}_{\mathbf{C}_{1:N}}\Big[\mathrm{tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{v}}\left(\boldsymbol{\Sigma}_{\boldsymbol{u}}+\boldsymbol{\Sigma}_{\boldsymbol{v}}\right)^{-1}\boldsymbol{\Sigma}_{\boldsymbol{u}}+\boldsymbol{\Sigma}_{\boldsymbol{v}}\left(\boldsymbol{\Sigma}_{\boldsymbol{u}}+\boldsymbol{\Sigma}_{\boldsymbol{v}}+\mathsf{m}_{\mathsf{C}_n}(N)\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}\right)^{-1}\\ &\quad\times\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}\left(\boldsymbol{\Sigma}_{\boldsymbol{u}}+\boldsymbol{\Sigma}_{\boldsymbol{v}}\right)^{-1}\boldsymbol{\Sigma}_{\boldsymbol{v}}\right]\Big]. &(5.158)\end{aligned}$$

We note that (5.158) functionally depends on $\mathsf{m}_{\mathsf{C}_n}(N)$ (see (5.152)), hence the expectation in (5.158) with respect to $\mathbf{C}_{1:N}$ can be expressed as

$$\begin{aligned}\mathrm{MSE}_{\min}^{(4)} = \frac{1}{D}\mathbb{E}_{\mathbf{C}_{1:N}\,|\,\mathsf{m}_{\mathsf{C}_n}(N)}\Big[\mathbb{E}_{\mathsf{m}_{\mathsf{C}_n}(N)}\Big[&\mathrm{tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{v}}\left(\boldsymbol{\Sigma}_{\boldsymbol{u}}+\boldsymbol{\Sigma}_{\boldsymbol{v}}\right)^{-1}\boldsymbol{\Sigma}_{\boldsymbol{u}}+\boldsymbol{\Sigma}_{\boldsymbol{v}}\left(\boldsymbol{\Sigma}_{\boldsymbol{u}}+\boldsymbol{\Sigma}_{\boldsymbol{v}}+\mathsf{m}_{\mathsf{C}_n}(N)\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}\right)^{-1}\\ &\times\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}\left(\boldsymbol{\Sigma}_{\boldsymbol{u}}+\boldsymbol{\Sigma}_{\boldsymbol{v}}\right)^{-1}\boldsymbol{\Sigma}_{\boldsymbol{v}}\right]\Big]\Big]. &(5.159)\end{aligned}$$

The $p_{\mathsf{m}_{\mathsf{C}_n}(N)}(m)$ of the cluster size $\mathsf{m}_{\mathsf{C}_n}(N)$, i.e., the number of objects for which $\mathsf{C}_{n'} = \mathsf{C}_n$ (see (5.152)), is shown in Appendix A.2 to be given by

$$p_{\mathsf{m}_{\mathsf{C}_n}(N)}(m) = \frac{\alpha^{\overline{N-m}}(N-1)!}{(\alpha+1)^{\overline{N-1}}(N-m)!}, \quad \text{for} \quad m\in\{1,\ldots,N\}, \tag{5.160}$$

with $\alpha^{\overline{m}} \triangleq \alpha(\alpha+1)\cdots(\alpha+m-1)$ denoting the Pochhammer symbol (with $\alpha^{\overline{0}}=1$ and $\alpha^{\overline{1}}=\alpha$) [36, Eq. 13.154]. Next, by exploiting the fact that given $\mathsf{m}_{\mathsf{C}_n}(N)=m$, the posterior covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y}_{1:N},\boldsymbol{C}_{1:N}}$ is independent of $\mathbf{C}_{1:N}$, we can write (5.159) as

$$\begin{aligned}\mathrm{MSE}_{\min}^{(4)} = \frac{1}{D}\,\mathbb{E}_{\mathsf{m}_{\mathsf{C}_n}(N)}\Big[&\mathrm{tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{v}}\left(\boldsymbol{\Sigma}_{\boldsymbol{u}}+\boldsymbol{\Sigma}_{\boldsymbol{v}}\right)^{-1}\boldsymbol{\Sigma}_{\boldsymbol{u}}+\boldsymbol{\Sigma}_{\boldsymbol{v}}\left(\boldsymbol{\Sigma}_{\boldsymbol{u}}+\boldsymbol{\Sigma}_{\boldsymbol{v}}+\mathsf{m}_{\mathsf{C}_n}(N)\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}\right)^{-1}\\ &\times\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}\left(\boldsymbol{\Sigma}_{\boldsymbol{u}}+\boldsymbol{\Sigma}_{\boldsymbol{v}}\right)^{-1}\boldsymbol{\Sigma}_{\boldsymbol{v}}\right]\Big], &(5.161)\end{aligned}$$

which using (5.160) becomes

$$\begin{aligned}\mathrm{MSE}_{\min}^{(4)} &= \sum_{m=1}^{N} \operatorname{tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{x}\,|\,\boldsymbol{y}_{1:N},\boldsymbol{C}_{1:N}}\right] p_{\mathsf{m}_{\mathsf{C}_n}(N)}(m) \\ &= \frac{1}{D}\sum_{m=1}^{N} \operatorname{tr}\Big[\boldsymbol{\Sigma}_{\boldsymbol{v}}\left(\boldsymbol{\Sigma}_{\boldsymbol{u}}+\boldsymbol{\Sigma}_{\boldsymbol{v}}\right)^{-1}\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}}\left(\boldsymbol{\Sigma}_{\boldsymbol{u}}+\boldsymbol{\Sigma}_{\boldsymbol{v}}+m\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}\right)^{-1} \\ &\quad \times \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}\left(\boldsymbol{\Sigma}_{\boldsymbol{u}}+\boldsymbol{\Sigma}_{\boldsymbol{v}}\right)^{-1}\boldsymbol{\Sigma}_{\boldsymbol{v}}\Big] p_{\mathsf{m}_{\mathsf{C}_n}(N)}(m). \end{aligned} \tag{5.162}$$

In the special case where $\boldsymbol{\theta}_l^*$, $\mathbf{u}_n$ and $\mathbf{v}_n$ are random vectors with i.i.d. components, i.e., $\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*} = \sigma_{\theta^*}^2\mathbf{I}_D$, $\boldsymbol{\Sigma}_{\boldsymbol{u}} = \sigma_u^2\mathbf{I}_D$, and $\boldsymbol{\Sigma}_{\boldsymbol{v}} = \sigma_v^2\mathbf{I}_D$, expression (5.162) reduces to

$$\begin{aligned}\mathrm{MSE}_{\min}^{(4)} &= \sum_{m=1}^{N} \operatorname{tr}\left[\frac{\sigma_u^2\sigma_v^2}{\sigma_u^2+\sigma_v^2}\mathbf{I}_D + \frac{(\sigma_v^2)^2\sigma_{\theta^*}^2}{(\sigma_u^2+\sigma_v^2)(\sigma_u^2+\sigma_v^2+m\sigma_{\theta^*}^2)}\mathbf{I}_D\right] p_{\mathsf{m}_{\mathsf{C}_n}(N)}(m) \\ &= \sum_{m=1}^{N}\left(\frac{\sigma_u^2\sigma_v^2}{\sigma_u^2+\sigma_v^2} + \frac{(\sigma_v^2)^2\sigma_{\theta^*}^2}{(\sigma_u^2+\sigma_v^2)(\sigma_u^2+\sigma_v^2+m\sigma_{\theta^*}^2)}\right)\frac{\alpha^{\overline{N-m}}(N-1)!}{(\alpha+1)^{\overline{N-1}}(N-m)!}. \end{aligned} \tag{5.163}$$

6 Simulation Results

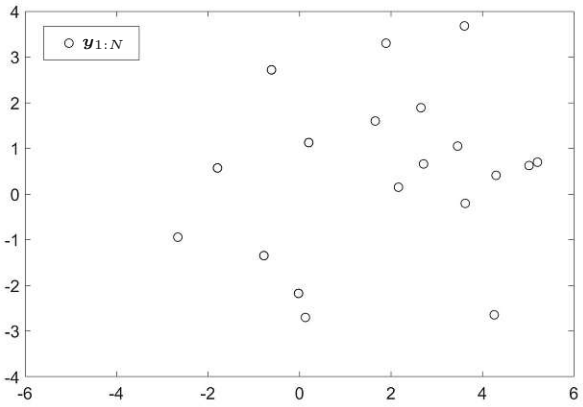
In this section, we evaluate and compare the performance of the estimators introduced in Sections 4 and 5. We will discuss the generation of the data, the simulation parameters, the performance metrics, and the obtained performance results.

6.1 Simulation Setup

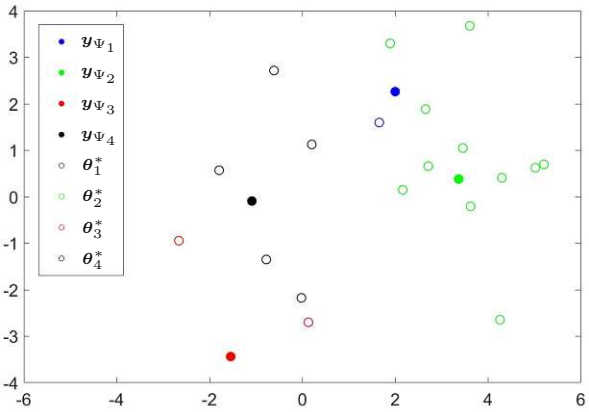
For each object n, we consider a parameter of interest $\mathbf{x}_n = (\mathsf{x}_{n,1}, \mathsf{x}_{n,2})^{\mathrm{T}} \in \mathbb{R}^2$, a hyperparameter $\boldsymbol{\theta}_n = (\theta_{n,1}, \theta_{n,2})^{\mathrm{T}} \in \mathbb{R}^2$, and a measurement $\mathbf{y}_n = (\mathsf{y}_{n,1}, \mathsf{y}_{n,2})^{\mathrm{T}} \in \mathbb{R}^2$. In each simulation run, we generate $\boldsymbol{\theta}_n$ from a DP assuming a Gaussian base distribution $f_{\mathrm{H}}(\boldsymbol{\theta}_l^*) = \mathcal{N}(\boldsymbol{\theta}_l^*; \boldsymbol{\mu}_{\boldsymbol{\theta}^*}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*})$, where $\boldsymbol{\mu}_{\boldsymbol{\theta}^*} = \mathbf{0}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*} = \sigma_{\theta^*}^2 \mathbf{I}_2$ with $\sigma_{\theta^*}^2 = 5$. In the first simulation, we consider three different values $\alpha = 0.5, 1, 5$ of the concentration parameter. Later, we will investigate the performance in terms of the MSE and how the MSE depends on the concentration parameter α, and we will consider $\alpha = 0.1, 0.5, 1 \ldots, 5$. We recall that the hyperparameter vector $\boldsymbol{\theta}_n$ and the parameter of interest $\mathbf{x}_n$ are statistically related according to (4.1), where we assume that the parameter noise $\mathbf{u}_n$ is zero-mean Gaussian (see (4.2)) with $\boldsymbol{\Sigma}_{\boldsymbol{u}} = \sigma_u^2 \mathbf{I}_2$ where $\sigma_u^2 = 1$. Similarly, the measurement vector $\mathbf{y}_n$ and $\mathbf{x}_n$ are statistically related according to (4.4), where we assume that the measurement noise $\mathbf{v}_n$ is zero-mean Gaussian (see (4.5)) with $\boldsymbol{\Sigma}_{\boldsymbol{v}} = \sigma_v^2 \mathbf{I}_2$ where $\sigma_v^2 = 1$. To investigate the performance in terms of the MSE and how the MSE depends on the total number of objects N, we simulate up to $N = 50$ objects, i.e., $N = 1, 2 \ldots, 50$. Each simulation result for one value of N is averaged over $J = 500$ simulation runs, where in each run, we create new hyperparameters $\boldsymbol{\theta}_{1:N}$, parameters of interest $\mathbf{x}_{1:N}$, and measurements $\mathbf{y}_{1:N}$. We create $Q = 1000$ samples $\boldsymbol{x}_n^{(q)}$ from the posterior distribution $f_{\mathbf{x}_n \,|\, \mathbf{y}_{1:N}}(\boldsymbol{x}_n \,|\, \boldsymbol{y}_{1:N})$ by means of the Gibbs sampler using

Table 2: Simulation parameters.

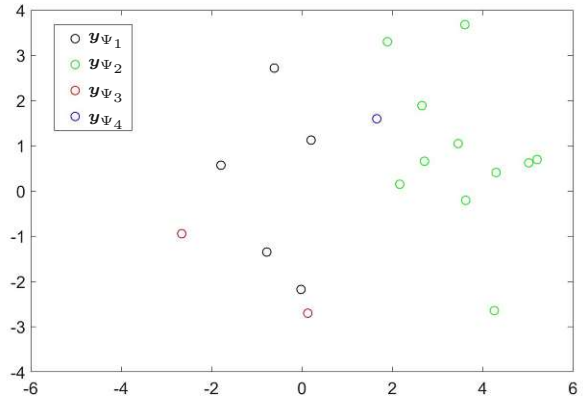
Parameter	Value
$\boldsymbol{\mu}_{\boldsymbol{\theta}^*}$	$\mathbf{0}$
$\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}$	$5\mathbf{I}_2$
$\boldsymbol{\Sigma}_{\boldsymbol{u}}$	$\mathbf{I}_2$
$\boldsymbol{\Sigma}_{\boldsymbol{v}}$	$\mathbf{I}_2$
α	$\{0.1, 0.5, 1 \ldots, 5\}$
N	$\{1, 2, \ldots, 50\}$
Q	1000
J	500



(a) Scenarios 1 and 3.



(b) Scenario 2.



(c) Scenario 4.

Figure 10: One realization of the data available to the estimator for each of the four scenarios, for $N{=}20$ and $\alpha{=}1$. In (a) (Scenarios 1 and 3), these data are only the measurement vector $\boldsymbol{y}_{1:N}$. In (b) (Scenario 2), the data are the measurement vector $\boldsymbol{y}_{1:N}$ and the hyperparameters $\boldsymbol{\theta}_{1:N}$, or, equivalently because $\boldsymbol{\theta}_n = \boldsymbol{\theta}^*_{C_n}$, the hyperparameters $\boldsymbol{\theta}^*_{C_n}$ and cluster assignment variables C_n for $n = 1, \ldots, N$. Using the cluster assignment variables C_n, we group all measurements $\boldsymbol{y}_n$ for which $C_n = l$, i.e., $(\boldsymbol{y}_n)_{n:C_n=l}$, which we denote as $\boldsymbol{y}_{\Psi_l}$ (see (5.142)). The cluster assignment variables C_n are represented by the color of the respective cluster. Lastly, in (c) (Scenario 4), the data are the measurement vector $\boldsymbol{y}_{1:N}$ and the cluster assignment variables C_n for $n{=}1, \ldots, N$. Similar to (b), we group all measurements of the l^{th} cluster into $\boldsymbol{y}_{\Psi_l}$.

cluster assignment variables (see Algorithm 3). The simulation parameters can be seen in Table 2. Figure 10 shows one realization of the data available to the respective estimator for $N = 20$ objects and $\alpha = 1$.

