SIAM J. MATH. DATA SCI.
Vol. 2, No. 3, pp. 583–606

A Rigorous Theory of Conditional Mean Embeddings[*]

Ilja Klebanov[†], Ingmar Schuster[‡], and T. J. Sullivan[§]

Abstract. Conditional mean embeddings (CMEs) have proven themselves to be a powerful tool in many machine learning applications. They allow the efficient conditioning of probability distributions within the corresponding reproducing kernel Hilbert spaces by providing a linear-algebraic relation for the kernel mean embeddings of the respective joint and conditional probability distributions. Both centered and uncentered covariance operators have been used to define CMEs in the existing literature. In this paper, we develop a mathematically rigorous theory for both variants, discuss the merits and problems of each, and significantly weaken the conditions for applicability of CMEs. In the course of this, we demonstrate a beautiful connection to Gaussian conditioning in Hilbert spaces.

Key words. conditional mean embedding, kernel mean embedding, Gaussian measure, reproducing kernel Hilbert space

AMS subject classifications. 46E22, 62J02, 28C20

DOI. 10.1137/19M1305069

1. Introduction.

Reproducing kernel Hilbert spaces (RKHSs) have long been popular tools in machine learning because of the powerful property—often called the "kernel trick"—that many problems posed in terms of the base set $\mathcal{X}$ of the RKHS $\mathcal{H}$ (e.g., classification into two or more classes) become linear-algebraic problems in $\mathcal{H}$ under the embedding of $\mathcal{X}$ into $\mathcal{H}$ induced by the reproducing kernel $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. This insight has been used to define the *kernel mean embedding* (KME; [3], [24]) $\mu_X \in \mathcal{H}$ of an $\mathcal{X}$-valued random variable X as the $\mathcal{H}$-valued mean of the embedded random variable $k(X, \cdot)$, and also the *conditional mean embedding* (CME; [10], [27]), which seeks to perform conditioning of the original random variable X through application of the Gaussian conditioning formula (also known as the Kálmán update) to the embedded *non-Gaussian* random variable $k(X, \cdot)$. This article aims to provide rigorous mathematical foundations for this attractive but apparently naïve approach to conditional probability, and hence to Bayesian inference.

To be somewhat more precise—while deferring technical points such as topological considerations, existence and uniqueness of conditional distributions, etc., to section 2—let us

[*] Received by the editors December 6, 2019; accepted for publication (in revised form) April 1, 2020; published electronically July 13, 2020.
https://doi.org/10.1137/19M1305069
Funding: The research presented here was supported in part by the German Research Foundation (Deutsche Forschungsgemeinschaft) through project TrU-2 "Demand modelling and control for e-commerce using RKHS transfer operator approaches" of the Excellence Cluster "MATH+ The Berlin Mathematics Research Centre" (EXC-2046/1, project ID: 390685689).

[†] Zuse Institute Berlin, Takustraße 7, 14195 Berlin, Germany (klebanov@zib.de).
[‡] Zalando SE, 11501 Berlin, Germany (ingmar.schuster@zalando.de).
[§] Freie Universität Berlin, Arnimallee 6, 14195 Berlin, Germany (t.j.sullivan@fu-berlin.de); Mathematics Institute and School of Engineering, The University of Warwick, Coventry, CV4 7AL, United Kingdom (t.j.sullivan@warwick.ac.uk); and Zuse Institute Berlin, Takustraße 7, 14195 Berlin, Germany (sullivan@zib.de).

$$\begin{tikzcd}
\textbf{original spaces } \mathcal{X},\mathcal{Y} & \textbf{RKHS feature spaces } \mathcal{H},\mathcal{G} \\
\left\{\begin{matrix} x\in\mathcal{X} \\ X\sim\mathbb{P}_X \\ Y\sim\mathbb{P}_Y \\ (X,Y)\sim\mathbb{P}_{XY} \end{matrix}\right\} \arrow[r, "\text{embed}", "\psi{,}\varphi"'] \arrow[d, "\text{conditioning on}", "X=x"'] & \left\{\begin{matrix} \varphi(x) \\ \psi(Y),\ \varphi(X) \\ \mu_Y,\ C_Y,\ C_{YX} \\ \mu_X,\ C_{XY},\ C_X \end{matrix}\right\} \arrow[d, "\text{conditional mean}", "\text{embedding}"'] \\
(Y|X=x)\sim\mathbb{P}_{Y|X=x} \arrow[r, "\text{embed}"] & \boxed{\mu_{Y|X=x} = \begin{cases} C_{YX}C_X^{-1}\varphi(x) & \text{according to (1.2)} \\ \mu_Y + (C_X^{\dagger}C_{XY})^*\,(\varphi(x)-\mu_X) & \text{by (1.3)} \\ ({}^uC_X^{\dagger}\,{}^uC_{XY})^*\,\varphi(x) & \text{by (1.4)} \end{cases}}
\end{tikzcd}$$