6.2 Performance Metrics

In this subsection, we discuss the metrics we use to analyze the performance of the estimators.

MSE

In order to quantify and compare the performance of the estimators for our four scenarios, we calculate their respective MSEs under the assumption that $\boldsymbol{\theta}_n$, $\mathbf{u}_n$, and $\mathbf{v}_n$ are random vectors with i.i.d. components. For the first, second, and fourth scenarios, the MSE is given in closed form (see (4.62), (4.88), and (5.163), respectively). In the third scenario, no closed-form expression for the MSE or the estimate is available. Therefore, we calculate the empirical MSE, by averaging the squared estimation error over all times $n = 1, \ldots, N$ and all simulation runs $j = 1, \ldots, J$, i.e.,

$$\mathrm{MSE}_{\mathrm{min}}^{(3)} \approx \mathrm{MSE}_{\mathrm{MC}}^{(3)} \triangleq \frac{1}{JN} \sum_{j=1}^{J} \sum_{n=1}^{N} \|\hat{\boldsymbol{x}}_{n,\mathrm{MC}}^{(3)}(\boldsymbol{y}_{1:N}^{(j)}) - \boldsymbol{x}_{n,j}\|_2^2. \tag{6.1}$$

We can approximate the posterior mean (see (5.16) and (5.17)) as

$$\hat{\boldsymbol{x}}_n^{(3)}(\boldsymbol{y}_{1:N,j}) \approx \hat{\boldsymbol{x}}_{n,\mathrm{MC}}^{(3)}(\boldsymbol{y}_{1:N}^{(j)}) \triangleq \frac{1}{Q} \sum_{q=1}^{Q} \boldsymbol{x}_{n,j}^{(q)}, \tag{6.2}$$

where $\boldsymbol{y}_{1:N}^{(j)}$ denotes the measurement vector $\boldsymbol{y}_{1:N}$ obtained in the j^{th} simulation run. Similarly, $\boldsymbol{x}_{n,j}$ stands for the true parameter in the j^{th} simulation run and $\hat{\boldsymbol{x}}_{n,\mathrm{MC}}^{(3)}(\boldsymbol{y}_{1:N}^{(j)})$ is the estimate of the parameter of interest, calculated using the measurements created in the j^{th} simulation run. Table 3 lists the MSE expressions for the four scenarios.

Clustering Gain

Since we expect an improved MSE performance due to clustering, we introduce a new performance metric called *clustering gain* (CG), which will allow us to quantify such improvement.

We consider the MSE in the first scenario, $\mathrm{MSE}_{\mathrm{min}}^{(1)} = \frac{\sigma_v^2(\sigma_\theta^2+\sigma_u^2)}{\sigma_\theta^2+\sigma_u^2+\sigma_v^2}$, as a baseline since it does not make use of the cluster structure and only uses the measurement vector $\mathbf{y}_{1:N}$, which is available to all estimators. The CG is therefore defined as

$$\mathrm{CG} \triangleq 10\log_{10}\left(\frac{\mathrm{MSE}_{\mathrm{min}}^{(1)}}{\mathrm{MSE}^{(\mathrm{clustering})}}\right), \tag{6.3}$$

where $\mathrm{MSE}^{(\mathrm{clustering})}$ is an MSE of the estimator that uses clustering, i.e., $\mathrm{MSE}_{\mathrm{MC}}^{(3)}$ or $\mathrm{MSE}_{\mathrm{min}}^{(4)}$. The clustering gain for the third scenario ($\mathrm{CG}^{(3)}$) has to be approximated, whereas for the fourth scenario, the clustering gain can be obtained in closed form by inserting (4.62) and (5.163) into (6.3), i.e.,

$$\mathrm{CG}^{(4)} \triangleq 10\log_{10}\left(\frac{\frac{\sigma_v^2(\sigma_\theta^2+\sigma_u^2)}{\sigma_\theta^2+\sigma_u^2+\sigma_v^2}}{\sum_{m=1}^{N}\left(\frac{\sigma_u^2\sigma_v^2}{\sigma_u^2+\sigma_v^2}+\frac{(\sigma_v^2)^2\sigma_{\theta*}^2}{(\sigma_u^2+\sigma_v^2)(\sigma_u^2+\sigma_v^2+m\sigma_{\theta*}^2)}\right)\frac{\alpha^{\overline{N-m}}(N-1)!}{(\alpha+1)^{\overline{N-1}}(N-m)!}}\right). \tag{6.4}$$

We note that this expression depends on the number of objects N; in particular, for $N=1$ we obtain $\mathrm{CG}^{(4)} = 0$.

6.3 Performance Evaluation

First, we compare the performance of the four estimators in terms of MSE. Subsequently, we will compare the performance in terms of MSE and CG achieved by the estimators in Scenarios 3 and 4 as opposed to modification of the estimator in Scenario 4, which uses a clustering algorithm followed by a pure estimation algorithm.

Table 3: MSE expressions for the four scenarios.

Scenario	MSE	Estimator uses clustering
1	$\frac{\sigma_v^2(\sigma_\theta^2+\sigma_u^2)}{\sigma_\theta^2+\sigma_u^2+\sigma_v^2}$ (see (4.62))	no
2	$\frac{\sigma_u^2\sigma_v^2}{\sigma_u^2+\sigma_v^2}$ (see (4.89))	no
3	$\approx \frac{1}{JN}\sum_{j=1}^{J}\sum_{n=1}^{N}\lVert\hat{\boldsymbol{x}}_{n,\mathrm{MC}}^{(3)}(\boldsymbol{y}_{1:N}^{(j)}) - \boldsymbol{x}_{n,j}\rVert_2^2$ (see (6.1))	yes
4	$\sum_{m=1}^{N}\left(\frac{\sigma_u^2\sigma_v^2}{\sigma_u^2+\sigma_v^2}+\frac{(\sigma_v^2)^2\sigma_{\theta*}^2}{(\sigma_u^2+\sigma_v^2)(\sigma_u^2+\sigma_v^2+m\sigma_{\theta*}^2)}\right)\frac{\alpha^{\overline{N-m}}(N-1)!}{(\alpha+1)^{\overline{N-1}}(N-m)!}$ (see (5.163))	yes

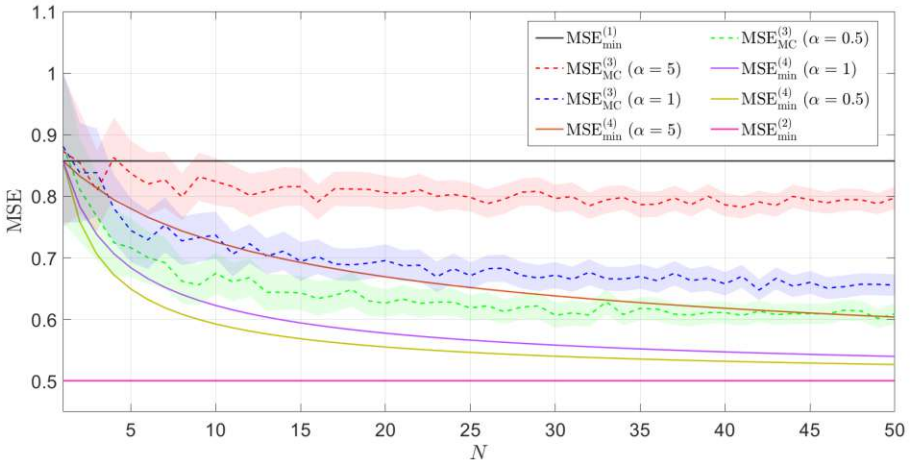
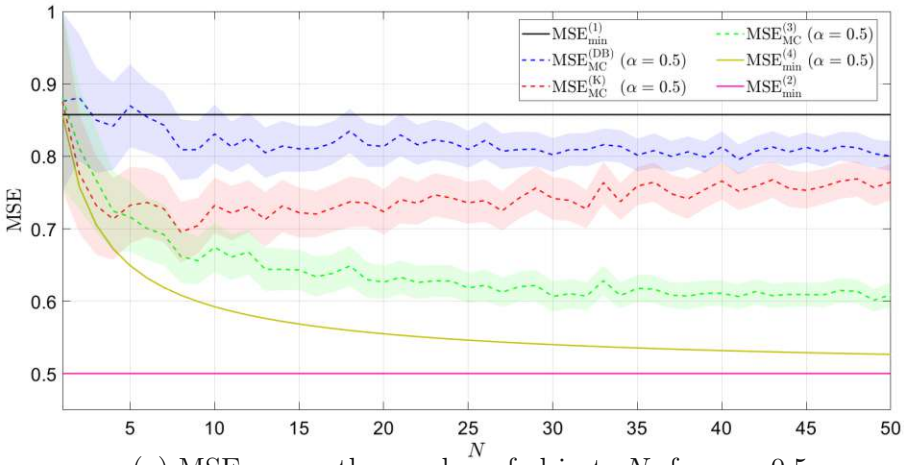


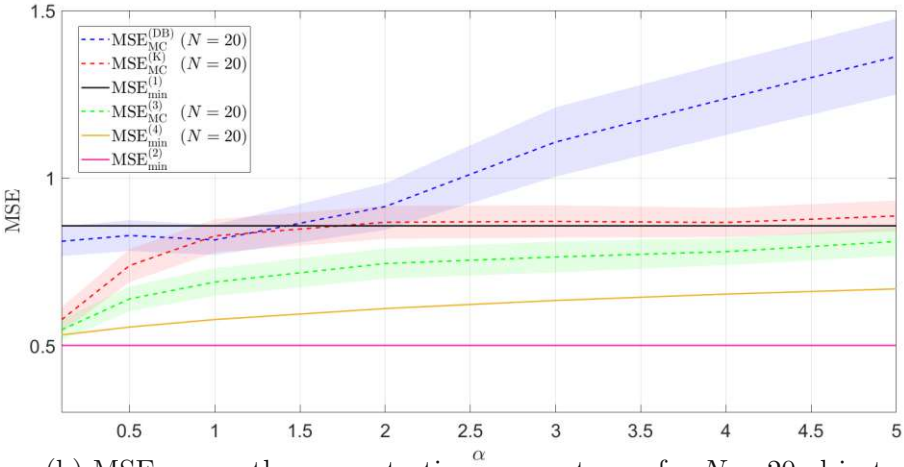
Figure 11: MSE of the four estimators as a function of the number of objects N. Closed-form expressions are shown using solid lines, whereas Monte Carlo simulation results ($\text{MSE}^{(3)}_{\text{MC}}$) are shown using dashed lines. The shaded regions indicate the empirical standard deviation of the simulated MSE.

6.3.1 MSE of the Four Scenarios

Figure 11 presents the MSE as a function of the number of objects N for all four scenarios. As expected, the MSE achieved by the first estimator, $\text{MSE}^{(1)}_{\text{min}}$ in (4.62), is higher than the MSE of any of the other estimators, as the first estimator makes no use of the cluster structure. On the other hand, the MSE achieved by the second estimator, $\text{MSE}^{(2)}_{\text{min}}$ in (4.89), is the lowest of all MSEs, since in Scenario 2 the hyperparameter $\boldsymbol{\theta}_n$ is known. Note also that, in line with (4.62) and (4.89), $\text{MSE}^{(1)}_{\text{min}}$ and $\text{MSE}^{(2)}_{\text{min}}$ do not depend on N. By contrast, $\text{MSE}^{(3)}_{\text{MC}}$ and $\text{MSE}^{(4)}_{\text{min}}$ depend on N and on the concentration parameter α. As the number of objects N increases, $\text{MSE}^{(3)}_{\text{MC}}$ and $\text{MSE}^{(4)}_{\text{min}}$ decrease. This can be explained by the fact that with more objects, more measurements associated with each cluster are available to the estimators and more objects share the same hyperparameter $\boldsymbol{\theta}^*$, which results in an improved estimation. For the small concentration parameter value $\alpha = 0.5$, the hyperparameters are more concentrated; in other words, there are fewer distinct clusters and more objects share the same hyperparameter $\boldsymbol{\theta}^*$. Here, we observe a strong decrease of the MSE with increasing N. On the contrary, for the large concentration parameter value $\alpha = 5$, there are more clusters and fewer objects share the same hyperparameter $\boldsymbol{\theta}^*$; here, the decrease of the MSE is less strong. Finally, $\text{MSE}^{(4)}_{\text{min}}$ is seen to be lower than $\text{MSE}^{(3)}_{\text{MC}}$. This is an expected result, since in the fourth scenario the cluster assignment variables $\mathbf{C}_{1:N}$ are considered to be known, and thus the estimator can use more measurements than in the



(a) MSE versus the number of objects N, for $\alpha = 0.5$.



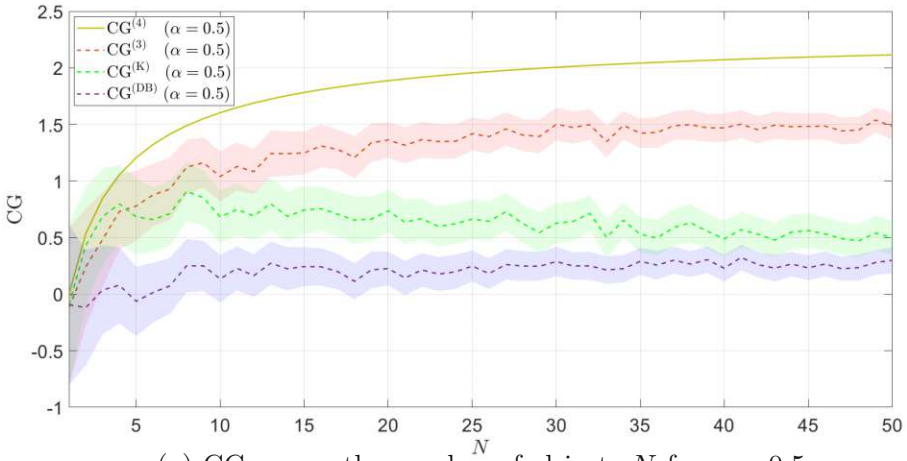
(b) MSE versus the concentration parameter α, for $N = 20$ objects.

Figure 12: MSE of estimators 1 through 4 as well as of two versions of estimator 4 using estimates of $\boldsymbol{C}_{1:N}$ ($\mathrm{MSE}_{\mathrm{MC}}^{(\mathrm{DB})}$ and $\mathrm{MSE}_{\mathrm{MC}}^{(\mathrm{K})}$). Closed-form expressions are shown using solid lines, whereas the Monte Carlo simulation results ($\mathrm{MSE}_{\mathrm{MC}}^{(3)}$, $\mathrm{MSE}_{\mathrm{MC}}^{(\mathrm{DB})}$, and $\mathrm{MSE}_{\mathrm{MC}}^{(\mathrm{K})}$) are shown using dashed lines. The shaded regions indicate the empirical standard deviation of the simulated MSE.

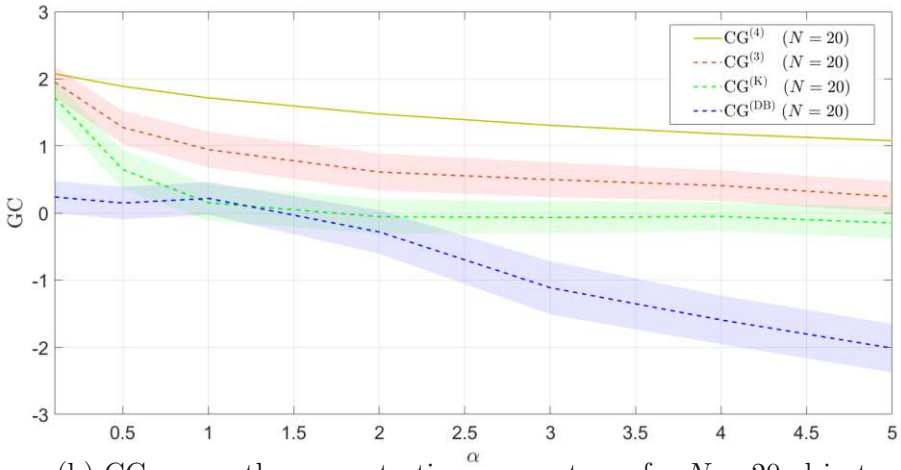
third scenario.

6.3.2 Comparison with Other Clustering Algorithms

The fourth estimator outperforms the third estimator because it exploits knowledge of the cluster assignment variables $\mathbf{C}_{1:N}$. Let us now consider a modification of the fourth estimator that uses an estimate $\hat{\boldsymbol{C}}_{1:N}$ of $\mathbf{C}_{1:N}$. The estimate $\hat{\boldsymbol{C}}_{1:N}$ is provided by two standard clustering algorithms, namely, the *DBSCAN* algorithm [16] and the *K-Means ++* algorithm [17]. The resulting MSEs, denoted as $\mathrm{MSE}_{\mathrm{MC}}^{(\mathrm{DB})}$ and $\mathrm{MSE}_{\mathrm{MC}}^{(\mathrm{K})}$, respectively, are shown in Figure 12 along with $\mathrm{MSE}_{\min}^{(1)}$, $\mathrm{MSE}_{\min}^{(2)}$, $\mathrm{MSE}_{\mathrm{MC}}^{(3)}$, and $\mathrm{MSE}_{\min}^{(4)}$.



(a) CG versus the number of objects N for $\alpha = 0.5$.



(b) CG versus the concentration parameter α for $N = 20$ objects.

Figure 13: CG of estimators 3 and 4 ($\mathrm{CG}^{(3)}$ and $\mathrm{CG}^{(4)}$) as well as of two versions of estimator 4 using estimates of $\boldsymbol{C}_{1:N}$ ($\mathrm{CG}^{(\mathrm{DB})}$ and $\mathrm{CG}^{(\mathrm{K})}$). The shaded regions indicate the empirical standard deviation.

We can observe in Figure 12a that $\mathrm{MSE}_{\mathrm{MC}}^{(3)}$ is lower than $\mathrm{MSE}_{\mathrm{MC}}^{(\mathrm{DB})}$ and (for $N \geq 5$) also $\mathrm{MSE}_{\mathrm{MC}}^{(\mathrm{K})}$, i.e., the third estimator outperforms both versions of the modified fourth estimator; furthermore, $\mathrm{MSE}_{\mathrm{MC}}^{(\mathrm{K})}$ is lower than $\mathrm{MSE}_{\mathrm{MC}}^{(\mathrm{DB})}$. The latter result can be explained by the fact that the K-Means ++ algorithm uses knowledge of the number of clusters, whereas the DBSCAN algorithm does not. Furthermore, Figure 12b shows that for concentration parameter $\alpha \geq 1.5$, $\mathrm{MSE}_{\mathrm{MC}}^{(\mathrm{DB})}$ is higher than $\mathrm{MSE}_{\mathrm{min}}^{(1)}$. This can be explained by the incorrect clustering performed by the DBSCAN algorithm, which for larger values of α tends to assign all samples into one cluster. Note that in Figure 12b, $N = 20$ is assumed; however, a similar performance was observed for both larger and smaller values of objects N.

In Figure 13a, we show the CG versus the number of objects N (see (6.3)). As in Figure 12a, the concentration parameter was chosen as $\alpha = 0.5$. As expected, the fourth

estimator achieves the largest CG, followed by the third estimator. It is also seen that the modified versions of the fourth estimator using the K-Means ++ and DBSCAN clustering algorithms achieve a considerably lower CG. Lastly, Figure 13b shows the GC versus the concentration parameter α. Similar to Figure 12b, we note that for approximately $\alpha \geq 1.5$, the $\mathrm{CG}^{(\mathrm{DB})}$ is negative due to incorrect clustering performed by DBSCAN algorithm.

7 Conclusion

In this Master's thesis, we investigated the application of DPMs in the Bayesian estimation framework, specifically exploring their use for estimation within a simple hierarchical Gaussian model. By considering a DP prior, we overcame the limitations of fixed-size models and thus avoided the need to pre-specify the number of clusters. Because the posterior distribution cannot be calculated in closed form, we used a Monte Carlo approximation of the MMSE estimator and derived a Gibbs sampling algorithm for our estimation problem. A notable distinction from prior works, such as [6] and [7], is that our model yields closed-form performance bounds. This feature facilitates the quantification of improvements in estimation performance due to the use of the DP prior in relation to the theoretically achievable performance.

We commenced with an introduction to the Bayesian framework of estimation and a discussion of key properties of the Gaussian distribution. Building on this foundation and on [21] and [6], we explored the basic theory of DPs and DPMs. In particular, we discussed certain distributions associated with a DP. Furthermore, we presented four procedures for generating samples from a DP, and we studied the clustering property intrinsic to DPs.

Shifting the focus to our Gaussian estimation problem, we introduced our general Gaussian model and associated independence assumptions. Within this general statistical framework, we considered four different scenarios. In Scenario 1, the hyperparameters are assumed to be i.i.d. and Gaussian distributed. In Scenario 2, they are still modeled as i.i.d. and Gaussian distributed but are now considered to be known. For both scenarios, we derived closed-form expressions for the MMSE estimator and the corresponding MMSE. Scenario 3 features a DP prior for the hyperparameters. Because the complexity of the DP prior does not allow for closed-form solutions, we developed two Gibbs sampling algorithms that provide Monte Carlo approximations of the MMSE estimator. In contrast to the estimators for Scenario 1 and 2, the estimator for Scenario 3 leverages the cluster structure induced by the DP prior. In Scenario 4, the hyperparameters are still distributed according to a DP prior but the object-cluster associations are now considered to be known. Here again, we derived closed-form expressions for the MMSE estimator and the corresponding MMSE.

The closed-form MMSEs of Scenarios 1, 2, and 4 provide lower and upper bounds on the MSE of Scenario 3. Through simulations, we demonstrated the effectiveness of the Gibbs sampler-based estimator in leveraging the cluster structure induced by the DP prior. Our

proposed Monte Carlo approximation of the MMSE estimator in Scenario 3 consistently achieves a lower MSE than the MMSE estimator in Scenario 1, which makes no use of the cluster structure; that is, in all simulation settings within Scenario 3, we always obtained a reduction in MSE due to clustering, or equivalently, a positive "clustering gain". Moreover, this clustering gain is only 0.5dB lower than the bound in Scenario 4, which uses knowledge of the object-cluster associations. Finally, our simulations showed that our approach of joint estimation and clustering outperforms the conventional approach of using a clustering algorithm (K-Means ++ or DBSCAN) followed by an estimation algorithm.

On the other hand, the computational complexity of the proposed Gibbs sampling algorithm is considerably higher than that of the other methods. This issue was addressed in [15] by using the CAVI algorithm instead of the Gibbs sampler to approximate the posterior distribution. For a small concentration parameter $\alpha = 0.5$, the CAVI algorithm proposed in [15] achieves a clustering gain that is only 16% lower than that achieved with our method, while the computational complexity is considerably smaller. However, for $\alpha = 5$, the clustering gain achieved with the CAVI algorithm is negative, which means that the algorithm performs worse than the MMSE estimator without clustering (Scenario 1), whereas our method consistently achieves a positive clustering gain for any α. Thus, our proposed method is most suitable for applications where the data is spread across a larger number of clusters (corresponding to a large concentration parameter α), and for applications that require a high accuracy of estimation even at the cost of a higher computational complexity.

The primary limitation of our approach lies in the computational complexity of the Gibbs sampler. This complexity can be reduced by adopting advanced sampling techniques such as the Hogwild Parallel Gaussian Gibbs Sampler [37] or the Fast Asynchronous MCMC Sampler [38]. Furthermore, our current model is limited in that it assumes that each object is associated with a single latent variable and thus cannot belong to more than one cluster. In certain applications, it would be advantageous to extend our model to accommodate objects belonging to multiple clusters defined by different features. Such an extension, described in [39], is constituted by a distribution that enables the construction of probabilistic models for objects with an infinite number of binary latent variables, and that can be seamlessly combined with priors on the latent variables' values. This distribution is based on a Bayesian non-parametric model known as the Indian buffet process, which is somewhat analogous to the Chinese restaurant process considered in this thesis.

A Proofs

A.1 Proof of (3.5)

We prove that the recursive construction of weights sequence $(\mathsf{Q}_l)_{l=1}^{\infty}$ as presented in (3.5) is equivalent to the definition in (3.3). Thus, in this proof, we can take (3.5) as being true, but not (3.3). For $l = 1$, (3.5) states that

$$\mathsf{Q}_1 = \mathsf{V}_1 \tag{A.1}$$

This is also stated in (3.3), and thus (3.5) is equivalent to (3.3) for $l = 1$.

For $l \geq 2$, we use mathematical induction. We claim that the expressions stated in (3.3) and (3.5) for $l \geq 2$ are equal, i.e.,

$$\mathsf{V}_l \prod_{l'=1}^{l-1} (1 - \mathsf{V}_{l'}) = \mathsf{V}_l \left(1 - \sum_{l'=1}^{l-1} \mathsf{Q}_{l'} \right) \tag{A.2}$$

or, equivalently,

$$\prod_{l'=1}^{l-1} (1 - \mathsf{V}_{l'}) = 1 - \sum_{l'=1}^{l-1} \mathsf{Q}_{l'}. \tag{A.3}$$

For the base case $l = 2$, (A.3) gives

$$1 - \mathsf{V}_1 = 1 - \mathsf{Q}_1, \tag{A.4}$$

which is clearly true because of (A.1). In the induction step ($l \to l + 1$), we assume that (A.3) is true and wish to show that it remains true when l is replaced by $l + 1$, i.e., we have to show

$$\prod_{l'=1}^{l} (1 - \mathsf{V}_{l'}) = 1 - \sum_{l'=1}^{l} \mathsf{Q}_{l'}. \tag{A.5}$$

The left hand side of (A.5) becomes

$$\prod_{l'=1}^{l} (1 - \mathsf{V}_{l'}) = (1 - \mathsf{V}_l) \prod_{l'=1}^{l-1} (1 - \mathsf{V}_{l'}) \,. \tag{A.6}$$

Using (A.3), we obtain further

$$\begin{aligned}\prod_{l'=1}^{l}(1-\mathsf{V}_{l'}) &= (1-\mathsf{V}_l)\left(1-\sum_{l'=1}^{l-1}\mathsf{Q}_{l'}\right)\\ &= 1-\mathsf{V}_l-\sum_{l'=1}^{l-1}\mathsf{Q}_{l'}+\mathsf{V}_l\sum_{l'=1}^{l-1}\mathsf{Q}_{l'}\\ &= 1-\mathsf{V}_l\left(1-\sum_{l'=1}^{l-1}\mathsf{Q}_{l'}\right)-\sum_{l'=1}^{l-1}\mathsf{Q}_{l'}\end{aligned} \tag{A.7}$$

Using (3.5), this finally becomes

$$\begin{aligned}\prod_{l'=1}^{l}(1-\mathsf{V}_{l'}) &= 1-\mathsf{Q}_l-\sum_{l'=1}^{l-1}\mathsf{Q}_{l'}\\ &= 1-\sum_{l'=1}^{l}\mathsf{Q}_{l'},\end{aligned} \tag{A.8}$$

which is seen to be equal to the right-hand side of (A.5).

A.2 Proof of (5.160)

Let us consider N samples from a DP with concentration parameter α. We wish to derive expression (5.160) for the probability of the cluster size $\mathsf{m}_l(N)$, i.e.,

$$p_{\mathsf{m}_l(N)}(m) = \mathbb{P}(\mathsf{m}_l(N)=m), \quad \text{for} \quad 1 \leq m \leq N. \tag{A.9}$$

We recall that $\boldsymbol{\theta}_n$ can equivalently be written using the cluster assignment variable $\tilde{\mathsf{C}}_n$ as $\boldsymbol{\theta}_n = \boldsymbol{\vartheta}^*_{\tilde{\mathsf{C}}_n}$ (see (3.67)). Let $s = \tilde{\mathsf{C}}_n$, i.e., the sample $\boldsymbol{\theta}_n$ belongs to the s^{th} cluster. The number of samples $\boldsymbol{\theta}_n$ belonging to the s^{th} cluster, is denoted as $\tilde{\mathsf{m}}_s(N)$, and given by

$$\tilde{m}_s(N) = \sum_{n=1}^{N}\mathbb{1}(\boldsymbol{\theta}_n = \boldsymbol{\vartheta}^*_s), \quad s=1,\ldots,S(N), \tag{A.10}$$

or equivalently, using the cluster assignment variables $\tilde{\mathsf{C}}_n$, we obtain

$$\tilde{m}_s(N) = \sum_{n=1}^{N}\mathbb{1}(\tilde{\mathsf{C}}_n = s), \quad s=1,\ldots,S(N). \tag{A.11}$$

Since the DP samples $(\boldsymbol{\theta}_n)_{n=1}^N$ are exchangeable, for simplicity, we assume $s = 1$, i.e., we consider the first cluster, according to the empirical ordering of the observed clusters (see (3.12)). Thus we want to derive an expression for the probability size $\mathbb{P}(\tilde{\mathsf{m}}_1(N) = m)$.

A.2.1 Recursive Construction

Consider once more the CRP analogy to seating customers in a restaurant, presented in Section 3.3.3. We recall that each customer n is assigned to a table (cluster), i.e., $\tilde{C}_n = s$. The conditional pmf of the $\tilde{C}_n$ is given in the expression (3.64).

We now examine the n^{th} customer entering the restaurant, with $n-1$ customers already seated and $S(n-1)$ tables occupied. Using (3.64) we conclude that the customer sits at a new table $S(n-1)+1$ with probability

$$\mathbb{P}\big(\tilde{\mathsf{C}}_n = S(n-1)+1 \,|\, \tilde{\mathbf{C}}_{1:n-1} = \tilde{\boldsymbol{C}}_{1:n-1}\big) = \frac{\alpha}{\alpha+n-1}, \tag{A.12}$$

or at an already occupied table, $s = 1, \ldots, S(n-1)$, with probability proportional to the number of customers already seated at the table,

$$\mathbb{P}\big(\tilde{\mathsf{C}}_n = s \,|\, \tilde{\mathbf{C}}_{1:n-1} = \tilde{\boldsymbol{C}}_{1:n-1}\big) = \frac{\sum_{n'=1}^{n-1} \mathbb{1}(\tilde{C}_{n'} = s)}{\alpha+n-1} = \frac{\tilde{m}_s(n-1)}{\alpha+n-1}. \tag{A.13}$$

We note that the probability of the n^{th} customer sitting at a new table in (A.12) only depends on n and concentration parameter α, and is independent of $\tilde{\mathbf{C}}_{1:n-1}$. Moreover, we note that the right-hand side of (A.13) only depends on the table size $\tilde{\mathsf{m}}_s(n-1)$ and α, i.e.,

$$\mathbb{P}\big(\tilde{\mathsf{C}}_n = s \,|\, \tilde{\mathsf{m}}_s(n-1) = \tilde{m}_s(n-1)\big) = \frac{\tilde{m}_s(n-1)}{\alpha+n-1}. \tag{A.14}$$

We can now calculate the joint probability of $\tilde{\mathsf{m}}_1(n)$, and $\tilde{\mathsf{C}}_n$. If the customer sits at the first table, we set $\tilde{\mathsf{C}}_n = 1$ and the number of customers seated at the first table ($s = 1$) increases by one, i.e., $\tilde{\mathsf{m}}_1(n) = \tilde{\mathsf{m}}_1(n-1) + 1$. This probability is given as

$$\mathbb{P}\big(\tilde{\mathsf{m}}_1(n) = m, \tilde{\mathsf{C}}_n = 1\big) = \mathbb{P}\big(\tilde{\mathsf{C}}_n = 1 \,|\, \tilde{\mathsf{m}}_1(n-1) = m-1\big)\mathbb{P}\big(\tilde{\mathsf{m}}_1(n-1) = m-1\big). \tag{A.15}$$

On the other hand, if the n^{th} customer sits at another table ($s \neq 1$), the number of customers seated at the first table remains $\tilde{\mathsf{m}}_1(n) = \tilde{\mathsf{m}}_1(n-1)$ and we set $\tilde{\mathsf{C}}_n \neq 1$. This conditional

probability is given as

$$\mathbb{P}\big(\tilde{\mathsf{m}}_1(n) = m, \tilde{\mathsf{C}}_n \neq 1\big) = \mathbb{P}\big(\tilde{\mathsf{C}}_n \neq 1 \,|\, \tilde{\mathsf{m}}_1(n-1) = m\big)\mathbb{P}\big(\tilde{\mathsf{m}}_1(n-1) = m\big). \tag{A.16}$$

This means we can write a general recursion for the table size $\mathbb{P}(\tilde{\mathsf{m}}_1(n) = m)$, by using the total probability theorem, i.e.,

$$\begin{aligned}\mathbb{P}(\tilde{\mathsf{m}}_1(n) = m) = \mathbb{P}\big(\tilde{\mathsf{C}}_n = 1 \,|\, \tilde{\mathsf{m}}_1(n-1) = m-1\big)\mathbb{P}\big(\tilde{\mathsf{m}}_1(n-1) = m-1\big) \\ + \mathbb{P}\big(\tilde{\mathsf{C}}_n \neq 1 \,|\, \tilde{\mathsf{m}}_1(n-1) = m\big)\mathbb{P}\big(\tilde{\mathsf{m}}_1(n-1) = m\big).\end{aligned} \tag{A.17}$$

Using (A.14) and $\mathbb{P}\big(\tilde{\mathsf{C}}_n \neq 1 \,|\, \tilde{\mathsf{m}}_1(n-1) = m\big) = \big(1 - \frac{m}{\alpha+n-1} = \frac{\alpha+n-1-m}{\alpha+n-1}\big)$ in (A.17), we obtain a recursive probability

$$\mathbb{P}(\tilde{\mathsf{m}}_1(N) = m) = \frac{m-1}{\alpha+n-1}\mathbb{P}\big(\tilde{\mathsf{m}}_1(n-1) = m-1\big) + \frac{\alpha+n-1-m}{\alpha+n-1}\mathbb{P}\big(\tilde{\mathsf{m}}_1(n-1) = m\big). \tag{A.18}$$

We also note that $1 \leq m \leq n$, therefore we need to consider two special cases of (A.18), specifically $m = 1$ and $m = n$. For $m = 1$ we obtain

$$\mathbb{P}\big(\tilde{\mathsf{m}}_1(n) = 1\big) = \frac{\alpha+n-2}{\alpha+n-1}\mathbb{P}\big(\tilde{\mathsf{m}}_1(n-1) = 1\big), \tag{A.19}$$

and for $m = n$,

$$\mathbb{P}\big(\tilde{\mathsf{m}}_1(n) = n\big) = \frac{n-1}{\alpha+n-1}\mathbb{P}\big(\tilde{\mathsf{m}}_1(n-1) = n-1\big), \tag{A.20}$$

since $\mathbb{P}\big(\tilde{\mathsf{m}}_1(n-1) = n\big) = 0$ due to the condition $1 \leq m \leq n$.

A.2.2 Proof by Induction

We claim that the probability $\mathbb{P}\big(\mathsf{m}_l(N) = m\big)$ is given by

$$\mathbb{P}\big(\mathsf{m}_l(N) = m\big) = \frac{(N-1)!\,\alpha^{\overline{N-m}}}{(N-m)!\,(\alpha+1)^{\overline{N-1}}}, \quad \text{for} \quad 1 \leq m \leq N. \tag{A.21}$$

We prove this claim using mathematical induction. For $N = 1$ we also obtain $m = 1$, since $1 \leq m \leq N$, therefore

$$\mathbb{P}\big(\mathsf{m}_l(1) = 1\big) = \frac{0!\,\alpha^{\overline{0}}}{0!\,(\alpha+1)^{\overline{0}}} = 1. \tag{A.22}$$

This is trivial, since for one sample $N = 1$, there exists only one cluster and the sample must belong to this cluster. In the induction step $N \to N+1$, we have to prove that from the induction assumption (A.21), it follows that

$$\mathbb{P}\big(\mathsf{m}_l(N+1) = m\big) = \frac{N!\,\alpha^{\overline{N+1-m}}}{(N+1-m)!\,(\alpha+1)^{\overline{N}}} \quad \text{for} \quad 1 \leq m \leq N+1. \tag{A.23}$$

Since the cluster sizes $\tilde{\mathsf{m}}_s(n)$ are exchangeable and $\tilde{\mathsf{m}}_s(n)$ is a permuted version of $\mathsf{m}_l(n)$, the recursive probabilities (A.18) and (A.20) are also valid for $\mathsf{m}_l(N)$. Using (A.18) (with $\mathsf{m}_l(N)$ instead of $\tilde{\mathsf{m}}_1(n)$), the left hand side of (A.23) becomes

$$\mathbb{P}\big(\mathsf{m}_l(N+1) = m\big) = \frac{\alpha+N-m}{\alpha+N}\,\mathbb{P}\big(\mathsf{m}_l(N) = m\big) + \frac{m-1}{\alpha+N}\,\mathbb{P}\big(\mathsf{m}_l(N) = m-1\big), \tag{A.24}$$

for $1 \leq m \leq N$ and

$$\mathbb{P}\big(\mathsf{m}_l(N+1) = m\big) = \frac{m-1}{\alpha+N}\mathbb{P}\big(\tilde{\mathsf{m}}_1(N) = m-1\big), \tag{A.25}$$

for $m = N+1$. Substituting (A.21) into (A.24) yields

$$\mathbb{P}\big(\mathsf{m}_l(N+1) = m\big) = \frac{\alpha+N-m}{\alpha+N}\frac{(N-1)!\,\alpha^{\overline{N-m}}}{(N-m)!\,(\alpha+1)^{\overline{N-1}}} + \frac{m-1}{\alpha+N}\frac{(N-1)!\,\alpha^{\overline{N+1-m}}}{(N+1-m)!\,(\alpha+1)^{\overline{N-1}}}. \tag{A.26}$$

Now $\alpha^{\overline{N-m}}(\alpha+N-m) = \alpha^{\overline{N-m+1}}$ and $(\alpha+1)^{\overline{N-1}}(\alpha+N) = (\alpha+1)^{\overline{N}}$, so that we further obtain

$$\begin{aligned}
\mathbb{P}\big(\mathsf{m}_l(N+1) = m\big) &= \frac{(N-1)!\,\alpha^{\overline{N+1-m}}}{(N-m)!\,(\alpha+1)^{\overline{N}}} + \frac{(N-1)!\,\alpha^{\overline{N+1-m}}}{(N+1-m)!\,(\alpha+1)^{\overline{N}}}(m-1) \\
&= \frac{(N-1)!\,\alpha^{\overline{N+1-m}}}{(N-m+1)!\,(\alpha+1)^{\overline{N}}}(N+1-m) + \frac{(N-1)!\,\alpha^{\overline{N+1-m}}}{(N+1-m)!\,(\alpha+1)^{\overline{N}}}(m-1) \\
&= \frac{(N-1)!\,\alpha^{\overline{N+1-m}}}{(N-m+1)!\,(\alpha+1)^{\overline{N}}}(N+\cancel{1-m+m-1})
\end{aligned}$$

$$= \frac{N!\,\alpha^{\overline{N+1-m}}}{(N-m+1)!\,(\alpha+1)^{\overline{N}}}, \tag{A.27}$$

which is seen to be equal to the right-hand side of (A.23).

Finally, for the case $m = N+1$, (A.23) becomes

$$\mathbb{P}\big(\mathsf{m}_l(N+1) = N+1\big) = \frac{N!\,\alpha^{\overline{0}}}{(0)!\,(\alpha+1)^{\overline{N}}} = \frac{N!}{(\alpha+1)^{\overline{N}}}. \tag{A.28}$$

Substituting (A.21) and $m = N+1$ into (A.25) gives

$$\begin{aligned}\mathbb{P}\big(\mathsf{m}_l(N+1) = N+1\big) &= \frac{m-1}{\alpha+N}\frac{(N-1)!\,\alpha^{\overline{N+1-m}}}{(N+1-m)!\,(\alpha+1)^{\overline{N-1}}} \\ &= \frac{N}{\alpha+N}\frac{(N-1)!\,\alpha^{\overline{0}}}{(0)!\,(\alpha+1)^{\overline{N-1}}} \\ &= \frac{N!}{(\alpha+1)^{\overline{N}}},\end{aligned} \tag{A.29}$$

which is seen to be equal to (A.28).

A.2.3 Further Considerations

To prepare the ground for a general case $n \in \mathbb{N}$, we first consider the cases $n = 1, 2, 3, 4$.

$\underline{n = 1}$

Consider $n = 1$, i.e., a customer is seated in an empty restaurant and therefore occupies one table. Evidently, then, $\tilde{\mathsf{m}}_1(1) = 1$ and

$$\mathbb{P}\big(\tilde{\mathsf{m}}_1(1) = 1\big) = 1.$$

$\underline{n = 2}$

Let a new customer enter the restaurant so that $n = 2$. The new customer either sits at a new table and the first table remains occupied by one customer ($\tilde{\mathsf{m}}_1(2) = 1$), i.e.,

$$\mathbb{P}\big(\tilde{\mathsf{m}}_1(2) = 1\big) = \frac{\alpha}{\alpha+1}, \tag{A.30}$$

or the new customer joins the first table, i.e., using (A.18) with $(\tilde{\mathsf{m}}_1(2-1) = 1)$ we have

$$\mathbb{P}\big(\tilde{\mathsf{m}}_1(2) = 2\big) = \frac{1}{\alpha+1}, \tag{A.31}$$

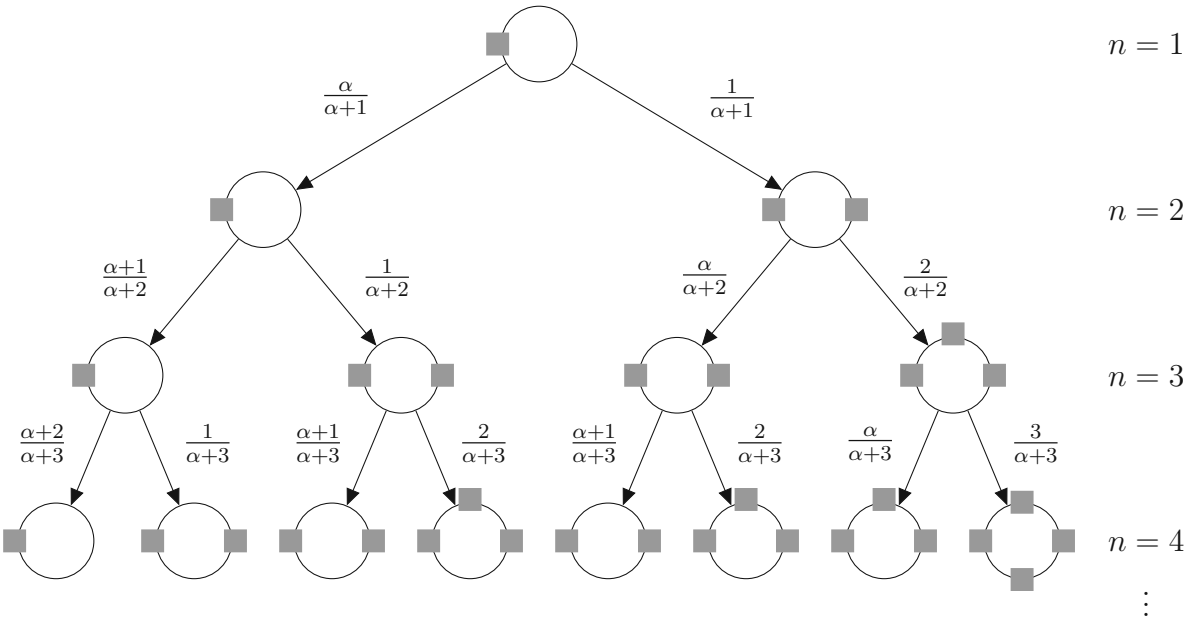


Figure 14: Probability tree diagram describing the events and probabilities related to the number of customers $\mathsf{m}_1(n)$ at the first table ($s = 1$). Each customer sitting at the first table is represented by a gray square.

using $\mathbb{P}\big(\tilde{\mathsf{m}}_1(1) = 1\big) = 1$ and $\mathbb{P}\big(\tilde{\mathsf{m}}_1(1) = 2\big) = 0$. The probability of the second customer $n = 2$ joining the first table is given by $\mathbb{P}\big(\tilde{\mathsf{C}}_2 = 1 \,|\, \tilde{\mathsf{m}}_1(1) = 1\big) = \frac{1}{\alpha+1}$, whereas the probability of not joining is given by $\mathbb{P}\big(\tilde{\mathsf{C}}_2 \neq 1 \,|\, \tilde{\mathsf{m}}_1(1) = 1\big) = \frac{\alpha}{\alpha+1}$. These exclusive events are depicted in the probability tree diagram in Figure 14.

$\underline{n = 3}$

For $n = 3$, the first table can remain occupied by one customer, which occurs when the third customer does not sit down at the first table, i.e., $\mathbb{P}\big(\tilde{\mathsf{C}}_3 \neq 1 \,|\, \tilde{\mathsf{m}}_1(2) = 1\big)$ and only one customer is seated at the first table, i.e., $\mathbb{P}\big(\tilde{\mathsf{m}}_1(2) = 1\big)$, as can be seen on the very left-hand side of the Figure 14. Using (A.31), the probability $\mathbb{P}\big(\tilde{\mathsf{m}}_1(3) = 1\big)$ is therefore given by

$$\begin{aligned}\mathbb{P}\big(\tilde{\mathsf{m}}_1(3) = 1\big) &= \mathbb{P}\big(\tilde{\mathsf{C}}_3 \neq 1 \,|\, \tilde{\mathsf{m}}_1(2) = 1\big)\mathbb{P}\big(\tilde{\mathsf{m}}_1(2) = 1\big) \\ &= \frac{\alpha+1}{\alpha+2}\frac{\alpha}{\alpha+1} \\ &= \frac{\alpha}{\alpha+2}.\end{aligned} \tag{A.32}$$

Secondly, the first table can be occupied by two customers $\tilde{\mathsf{m}}_1(3) = 2$. This happens when the third customer joins the first table with probability equal to $\mathbb{P}\big(\tilde{\mathsf{C}}_3 = 1 \,|\, \tilde{\mathsf{m}}_1(2) = 1\big)$ and the first table is occupied by only one customer with probability $\mathbb{P}\big(\tilde{\mathsf{m}}_1(2) = 1\big)$, or the

first table is already occupied by two customers $\mathbb{P}\big(\tilde{\mathsf{m}}_1(2) = 2\big)$ and the third customer one does not join the first table $\mathbb{P}\big(\tilde{\mathsf{C}}_3 \neq 1 \,|\, \tilde{\mathsf{m}}_1(2) = 2\big)$. This means that

$$\mathbb{P}\big(\tilde{\mathsf{m}}_1(3) = 2\big) = \mathbb{P}\big(\tilde{\mathsf{C}}_3 = 1 \,|\, \tilde{\mathsf{m}}_1(2) = 1\big)\mathbb{P}\big(\tilde{\mathsf{m}}_1(2) = 1\big) + \mathbb{P}\big(\tilde{\mathsf{C}}_3 \neq 1 \,|\, \tilde{\mathsf{m}}_1(2) = 2\big)\mathbb{P}\big(\tilde{\mathsf{m}}_1(2) = 2\big) \tag{A.33}$$

As can be read from Figure 14, the probability of this event is given by

$$\mathbb{P}\big(\tilde{\mathsf{m}}_1(3) = 2\big) = \frac{1}{\alpha+2}\frac{\alpha}{\alpha+1} + \frac{\alpha}{\alpha+2}\frac{1}{\alpha+1} = \frac{2\alpha}{(\alpha+2)(\alpha+1)}, \tag{A.34}$$

where we used (A.30) and (A.31).

Lastly, all three customers can be seated at the first table, i.e., $\tilde{\mathsf{m}}_1(3) = 3$. This event only occurs when the third customer joins the first table $\mathbb{P}\big(\tilde{\mathsf{C}}_3 = 1 \,|\, \tilde{\mathsf{m}}_1(2) = 2\big)$ that is already occupied by two customers $\mathbb{P}\big(\tilde{\mathsf{m}}_1(2) = 2\big)$, i.e., using (A.31) we obtain

$$\begin{aligned}\mathbb{P}\big(\tilde{\mathsf{m}}_1(3) = 3\big) &= \mathbb{P}\big(\tilde{\mathsf{C}}_3 = 1 \,|\, \tilde{\mathsf{m}}_1(2) = 2\big)\mathbb{P}\big(\tilde{\mathsf{m}}_1(2) = 2\big) \\ &= \frac{2}{\alpha+2}\frac{1}{\alpha+1} \\ &= \frac{2}{(\alpha+2)(\alpha+1)}.\end{aligned} \tag{A.35}$$

$\underline{n = 4}$

From Figure 14 we can see that for $n = 4$, there is only one path leading to $\tilde{\mathsf{m}}_1(4) = 1$ and $\tilde{\mathsf{m}}_1(4) = 4$; however, there are three paths leading to $\tilde{\mathsf{m}}_1(4) = 2$ and $\tilde{\mathsf{m}}_1(4) = 3$. In order not to calculate each path separately, we can use the general recursion (A.18), together with the previous results for $n = 3$ to calculate the probabilities for $n = 4$. First, for $\mathbb{P}\big(\tilde{\mathsf{m}}_1(4) = 1\big)$, we insert (A.32) into (A.19) (for $n = 4$ and $m = 1$) and obtain

$$\begin{aligned}\mathbb{P}\big(\tilde{\mathsf{m}}_1(4) = 1\big) &= \frac{\alpha+2}{\alpha+3}\mathbb{P}\big(\tilde{\mathsf{m}}_1(3) = 1\big) \\ &= \frac{\alpha+2}{\alpha+3}\frac{\alpha+1}{\alpha+2}\frac{\alpha}{\alpha+1} \\ &= \frac{\alpha}{\alpha+3}.\end{aligned} \tag{A.36}$$

Similarly, in order to obtain an expression for $\mathbb{P}(\tilde{\mathsf{m}}_1(4) = 2)$, we insert (A.32) and (A.34)

into (A.18) (for $N = 4$ and $m = 2$), i.e.,

$$\begin{aligned}\mathbb{P}\big(\tilde{\mathsf{m}}_1(4) = 2\big) &= \frac{\alpha+1}{\alpha+3}\,\mathbb{P}(\tilde{\mathsf{m}}_1(3) = 2) + \frac{1}{\alpha+3}\,\mathbb{P}\big(\tilde{\mathsf{m}}_1(3) = 1\big)\\ &= \frac{\alpha+1}{\alpha+3}\frac{2\alpha}{(\alpha+2)(\alpha+1)} + \frac{1}{\alpha+3}\frac{\alpha}{\alpha+2}\\ &= \frac{3\alpha(\alpha+1)}{(\alpha+3)(\alpha+2)(\alpha+1)},\end{aligned} \tag{A.37}$$

and for $\mathbb{P}\big(\tilde{\mathsf{m}}_1(4) = 3\big)$, we insert (A.34) and (A.35) into (A.18) (for $n = 4$ and $m = 3$), i.e.,

$$\begin{aligned}\mathbb{P}\big(\tilde{\mathsf{m}}_1(4) = 3\big) &= \frac{\alpha}{\alpha+3}\mathbb{P}\big(\tilde{\mathsf{m}}_1(3) = 3\big) + \frac{2}{\alpha+3}\mathbb{P}\big(\tilde{\mathsf{m}}_1(3) = 2\big)\\ &= \frac{\alpha}{\alpha+3}\frac{2}{(\alpha+2)(\alpha+1)} + \frac{2}{\alpha+3}\frac{2\alpha}{(\alpha+2)(\alpha+1)}\\ &= \frac{6\alpha}{(\alpha+3)(\alpha+2)(\alpha+1)}.\end{aligned} \tag{A.38}$$

Lastly, to obtain $\mathbb{P}\big(\tilde{\mathsf{m}}_1(4) = 4\big)$, we insert (A.35) into (A.20) (for $n = 4$ and $m = 4$), i.e.,

$$\begin{aligned}\mathbb{P}\big(\tilde{\mathsf{m}}_1(4) = 4\big) &= \frac{3}{\alpha+3}\mathbb{P}\big(\tilde{\mathsf{m}}_1(3) = 3\big)\\ &= \frac{3}{\alpha+3}\frac{2}{\alpha+2}\frac{1}{\alpha+1}\\ &= \frac{6}{(\alpha+3)(\alpha+2)(\alpha+1)}.\end{aligned} \tag{A.39}$$

General $n \in \mathbb{N}$

We now consider $n \in \mathbb{N}$ customer entering the restaurant. From Figure 14 we conclude that only the path on the left edge leads to $\tilde{\mathsf{m}}_1(n) = 1$. This means that this event occurs only if no other customer joins the first table, i.e.,

$$\begin{aligned}\mathbb{P}\big(\tilde{\mathsf{m}}_1(n) = 1\big) = \mathbb{P}\big(\tilde{\mathsf{C}}_2 \neq 1 \,|\, \tilde{\mathsf{m}}_1(1) = 1\big)\mathbb{P}\big(\tilde{\mathsf{m}}_1(1) = 1\big)\mathbb{P}\big(\tilde{\mathsf{C}}_3 \neq 1 \,|\, \tilde{\mathsf{m}}_1(2) = 1\big)\mathbb{P}\big(\tilde{\mathsf{m}}_1(2) = 1\big)\cdots\\ \times\, \mathbb{P}\big(\tilde{\mathsf{C}}_n \neq 1 \,|\, \tilde{\mathsf{m}}_1(n-1) = 1\big)\mathbb{P}\big(\tilde{\mathsf{m}}_1(n-1) = 1\big),\end{aligned} \tag{A.40}$$

which as can be seen on the left-hand side of Figure 14 is given by

$$\mathbb{P}\big(\tilde{\mathsf{m}}_1(n) = 1\big) = \frac{\alpha}{\alpha+1}\frac{\alpha+1}{\alpha+2}\cdots\frac{\alpha+n-2}{\alpha+n-1} = \frac{\alpha^{\overline{n-1}}}{(\alpha+1)^{\overline{n-1}}}, \tag{A.41}$$

with $\alpha^{\overline{m}} = \alpha(\alpha+1)\cdots(\alpha+m-1)$ ($\alpha^{\overline{0}} = 1$ and $\alpha^{\overline{1}} = \alpha$), denoting the rising factorial, also

called the Pochhammer symbol. On the other hand, the probability $\mathbb{P}\big(\tilde{\mathsf{m}}_1(n) = n\big)$ of the first table being occupied by n customers, i.e., all customers that entered the restaurant are seated at the first table is given by

$$\begin{aligned}\mathbb{P}\big(\tilde{\mathsf{m}}_1(n) = \big) = \mathbb{P}\big(\tilde{\mathsf{C}}_2 = 1 \,|\, \tilde{\mathsf{m}}_1(1) = 1\big)\mathbb{P}\big(\tilde{\mathsf{m}}_1(1) = 1\big)\mathbb{P}\big(\tilde{\mathsf{C}}_3 = 1 \,|\, \tilde{\mathsf{m}}_1(2) = 2\big)\mathbb{P}\big(\tilde{\mathsf{m}}_1(2) = 2\big)\cdots \\ \times \mathbb{P}\big(\tilde{\mathsf{C}}_n = 1 \,|\, \tilde{\mathsf{m}}_1(n-1) = n-1\big)\mathbb{P}\big(\tilde{\mathsf{m}}_1(n-1) = n-1\big),\end{aligned} \tag{A.42}$$

which as can be seen on the right-hand side of Figure 14,i.e.,

$$\mathbb{P}\big(\tilde{\mathsf{m}}_1(n) = n\big) = \frac{1}{\alpha+1}\frac{2}{\alpha+2}\cdots\frac{n-1}{\alpha+n-1} = \frac{(n-1)!}{(\alpha+1)^{\overline{n-1}}}. \tag{A.43}$$

We can now formulate a general expression for $\mathbb{P}\big(\tilde{\mathsf{m}}_1(n) = m\big)$ for $1 \leq m \leq n$, $m \in \mathbb{N}$. From Figure 14 we note that there are $\binom{n-1}{m-1}$ possible paths leading to event $\tilde{\mathsf{m}}_1(n) = m$. Since the DP samples $(\boldsymbol{\theta}_{n'})_{n'=1}^{n}$ are exchangeable, the order in which the customers sit down at the first table does not affect the probability $\mathbb{P}\big(\tilde{\mathsf{m}}_1(n) = m\big)$ and each of the paths leading to the event $\tilde{\mathsf{m}}_1(n) = m$ is equally likely. Let $\mathcal{A}$ denote a scenario, where m customers join the first table, and subsequently, $n-m$ customers join another table. The probability of this scenario is given by

$$\begin{aligned}\mathbb{P}\big(\tilde{\mathsf{m}}_1(n) = m \,|\, \mathcal{A}\big) &= \mathbb{P}\big(\tilde{\mathsf{C}}_2 = 1 \,|\, \tilde{\mathsf{m}}_1(1) = 1\big)\mathbb{P}\big(\tilde{\mathsf{m}}_1(1) = 1\big)\cdots\mathbb{P}\big(\tilde{\mathsf{C}}_m = 1 \,|\, \tilde{\mathsf{m}}_1(m-1) = m-1\big) \\ &\quad\times \mathbb{P}\big(\tilde{\mathsf{m}}_1(m-1) = m-1\big)\mathbb{P}\big(\tilde{\mathsf{C}}_{m+1} \neq 1 \,|\, \tilde{\mathsf{m}}_1(m) = m\big)\mathbb{P}\big(\tilde{\mathsf{m}}_1(m) = m\big)\cdots \\ &\quad\times \mathbb{P}\big(\tilde{\mathsf{C}}_n \neq 1 \,|\, \tilde{\mathsf{m}}_1(n-1) = m\big)\mathbb{P}\big(\tilde{\mathsf{m}}_1(n-1) = m\big) \\ &= \frac{1}{\alpha+1}\frac{2}{\alpha+2}\cdots\frac{m-2}{\alpha+m-2}\frac{m-1}{\alpha+m-1}\frac{\alpha}{\alpha+m}\frac{\alpha+1}{\alpha+m+1}\cdots \\ &\quad\times \frac{\alpha+n-m-1}{\alpha+n-1} \\ &= \frac{\alpha^{\overline{n-m}}(m-1)!}{(\alpha+1)^{\overline{n-1}}}.\end{aligned} \tag{A.44}$$

The probability $\mathbb{P}\big(\tilde{\mathsf{m}}_1(n) = m\big)$ is finally obtained by considering the number of paths leading to the event $\tilde{\mathsf{m}}_1(n) = m$, given by $\binom{n-1}{m-1}$ and the probability of one path, given in (A.44), i.e.,

$$\mathbb{P}\big(\tilde{\mathsf{m}}_1(n) = m\big) = \binom{n-1}{m-1}\frac{\alpha^{\overline{n-m}}(m-1)!}{(\alpha+1)^{\overline{n-1}}}$$

$$= \frac{(n-1)!}{\cancel{(m-1)!}(n-m)!} \frac{\alpha^{\overline{n-m}} \cancel{(m-1)!}}{(\alpha+1)^{\overline{n-1}}}$$
$$= \frac{(n-1)!}{(n-m)!} \frac{\alpha^{\overline{n-m}}}{(\alpha+1)^{\overline{n-1}}}. \tag{A.45}$$

In [5, Proposition 4.11], it was shown that the cluster sizes $\tilde{\mathsf{m}}_s(n)$ are exchangeable and the joint pmf is given by

$$p_{\tilde{\mathsf{m}}_{1:S(n)}(n)}(\tilde{\boldsymbol{m}}_{1:S(n)}(n)) = \frac{\alpha^{S(n)}\Gamma(\alpha)\prod_{s=1}^{S(n)}\Gamma(\tilde{m}_s(n))}{\Gamma(\alpha+n)}. \tag{A.46}$$

We note that the joint pmf (A.46) is a product in $\tilde{m}_s(n)$ and is invariant to permutations (see (3.19)). Due to the exchangeability of the cluster sizes $\tilde{\mathsf{m}}_s(n)$, we now conclude that the marginal distribution of any cluster size $\tilde{\mathsf{m}}_s(n)$ is given by

$$p_{\tilde{\mathsf{m}}_s(n)}(m) = \frac{(n-1)!}{(n-m)!} \frac{\alpha^{\overline{n-m}}}{(\alpha+1)^{\overline{n-1}}}, \quad \text{for} \quad s = 1, \ldots, S(n), \quad \text{and} \quad \text{for} \quad m = 1, \ldots n. \tag{A.47}$$

Since the cluster size $\tilde{\mathsf{m}}_s(n)$ is a permuted version of $\mathsf{m}_l(n)$ (see (3.48)), the pmf $p_{\mathsf{m}_l(N)}(m)$ is finally obtained for $n = N$ samples as

$$p_{\mathsf{m}_l(N)}(m) = \frac{(N-1)!}{(N-m)!} \frac{\alpha^{\overline{N-m}}}{(\alpha+1)^{\overline{N-1}}}, \quad m = 1, \ldots N. \tag{A.48}$$

for any $l \in \mathcal{C}(N)$ and $1 \leq m \leq N$.

B Matrix Inversion Identities

We consider two positive definite matrices $\boldsymbol{\Sigma_A}$ and $\boldsymbol{\Sigma_B}$. Since positive definiteness implies nonsingularity, $\boldsymbol{\Sigma_A}^{-1}$ and $\boldsymbol{\Sigma_B}^{-1}$ exist, and we have $\boldsymbol{\Sigma_A}^{-1}\boldsymbol{\Sigma_A} = \mathbf{I}$ and $\boldsymbol{\Sigma_B}^{-1}\boldsymbol{\Sigma_B} = \mathbf{I}$. Therefore, we obtain

$$\begin{aligned}\left(\boldsymbol{\Sigma_A}^{-1} + \boldsymbol{\Sigma_B}^{-1}\right)^{-1} &= \left(\boldsymbol{\Sigma_A}^{-1}\boldsymbol{\Sigma_B}\boldsymbol{\Sigma_B}^{-1} + \boldsymbol{\Sigma_A}^{-1}\boldsymbol{\Sigma_A}\boldsymbol{\Sigma_B}^{-1}\right)^{-1} \\ &= \left(\boldsymbol{\Sigma_A}^{-1}\left(\boldsymbol{\Sigma_B} + \boldsymbol{\Sigma_A}\right)\boldsymbol{\Sigma_B}^{-1}\right)^{-1} \\ &= \boldsymbol{\Sigma_B}\left(\boldsymbol{\Sigma_A} + \boldsymbol{\Sigma_B}\right)^{-1}\boldsymbol{\Sigma_A}. \end{aligned} \tag{B.1}$$

Equivalently, since the addition of matrices is commutative, we also have

$$\left(\boldsymbol{\Sigma_A}^{-1} + \boldsymbol{\Sigma_B}^{-1}\right)^{-1} = \left(\boldsymbol{\Sigma_B}^{-1} + \boldsymbol{\Sigma_A}^{-1}\right)^{-1} = \boldsymbol{\Sigma_A}\left(\boldsymbol{\Sigma_B} + \boldsymbol{\Sigma_A}\right)^{-1}\boldsymbol{\Sigma_B}, \tag{B.2}$$

where (B.1) was used with $\boldsymbol{\Sigma_A}$ and $\boldsymbol{\Sigma_B}$ interchanged. Combining (B.1) and (B.2) gives

$$\boldsymbol{\Sigma_B}\left(\boldsymbol{\Sigma_A} + \boldsymbol{\Sigma_B}\right)^{-1}\boldsymbol{\Sigma_A} = \boldsymbol{\Sigma_A}\left(\boldsymbol{\Sigma_A} + \boldsymbol{\Sigma_B}\right)^{-1}\boldsymbol{\Sigma_B}. \tag{B.3}$$

Furthermore, we have

$$\begin{aligned}\boldsymbol{\Sigma_B}(\boldsymbol{\Sigma_A} + \boldsymbol{\Sigma_B})^{-1}\boldsymbol{\Sigma_A} &= \boldsymbol{\Sigma_A} - \boldsymbol{\Sigma_A} + \boldsymbol{\Sigma_B}(\boldsymbol{\Sigma_A} + \boldsymbol{\Sigma_B})^{-1}\boldsymbol{\Sigma_A} \\ &= \boldsymbol{\Sigma_A} - (\boldsymbol{\Sigma_A} + \boldsymbol{\Sigma_B})(\boldsymbol{\Sigma_A} + \boldsymbol{\Sigma_B})^{-1}\boldsymbol{\Sigma_A} + \boldsymbol{\Sigma_B}(\boldsymbol{\Sigma_A} + \boldsymbol{\Sigma_B})^{-1}\boldsymbol{\Sigma_A} \\ &= \boldsymbol{\Sigma_A} - (\boldsymbol{\Sigma_A} + \boldsymbol{\Sigma_B} - \boldsymbol{\Sigma_B})(\boldsymbol{\Sigma_A} + \boldsymbol{\Sigma_B})^{-1}\boldsymbol{\Sigma_A} \\ &= \boldsymbol{\Sigma_A} - \boldsymbol{\Sigma_A}(\boldsymbol{\Sigma_A} + \boldsymbol{\Sigma_B})^{-1}\boldsymbol{\Sigma_A}. \end{aligned} \tag{B.4}$$

C Product of Gaussian pdfs

We consider a Gaussian prior,

$$f_{\mathbf{x}}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_{\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{x}}), \tag{C.1}$$

and a Gaussian likelihood function

$$f_{\mathbf{y}_n \,|\, \mathbf{x}}(\boldsymbol{y}_n \,|\, \boldsymbol{x}) = \mathcal{N}(\boldsymbol{y}_n; \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{\mu}_{\boldsymbol{y}_n}, \boldsymbol{\Sigma}_{\boldsymbol{y}}), \quad \text{for} \quad n = 1, \ldots, N, \tag{C.2}$$

with square $D \times D$ matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. Let $\mathbf{y}_{1:N} = (\mathbf{y}_1^{\mathrm{T}}, \ldots, \mathbf{y}_N^{\mathrm{T}})^{\mathrm{T}}$ denote the stacked vector of measurements. We assume the measurements to be independent and identically distributed (i.i.d.), using (C.2) we have

$$f_{\mathbf{y}_{1:N} \,|\, \mathbf{x}}(\boldsymbol{y}_{1:N} \,|\, \boldsymbol{x}) = \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{y}_n; \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{\mu}_{\boldsymbol{y}_n}, \boldsymbol{\Sigma}_{\boldsymbol{y}}). \tag{C.3}$$

In what follows we derive an expression for the joint pdf of the measurements $\mathbf{y}_{1:N}$ and the parameter of interest $\mathbf{x}$, i.e., $f_{\mathbf{x}, \mathbf{y}_{1:N}}(\boldsymbol{x}, \boldsymbol{y}_{1:N})$ as well as the marginal pdf $f_{\mathbf{y}_{1:N}}(\boldsymbol{y}_{1:N})$.

C.1 Joint pdf

$$f_{\mathbf{x}, \mathbf{y}_{1:N}}(\boldsymbol{x}, \boldsymbol{y}_{1:N}) = f_{\mathbf{x}}(\boldsymbol{x}) f_{\mathbf{y}_{1:N} \,|\, \mathbf{x}}(\boldsymbol{y}_{1:N} \,|\, \boldsymbol{x}) \tag{C.4}$$

$$= \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_{\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{x}}) \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{y}_n; \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{\mu}_{\boldsymbol{y}_n}, \boldsymbol{\Sigma}_{\boldsymbol{y}}). \tag{C.5}$$

The expression in (C.5) is recognized to be a product of Gaussian pdfs, which is another Gaussian [18, 7.14]. We will now calculate this expression. Using the precision matrices $\boldsymbol{\Lambda}_{\boldsymbol{y}\boldsymbol{y}} = \boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}$ and $\boldsymbol{\Lambda}_{\boldsymbol{x}\boldsymbol{x}} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}$, (C.5) can be written as

$$\begin{aligned}
&f_{\mathbf{x}}(\boldsymbol{x}) f_{\mathbf{y}_{1:N} \,|\, \mathbf{x}}(\boldsymbol{y}_{1:N} \,|\, \boldsymbol{x}) \\
&\propto \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{x}})^{\mathrm{T}} \boldsymbol{\Lambda}_{\boldsymbol{x}\boldsymbol{x}} (\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{x}})\right) \prod_{n=1}^{N} \exp\left(-\frac{1}{2}(\boldsymbol{y}_n - \boldsymbol{A}\boldsymbol{x} - \boldsymbol{B}\boldsymbol{\mu}_{\boldsymbol{y}_n})^{\mathrm{T}} \boldsymbol{\Lambda}_{\boldsymbol{y}\boldsymbol{y}} (\boldsymbol{y}_n - \boldsymbol{A}\boldsymbol{x} - \boldsymbol{B}\boldsymbol{\mu}_{\boldsymbol{y}_n})\right) \\
&= \exp\left(-\frac{1}{2}\left((\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{x}})^{\mathrm{T}} \boldsymbol{\Lambda}_{\boldsymbol{x}\boldsymbol{x}} (\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{x}}) + \sum_{n=1}^{N} (\boldsymbol{y}_n - \boldsymbol{A}\boldsymbol{x} - \boldsymbol{B}\boldsymbol{\mu}_{\boldsymbol{y}_n})^{\mathrm{T}} \boldsymbol{\Lambda}_{\boldsymbol{y}\boldsymbol{y}} (\boldsymbol{y}_n - \boldsymbol{A}\boldsymbol{x} - \boldsymbol{B}\boldsymbol{\mu}_{\boldsymbol{y}_n})\right)\right)
\end{aligned}$$

$$= \exp\left(-\frac{1}{2}E\right), \tag{C.6}$$

with

$$E \triangleq (\boldsymbol{x}-\boldsymbol{\mu_x})^{\mathrm{T}}\boldsymbol{\Lambda_{xx}}(\boldsymbol{x}-\boldsymbol{\mu_x}) + \sum_{n=1}^{N}(\boldsymbol{y}_n - \boldsymbol{A}\boldsymbol{x} - \boldsymbol{B}\boldsymbol{\mu_{y_n}})^{\mathrm{T}}\boldsymbol{\Lambda_{yy}}(\boldsymbol{y}_n - \boldsymbol{A}\boldsymbol{x} - \boldsymbol{B}\boldsymbol{\mu_{y_n}}). \tag{C.7}$$

We claim that E can be written as a quadratic form (see (2.31) and [20, Eq. 2.70]), i.e.,

$$\begin{aligned}\tilde{E} \triangleq &\Big(\sum_{n=1}^{N}(\boldsymbol{y}_n - \tilde{\boldsymbol{\mu}}_{\boldsymbol{y}_n})^{\mathrm{T}}\tilde{\boldsymbol{\Lambda}}_{\boldsymbol{yy}}(\boldsymbol{y}_n - \tilde{\boldsymbol{\mu}}_{\boldsymbol{y}_n})\Big) + \sum_{n=1}^{N}(\boldsymbol{y}_n - \tilde{\boldsymbol{\mu}}_{\boldsymbol{y}_n})^{\mathrm{T}}\tilde{\boldsymbol{\Lambda}}_{\boldsymbol{yx}}(\boldsymbol{x}-\tilde{\boldsymbol{\mu}}_{\boldsymbol{x}}) \\ &+ (\boldsymbol{x}-\tilde{\boldsymbol{\mu}}_{\boldsymbol{x}})^{\mathrm{T}}\tilde{\boldsymbol{\Lambda}}_{\boldsymbol{xy}}\sum_{n=1}^{N}(\boldsymbol{y}_n - \tilde{\boldsymbol{\mu}}_{\boldsymbol{y}_n}) + (\boldsymbol{x}-\tilde{\boldsymbol{\mu}}_{\boldsymbol{x}})^{\mathrm{T}}\tilde{\boldsymbol{\Lambda}}_{\boldsymbol{xx}}(\boldsymbol{x}-\tilde{\boldsymbol{\mu}}_{\boldsymbol{x}}),\end{aligned} \tag{C.8}$$

with unknown means $\tilde{\boldsymbol{\mu}}_{\boldsymbol{y}_n}$ and $\tilde{\boldsymbol{\mu}}_{\boldsymbol{x}}$, precision matrices $\tilde{\boldsymbol{\Lambda}}_{\boldsymbol{xx}}$ and $\tilde{\boldsymbol{\Lambda}}_{\boldsymbol{yy}}$ as well as unknown cross-precision matrices $\tilde{\boldsymbol{\Lambda}}_{\boldsymbol{xy}}$ and $\tilde{\boldsymbol{\Lambda}}_{\boldsymbol{yx}}$.

Completing the Square

By expanding the expression (C.7) into a quadratic form we obtain

$$\begin{aligned}E =&\Big(\sum_{n=1}^{N}(\boldsymbol{y}_n - \boldsymbol{B}\boldsymbol{\mu_{y_n}})^{\mathrm{T}}\boldsymbol{\Lambda_{yy}}(\boldsymbol{y}_n - \boldsymbol{B}\boldsymbol{\mu_{y_n}})\Big) - \sum_{n=1}^{N}(\boldsymbol{y}_n - \boldsymbol{B}\boldsymbol{\mu_{y_n}})^{\mathrm{T}}\boldsymbol{\Lambda_{yy}}\boldsymbol{A}\boldsymbol{x} + \boldsymbol{\mu_x}^{\mathrm{T}}\boldsymbol{\Lambda_{xx}}\boldsymbol{\mu_x} \\ &- \boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\Lambda_{yy}}\sum_{n=1}^{N}(\boldsymbol{y}_n - \boldsymbol{B}\boldsymbol{\mu_{y_n}}) + N\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\Lambda_{yy}}\boldsymbol{A}\boldsymbol{x} + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Lambda_{xx}}\boldsymbol{x} - \boldsymbol{\mu_x}^{\mathrm{T}}\boldsymbol{\Lambda_{xx}}\boldsymbol{x} - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Lambda_{xx}}\boldsymbol{\mu_x} \\ =&\Big(\sum_{n=1}^{N}(\boldsymbol{y}_n - \boldsymbol{B}\boldsymbol{\mu_{y_n}})^{\mathrm{T}}\boldsymbol{\Lambda_{yy}}(\boldsymbol{y}_n - \boldsymbol{B}\boldsymbol{\mu_{y_n}})\Big) - \sum_{n=1}^{N}(\boldsymbol{y}_n - \boldsymbol{B}\boldsymbol{\mu_{y_n}})^{\mathrm{T}}\ \boldsymbol{\Lambda_{yy}}\boldsymbol{A}\boldsymbol{x} - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\Lambda_{yy}}\sum_{n=1}^{N}(\boldsymbol{y}_n - \boldsymbol{B}\boldsymbol{\mu_{y_n}}) \\ &+ \boldsymbol{x}^{\mathrm{T}}(\boldsymbol{\Lambda_{xx}} + N\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\Lambda_{yy}}\boldsymbol{A})\boldsymbol{x} - \boldsymbol{\mu_x}^{\mathrm{T}}(\boldsymbol{\Lambda_{xx}} + N\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\Lambda_{yy}}\boldsymbol{A})\boldsymbol{x} + \boldsymbol{\mu_x}^{\mathrm{T}}N\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\Lambda_{yy}}\boldsymbol{A}\boldsymbol{x} - \boldsymbol{x}^{\mathrm{T}}(\boldsymbol{\Lambda_{xx}} \\ &+ N\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\Lambda_{yy}}\boldsymbol{A})\boldsymbol{\mu_x} + \boldsymbol{x}^{\mathrm{T}}N\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\Lambda_{yy}}\boldsymbol{A}\boldsymbol{\mu_x} + \boldsymbol{\mu_x}^{\mathrm{T}}(\boldsymbol{\Lambda_{xx}} + N\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\Lambda_{yy}}\boldsymbol{A})\boldsymbol{\mu_x} - \boldsymbol{\mu_x}^{\mathrm{T}}N\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\Lambda_{yy}}\boldsymbol{A}\boldsymbol{\mu_x} \\ =&\Big(\sum_{n=1}^{N}(\boldsymbol{y}_n - \boldsymbol{B}\boldsymbol{\mu_{y_n}} - \boldsymbol{A}\boldsymbol{\mu_x})^{\mathrm{T}}\boldsymbol{\Lambda_{yy}}(\boldsymbol{y}_n - \boldsymbol{B}\boldsymbol{\mu_{y_n}} - \boldsymbol{A}\boldsymbol{\mu_x})\Big) - \sum_{n=1}^{N}(\boldsymbol{y}_n - \boldsymbol{B}\boldsymbol{\mu_{y_n}} - \boldsymbol{A}\boldsymbol{\mu_x})^{\mathrm{T}}\boldsymbol{\Lambda_{yy}}\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{\mu_x}) \\ &- (\boldsymbol{x}-\boldsymbol{\mu_x})^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\Lambda_{yy}}\sum_{n=1}^{N}(\boldsymbol{y}_n - \boldsymbol{B}\boldsymbol{\mu_{y_n}} - \boldsymbol{A}\boldsymbol{\mu_x}) + (\boldsymbol{x}-\boldsymbol{\mu_x})^{\mathrm{T}}(\boldsymbol{\Lambda_{xx}} + N\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\Lambda_{yy}}\boldsymbol{A})(\boldsymbol{x}-\boldsymbol{\mu_x}).\end{aligned} \tag{C.9}$$

Comparing Coefficients

By comparing the coefficients in (C.8) with (C.9), so that $E \stackrel{!}{=} \tilde{E}$ we obtain

$$\tilde{\boldsymbol{\mu}}_{\boldsymbol{y}_n} = \boldsymbol{B}\boldsymbol{\mu}_{\boldsymbol{y}_n} + \boldsymbol{A}\boldsymbol{\mu}_{\boldsymbol{x}}, \tag{C.10}$$

and

$$\tilde{\boldsymbol{\mu}}_{\boldsymbol{x}} = \boldsymbol{\mu}_{\boldsymbol{x}}. \tag{C.11}$$

For the precision matrices, we obtain

$$\tilde{\boldsymbol{\Lambda}}_{\boldsymbol{yy}} = \boldsymbol{\Lambda}_{\boldsymbol{yy}}, \tag{C.12}$$

and

$$\tilde{\boldsymbol{\Lambda}}_{\boldsymbol{xx}} = \boldsymbol{\Lambda}_{\boldsymbol{xx}} + N\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\Lambda}_{\boldsymbol{yy}}\boldsymbol{A}. \tag{C.13}$$

For the cross-precision matrices, we obtain

$$\tilde{\boldsymbol{\Lambda}}_{\boldsymbol{yx}} = -\boldsymbol{\Lambda}_{\boldsymbol{yy}}\boldsymbol{A}, \tag{C.14}$$

and

$$\tilde{\boldsymbol{\Lambda}}_{\boldsymbol{xy}} = -\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\Lambda}_{\boldsymbol{yy}}. \tag{C.15}$$

Next, using (C.10) we define the vector of the means

$$\tilde{\boldsymbol{\mu}}_{\boldsymbol{y}_{1:N}} \triangleq (\tilde{\boldsymbol{\mu}}_{\boldsymbol{y}_1}^{\mathrm{T}}, \ldots, \tilde{\boldsymbol{\mu}}_{\boldsymbol{y}_N})^{\mathrm{T}}, \tag{C.16}$$

so we can write (C.8) as

$$\begin{aligned}\tilde{E} = (\boldsymbol{y}_{1:N} - \tilde{\boldsymbol{\mu}}_{\boldsymbol{y}_{1:N}})^{\mathrm{T}}\hat{\boldsymbol{\Lambda}}_{\boldsymbol{yy}}(\boldsymbol{y}_{1:N} - \tilde{\boldsymbol{\mu}}_{\boldsymbol{y}_{1:N}}) + (\boldsymbol{y}_{1:N} - \tilde{\boldsymbol{\mu}}_{\boldsymbol{y}_{1:N}})^{\mathrm{T}}\hat{\boldsymbol{\Lambda}}_{\boldsymbol{yx}}(\boldsymbol{x} - \tilde{\boldsymbol{\mu}}_{\boldsymbol{x}}) \\ + (\boldsymbol{x} - \tilde{\boldsymbol{\mu}}_{\boldsymbol{x}})^{\mathrm{T}}\hat{\boldsymbol{\Lambda}}_{\boldsymbol{xy}}(\boldsymbol{y}_{1:N} - \tilde{\boldsymbol{\mu}}_{\boldsymbol{y}_{1:N}}) + (\boldsymbol{x} - \tilde{\boldsymbol{\mu}}_{\boldsymbol{x}})^{\mathrm{T}}\hat{\boldsymbol{\Lambda}}_{\boldsymbol{xx}}(\boldsymbol{x} - \tilde{\boldsymbol{\mu}}_{\boldsymbol{x}}),\end{aligned} \tag{C.17}$$

with (see (C.12))

$$\hat{\boldsymbol{\Lambda}}_{\boldsymbol{yy}} = \mathbf{I}_N \otimes \tilde{\boldsymbol{\Lambda}}_{\boldsymbol{yy}} = \mathbf{I}_N \otimes \boldsymbol{\Lambda}_{\boldsymbol{yy}}, \tag{C.18}$$

and (see (C.13))

$$\hat{\boldsymbol{\Lambda}}_{\boldsymbol{xx}} = \tilde{\boldsymbol{\Lambda}}_{\boldsymbol{xx}} = \boldsymbol{\Lambda}_{\boldsymbol{xx}} + N\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\Lambda}_{\boldsymbol{yy}}\boldsymbol{A}. \tag{C.19}$$

Furthermore, using a all-ones vector $\mathbf{1}_N$ of size $N \times 1$, we obtain (see (C.14))

$$\hat{\boldsymbol{\Lambda}}_{\boldsymbol{yx}} = \mathbf{1}_N \otimes \tilde{\boldsymbol{\Lambda}}_{\boldsymbol{yx}} = -\mathbf{1}_N \otimes \boldsymbol{\Lambda}_{\boldsymbol{yy}} \boldsymbol{A}, \tag{C.20}$$

and (see (C.15))

$$\hat{\boldsymbol{\Lambda}}_{\boldsymbol{xy}} = \mathbf{1}_N^{\mathrm{T}} \otimes \tilde{\boldsymbol{\Lambda}}_{\boldsymbol{xy}} = -\mathbf{1}_N^{\mathrm{T}} \otimes \boldsymbol{A}^{\mathrm{T}} \boldsymbol{\Lambda}_{\boldsymbol{yy}}. \tag{C.21}$$

Finally, since $\tilde{E} = E$, we can write (C.6) and thus also the right hand side of (C.5) as a joint Gaussian distribution in terms of $\boldsymbol{y}_{1:N}$ and $\boldsymbol{x}$, i.e.,

$$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_{\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{x}}) \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{y}_n; \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{\mu}_{\boldsymbol{y}_n}, \boldsymbol{\Sigma}_{\boldsymbol{y}}) = \mathcal{N}\left(\begin{pmatrix} \boldsymbol{y}_{1:N} \\ \boldsymbol{x} \end{pmatrix}; \begin{pmatrix} \tilde{\boldsymbol{\mu}}_{\boldsymbol{y}_{1:N}} \\ \tilde{\boldsymbol{\mu}}_{\boldsymbol{x}} \end{pmatrix}, \begin{pmatrix} \hat{\boldsymbol{\Lambda}}_{\boldsymbol{yy}} & \hat{\boldsymbol{\Lambda}}_{\boldsymbol{yx}} \\ \hat{\boldsymbol{\Lambda}}_{\boldsymbol{xy}} & \hat{\boldsymbol{\Lambda}}_{\boldsymbol{xx}} \end{pmatrix}^{-1} \right), \tag{C.22}$$

or equivalently,

$$\begin{pmatrix} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{yy}} & \hat{\boldsymbol{\Sigma}}_{\boldsymbol{yx}} \\ \hat{\boldsymbol{\Sigma}}_{\boldsymbol{xy}} & \hat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}} \end{pmatrix} = \begin{pmatrix} \hat{\boldsymbol{\Lambda}}_{\boldsymbol{yy}} & \hat{\boldsymbol{\Lambda}}_{\boldsymbol{yx}} \\ \hat{\boldsymbol{\Lambda}}_{\boldsymbol{xy}} & \hat{\boldsymbol{\Lambda}}_{\boldsymbol{xx}} \end{pmatrix}^{-1}. \tag{C.23}$$

Using [40, Eq. 2.2], the joint-covariance matrix $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{yy}}$ is given by

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{yy}} = \hat{\boldsymbol{\Lambda}}_{\boldsymbol{yy}}^{-1} + \hat{\boldsymbol{\Lambda}}_{\boldsymbol{yy}}^{-1} \hat{\boldsymbol{\Lambda}}_{\boldsymbol{yx}} \left(\hat{\boldsymbol{\Lambda}}_{\boldsymbol{xx}} - \hat{\boldsymbol{\Lambda}}_{\boldsymbol{xy}} \hat{\boldsymbol{\Lambda}}_{\boldsymbol{yy}}^{-1} \hat{\boldsymbol{\Lambda}}_{\boldsymbol{yx}} \right)^{-1} \hat{\boldsymbol{\Lambda}}_{\boldsymbol{xy}} \hat{\boldsymbol{\Lambda}}_{\boldsymbol{yy}}^{-1}. \tag{C.24}$$

Inserting (C.18)-(C.21) into (C.24) we obtain

$$\begin{aligned} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{yy}} &= (\mathbf{I}_N \otimes \boldsymbol{\Lambda}_{\boldsymbol{yy}})^{-1} + (\mathbf{I}_N \otimes \boldsymbol{\Lambda}_{\boldsymbol{yy}})^{-1} (-\mathbf{1}_N \otimes \boldsymbol{\Lambda}_{\boldsymbol{yy}} \boldsymbol{A}) \\ &\quad \Big(\boldsymbol{\Lambda}_{\boldsymbol{xx}} - N \boldsymbol{A}^{\mathrm{T}} \boldsymbol{\Lambda}_{\boldsymbol{yy}} \boldsymbol{A} - \left(\mathbf{1}_N^{\mathrm{T}} \otimes \boldsymbol{A}^{\mathrm{T}} \boldsymbol{\Lambda}_{\boldsymbol{yy}} \right) (\mathbf{I}_N \otimes \boldsymbol{\Lambda}_{\boldsymbol{yy}})^{-1} (\mathbf{1}_N \otimes \boldsymbol{\Lambda}_{\boldsymbol{yy}} \boldsymbol{A}) \Big)^{-1} \\ &\quad \left(-\mathbf{1}_N^{\mathrm{T}} \otimes \boldsymbol{A}^{\mathrm{T}} \boldsymbol{\Lambda}_{\boldsymbol{yy}} \right) (\mathbf{I}_N \otimes \boldsymbol{\Lambda}_{\boldsymbol{yy}})^{-1} \\ &= \mathbf{I}_N \otimes \boldsymbol{\Lambda}_{\boldsymbol{yy}}^{-1} + \mathbf{1}_N \mathbf{1}_N^{\mathrm{T}} \otimes \boldsymbol{A} \boldsymbol{\Lambda}_{\boldsymbol{xx}}^{-1} \boldsymbol{A}^{\mathrm{T}} \\ &= \mathbf{I}_N \otimes \boldsymbol{\Sigma}_{\boldsymbol{y}} + \mathbf{1}_N \mathbf{1}_N^{\mathrm{T}} \otimes \boldsymbol{A} \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{A}^{\mathrm{T}}, \end{aligned} \tag{C.25}$$

where in the last step $\boldsymbol{\Lambda}_{\boldsymbol{yy}}^{-1} = \boldsymbol{\Sigma}_{\boldsymbol{y}}$ and $\boldsymbol{\Lambda}_{\boldsymbol{xx}}^{-1} = \boldsymbol{\Sigma}_{\boldsymbol{x}}$ was used (see (C.5)). Similar to (C.24), the covariance matrix $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}}$ is given by [40, Eq. 2.2]

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}} = \left(\hat{\boldsymbol{\Lambda}}_{\boldsymbol{xx}} - \hat{\boldsymbol{\Lambda}}_{\boldsymbol{xy}} \hat{\boldsymbol{\Lambda}}_{\boldsymbol{yy}}^{-1} \hat{\boldsymbol{\Lambda}}_{\boldsymbol{yx}} \right)^{-1}. \tag{C.26}$$

Inserting (C.18)-(C.21) into (C.26) finally gives

$$\begin{aligned}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}} &= \left(\boldsymbol{\Lambda}_{\boldsymbol{xx}} + N\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\Lambda}_{\boldsymbol{yy}}\boldsymbol{A} - (-\mathbf{1}_N^{\mathrm{T}} \otimes \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\Lambda}_{\boldsymbol{yy}})(\mathbf{I}_N \otimes \boldsymbol{\Lambda}_{\boldsymbol{yy}})^{-1}(-\mathbf{1}_N \otimes \boldsymbol{\Lambda}_{\boldsymbol{yy}}\boldsymbol{A})\right)^{-1} \\ &= \left(\boldsymbol{\Lambda}_{\boldsymbol{xx}} + N\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\Lambda}_{\boldsymbol{yy}}\boldsymbol{A} - N\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\Lambda}_{\boldsymbol{yy}}\boldsymbol{A}\right)^{-1} \\ &= \boldsymbol{\Sigma}_{\boldsymbol{x}}, \end{aligned} \tag{C.27}$$

where in the last step $\boldsymbol{\Lambda}_{\boldsymbol{xx}}^{-1} = \boldsymbol{\Sigma}_{\boldsymbol{x}}$ was used (see (C.5)). Furthermore, we have

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{xy}} = -\left(\hat{\boldsymbol{\Lambda}}_{\boldsymbol{xx}} - \hat{\boldsymbol{\Lambda}}_{\boldsymbol{xy}}\hat{\boldsymbol{\Lambda}}_{\boldsymbol{yy}}^{-1}\hat{\boldsymbol{\Lambda}}_{\boldsymbol{yx}}\right)^{-1}\hat{\boldsymbol{\Lambda}}_{\boldsymbol{xy}}\hat{\boldsymbol{\Lambda}}_{\boldsymbol{yy}}^{-1}, \tag{C.28}$$

and

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{yx}} = -\hat{\boldsymbol{\Lambda}}_{\boldsymbol{yy}}^{-1}\hat{\boldsymbol{\Lambda}}_{\boldsymbol{yx}}\left(\hat{\boldsymbol{\Lambda}}_{\boldsymbol{xx}} - \hat{\boldsymbol{\Lambda}}_{\boldsymbol{xy}}\hat{\boldsymbol{\Lambda}}_{\boldsymbol{yy}}^{-1}\hat{\boldsymbol{\Lambda}}_{\boldsymbol{yx}}\right)^{-1}. \tag{C.29}$$

By inserting (C.18), (C.21) and (C.26) into (C.28) we obtain

$$\begin{aligned}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{xy}} &= -\hat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}}\left(-\mathbf{1}_N^{\mathrm{T}} \otimes \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\Lambda}_{\boldsymbol{yy}}\right)\left(\mathbf{I}_N \otimes \boldsymbol{\Lambda}_{\boldsymbol{yy}}^{-1}\right) \\ &= \boldsymbol{\Sigma}_{\boldsymbol{x}}\left(\mathbf{1}_N^{\mathrm{T}} \otimes \boldsymbol{A}^{\mathrm{T}}\right), \end{aligned} \tag{C.30}$$

where in the last step (C.27) was used. Similarly, inserting (C.18), (C.20) and (C.26) into (C.29) yields

$$\begin{aligned}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{yx}} &= -\left(\mathbf{I}_N \otimes \boldsymbol{\Lambda}_{\boldsymbol{yy}}^{-1}\right)\left(-\mathbf{1}_N \otimes \boldsymbol{\Lambda}_{\boldsymbol{yy}}\boldsymbol{A}\right)\hat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}} \\ &= \left(\mathbf{1}_N \otimes \boldsymbol{A}\right)\boldsymbol{\Sigma}_{\boldsymbol{x}}. \end{aligned} \tag{C.31}$$

Lastly, we can write (C.4) using (C.22) and (C.23) as a joint Gaussian, i.e.,

$$f_{\mathbf{y}_{1:N},\mathbf{x}}(\boldsymbol{y}_{1:N}, \boldsymbol{x}) = \mathcal{N}\left(\begin{pmatrix}\boldsymbol{y}_{1:N} \\ \boldsymbol{x}\end{pmatrix}; \begin{pmatrix}\tilde{\boldsymbol{\mu}}_{\boldsymbol{y}_{1:N}} \\ \tilde{\boldsymbol{\mu}}_{\boldsymbol{x}}\end{pmatrix}, \begin{pmatrix}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{yy}} & \hat{\boldsymbol{\Sigma}}_{\boldsymbol{yx}} \\ \hat{\boldsymbol{\Sigma}}_{\boldsymbol{xy}} & \hat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}}\end{pmatrix}\right), \tag{C.32}$$

with means $\tilde{\boldsymbol{\mu}}_{\boldsymbol{y}_{1:N}}$ and $\tilde{\boldsymbol{\mu}}_{\boldsymbol{x}}$ given in (C.16) and (C.11) respectively. The components of the partitioned covariance matrix are given in (C.25), (C.27), (C.30) and (C.31).

C.2 Marginal pdf

We are now interested in the marginal distribution, given by

$$f_{\mathbf{y}_{1:N}}(\boldsymbol{y}_{1:N}) = \int_{\boldsymbol{x}} f_{\mathbf{x},\mathbf{y}_{1:N}}(\boldsymbol{x}, \boldsymbol{y}_{1:N}) d\boldsymbol{x} \tag{C.33}$$

$$= \int_{\boldsymbol{x}} f_{\mathbf{x}}(\boldsymbol{x}) f_{\mathbf{y}_{1:N}|\mathbf{x}}(\boldsymbol{y}_{1:N} \,|\, \boldsymbol{x}) d\boldsymbol{x} \tag{C.34}$$

$$= \int_{\boldsymbol{x}} \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_{\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{x}}) \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{y}_n; \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{\mu}_{\boldsymbol{y}_n}, \boldsymbol{\Sigma}_{\boldsymbol{y}}) d\boldsymbol{x}. \tag{C.35}$$

Using (C.32) and [20, 2.98], the marginal pdf $f_{\mathbf{y}_{1:N}}(\boldsymbol{y}_{1:N})$ is obtained as

$$f_{\mathbf{y}_{1:N}}(\boldsymbol{y}_{1:N}) = \mathcal{N}\left(\boldsymbol{y}_{1:N}; \tilde{\boldsymbol{\mu}}_{\boldsymbol{y}_{1:N}}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{y}\boldsymbol{y}}\right), \tag{C.36}$$

with mean $\tilde{\boldsymbol{\mu}}_{\boldsymbol{y}_{1:N}}$ (see (C.16) and (C.10))

$$\tilde{\boldsymbol{\mu}}_{\boldsymbol{y}_{1:N}} = \begin{pmatrix} \boldsymbol{B}\boldsymbol{\mu}_{\boldsymbol{y}_1} + \boldsymbol{A}\boldsymbol{\mu}_{\boldsymbol{x}} \\ \vdots \\ \boldsymbol{B}\boldsymbol{\mu}_{\boldsymbol{y}_N} + \boldsymbol{A}\boldsymbol{\mu}_{\boldsymbol{x}}, \end{pmatrix} \tag{C.37}$$

and covariance matrix $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{y}\boldsymbol{y}}$ (see (C.25))

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{y}\boldsymbol{y}} = \begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{y}} + \boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{A}^{\mathrm{T}} & \boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{A}^{\mathrm{T}} & \dots & \boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{A}^{\mathrm{T}} \\ \boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{A}^{\mathrm{T}} & \boldsymbol{\Sigma}_{\boldsymbol{y}} + \boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{A}^{\mathrm{T}} & \ddots & \boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{A}^{\mathrm{T}} \\ \vdots & \ddots & \ddots & \vdots \\ \boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{A}^{\mathrm{T}} & \dots & \boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{A}^{\mathrm{T}} & \boldsymbol{\Sigma}_{\boldsymbol{y}} + \boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{A}^{\mathrm{T}} \end{pmatrix}. \tag{C.38}$$

References

[1] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice Hall, 1997.

[2] C. Andrieu, N. de Freitas, A. Doucet, and M. I. Jordan, "An Introduction to MCMC for Machine Learning," *Machine Learning*, vol. 50, no. 1–2, pp. 5–43, 2003.

[3] J. Sethuraman, "A Constructive Definition of Dirichlet Priors," *Statistica Sinica*, vol. 4, no. 2, pp. 639–650, 1994.

[4] T. S. Ferguson, "A Bayesian Analysis of Some Nonparametric Problems," *The Annals of Statistics*, vol. 1, no. 2, pp. 209–230, 1973.

[5] S. Ghosal and A. van der Vaart, *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press, 2017.

[6] B. Kreidl, "Bayesian Nonparametric Inference in State-Space Models with an Application to Extended Target Tracking," Master's thesis, TU Wien, 2021.

[7] T. J. Bucco, "Extended Multi-Target Tracking Using Probabilistic Data Association and Bayesian Nonparametric Inference," Master's thesis, TU Wien, 2020.

[8] S. P. Chatzis, D. Korkinof, and Y. Demiris, "A Nonparametric Bayesian Approach toward Robot Learning by Demonstration," *Robotics and Autonomous Systems*, vol. 60, no. 6, pp. 789–802, 2012.

[9] T. Taniguchi, R. Yoshino, and T. Takano, "Multimodal Hierarchical Dirichlet Process-Based Active Perception by a Robot," *Frontiers in Neurorobotics*, vol. 12, 2018.

[10] A. R. Ferreira da Silva, "A Dirichlet Process Mixture Model for Brain MRI Tissue Classification," *Medical Image Analysis*, vol. 11, no. 2, pp. 169–182, 2007.

[11] Z. Zhang, M. Descoteaux, and D. B. Dunson, "Nonparametric Bayes Models of Fiber Curves Connecting Brain Regions," *Journal of the American Statistical Association*, vol. 114, no. 528, pp. 1505–1517, 2019.

[12] D. M. Blei and M. I. Jordan, "Variational Inference for Dirichlet Process Mixtures," *Bayesian Analysis*, vol. 1, no. 1, pp. 121–143, 2006.

[13] H. Ishwaran and L. F. James, "Gibbs Sampling Methods for Stick-Breaking Priors," *Journal of the American Statistical Association*, vol. 96, no. 453, pp. 161–173, 2001.

[14] S. Chib and E. Greenberg, "Understanding the Metropolis-Hastings Algorithm," *The American Statistician*, vol. 49, no. 4, pp. 327–335, 1995.

[15] T. Lipovec, "Variational Inference for Dirichlet Process Mixtures and Application to Gaussian Estimation," Master's thesis, TU Wien, 2023.

[16] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, p. 226–231, AAAI Press, 1996.

[17] D. Arthur and S. Vassilvitskii, "K-Means++: The Advantages of Careful Seeding," in *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, (USA), p. 1027–1035, Society for Industrial and Applied Mathematics, 2007.

[18] E. T. Jaynes, *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.

[19] A. Papoulis and S. Pillai, *Probability, Random Variables, and Stochastic Processes*. Tata McGraw-Hill, 2002.

[20] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2007.

[21] P. Orbanz, "Lecture Notes on Bayesian Nonparametrics." `https://www.gatsby.ucl.ac.uk/~porbanz/papers/porbanz_BNP_draft.pdf`, May 2014.

[22] J. Pitman, "Some Developments of the Blackwell-MacQueen Urn Scheme," *Lecture Notes-Monograph Series*, vol. 30, pp. 245–267, 1996.

[23] F. M. Hoppe, "Pólya-like Urns and the Ewens' Sampling Formula," *Journal of Mathematical Biology*, vol. 20, pp. 91–94, Aug 1984.

[24] F. Eggenberger and G. Pólya, "Über die Statistik verketteter Vorgänge," *ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik*, vol. 3, no. 4, pp. 279–289, 1923.

[25] L. Serafino, "On the de Finetti's Representation Theorem: An Evergreen (and Often Misunderstood) Result at the Foundation of Statistics." `https://philsci-archive.pitt.edu/id/eprint/12059`, March 2016.

[26] I. Berkes and E. Péter, "Exchangeable Random Variables and the Subsequence Principle," *Probability Theory and Related Fields*, vol. 73, pp. 395–413, Sep. 1986.

[27] D. J. Aldous, "Exchangeability and Related Topics," in *École d'Été de Probabilités de Saint-Flour XIII — 1983* (P. L. Hennequin, ed.), pp. 1–198, Springer, 1985.

[28] C. E. Antoniak, "Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems," *The Annals of Statistics*, vol. 2, no. 6, pp. 1152 – 1174, 1974.

[29] S. J. Gershman and D. M. Blei, "A Tutorial on Bayesian Nonparametric Models," *Journal of Mathematical Psychology*, vol. 56, no. 1, pp. 1–12, 2012.

[30] J. F. C. Kingman, "The Representation of Partition Structures," *Journal of the London Mathematical Society*, vol. s2-18, no. 2, pp. 374–380, 1978.

[31] D. Vats, F. Acosta, M. L. Huber, and G. L. Jones, "Understanding Linchpin Variables in Markov Chain Monte Carlo." `https://doi.org/10.48550/arXiv.2210.13574`, 2022.

[32] S. Hess, M. Bierlaire, and J. Polak, "A Systematic Comparison of Continuous and Discrete Mixture Models," *European Transport \ Trasporti Europei*, no. 37, pp. 35–61, 2007.

[33] S. N. Maceachern, "Estimating Normal Means with a Conjugate Style Dirichlet Process Prior," *Communications in Statistics – Simulation and Computation*, vol. 23, no. 3, pp. 727–741, 1994.

[34] S. N. Maceachern and P. Müller, "Estimating Mixture of Dirichlet Process Models," *Journal of Computational and Graphical Statistics*, vol. 7, no. 2, pp. 223–238, 1998.

[35] M. D. Escobar and M. West, "Bayesian Density Estimation and Inference Using Mixtures," *Journal of the American Statistical Association*, vol. 90, no. 430, pp. 577–588, 1995.

[36] S. Blinder, *Guide to Essential Math*. Elsevier, 2 ed., 2013.

[37] M. J. Johnson, J. Saunderson, and A. Willsky, "Analyzing Hogwild Parallel Gaussian Gibbs Sampling," in *Advances in Neural Information Processing Systems* (C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, eds.), vol. 26, Curran Associates, 2013.

[38] Y. Atchadé and L. Wang, "A Fast Asynchronous MCMC Sampler for Sparse Bayesian Inference." `https://arxiv.org/abs/2108.06446`, 2021.

[39] T. L. Griffiths and Z. Ghahramani, "The Indian Buffet Process: An Introduction and Review," *Journal of Machine Learning Research*, vol. 12, no. 32, pp. 1185–1224, 2011.

[40] T.-T. Lu and S.-H. Shiou, "Inverses of 2 × 2 Block Matrices," *Computers and Mathematics with Applications*, vol. 43, no. 1, pp. 119–129, 2002.