Figure 1.1. *While conditioning of the probability distributions in the original spaces $\mathcal{X},\mathcal{Y}$ is a possibly complicated, nonlinear problem, the corresponding formula for their KMEs reduces to elementary linear algebra—a common guiding theme when working with RKHSs.*

fix two RKHSs $\mathcal{H}$ and $\mathcal{G}$ over $\mathcal{X}$ and $\mathcal{Y}$, respectively, with reproducing kernels k and ℓ and canonical feature maps $\varphi(x) := k(x,\,\cdot\,)$ and $\psi(y) := \ell(y,\,\cdot\,)$. Let X and Y be random variables taking values in $\mathcal{X}$ and $\mathcal{Y}$, respectively, with joint distribution $\mathbb{P}_{XY}$ on $\mathcal{X}\times\mathcal{Y}$. Let μ_X, μ_Y, and $\mu_{Y|X=x}$ denote the KMEs of the marginal distributions $\mathbb{P}_X$ of X, $\mathbb{P}_Y$ of Y, and the conditional distribution $\mathbb{P}_{Y|X=x}$ of Y given $X=x$ given by

$$\mu_X := \mathbb{E}[\varphi(X)] \in \mathcal{H}, \qquad \mu_Y := \mathbb{E}[\psi(Y)] \in \mathcal{G}, \qquad \mu_{Y|X=x} := \mathbb{E}[\psi(Y)|X=x] \in \mathcal{G}. \tag{1.1}$$

The CME offers a way to perform conditioning of probability distributions on $\mathcal{X}$ and $\mathcal{Y}$ by means of linear algebra in the corresponding feature spaces $\mathcal{H}$ and $\mathcal{G}$ (Figure 1.1). In terms of the kernel covariance operator C_X and cross-covariance operator C_{YX} defined later in (2.3), if C_X is invertible and $\mathbb{E}[g(Y)|X=\,\cdot\,]$ is an element of $\mathcal{H}$ whenever $g\in\mathcal{G}$, then the well-known formula for the CME [27, Theorem 4] is

$$\mu_{Y|X=x} = C_{YX}C_X^{-1}\varphi(x), \qquad x\in\mathcal{X}. \tag{1.2}$$

(We emphasize here that the CME $\mu_{Y|X=x}$ is *defined* in (1.1) as the KME of $\mathbb{P}_{Y|X=x}$; the claim implicit in (1.2) is that $\mu_{Y|X=x}$ can be *realized* through simple linear algebra involving cross-covariance operators; cf. the discussion of [20].) Note that there are in fact two theories of CMEs, one working with *centered* covariance operators [10, 27] and the other with *uncentered* ones [14]. We will discuss both theories in detail, but let us focus for a moment on the centered case for which the above formula was originally derived.

In the trivial case where X and Y are independent, the CME should yield $\mu_{Y|X=x} = \mu_Y$. However, independence implies that $C_{YX}=0$, and so (1.2) yields $\mu_{Y|X=x}=0$, regardless of

x. In order to understand what has gone wrong it is helpful to consider in turn the two cases in which the constant function $\mathbb{1}_{\mathcal{X}}\colon x \mapsto 1$ is, or is not, an element of $\mathcal{H}$.

- If $\mathbb{1}_{\mathcal{X}} \in \mathcal{H}$, then C_X cannot be injective, since $C_X \mathbb{1}_{\mathcal{X}} = 0$, and (1.2) is not applicable.
- If $\mathbb{1}_{\mathcal{X}} \notin \mathcal{H}$ and X and Y are independent, then the assumption $\mathbb{E}[g(Y)|X = \cdot] \in \mathcal{H}$ for $g \in \mathcal{G}$ cannot be fulfilled (except for those special elements $g \in \mathcal{G}$ for which $\mathbb{E}[g(Y)] = 0$ or if $\mathbb{E}[\ell(y, Y)] = 0$ for all $y \in \mathcal{Y}$, respectively), and (1.2) is again not applicable.

In summary, (1.2) is never applicable for independent random variables except in certain degenerate cases. Note that this problem does not occur in the case of uncentered operators, where ${}^{u}C_X$ (defined in (2.5)) is typically injective.

Therefore, this paper aims to provide a rigorous theory of CMEs that not only addresses the above-mentioned pathology but also substantially generalizes the assumptions under which CME can be performed. We will treat both centered and uncentered (cross-)covariance operators, with particular emphasis on the centered case, and we will also exhibit a connection to Gaussian conditioning in general Hilbert spaces.

(1) The standard assumption $\mathbb{E}[g(Y)|X = \cdot] \in \mathcal{H}$ for CME is rather restrictive.[1] We show in section 4 that this assumption can be significantly weakened in the case of centered kernel (cross-)covariance operators as defined in (2.3): only $\mathbb{E}[g(Y)|X = \cdot]$ shifted by some constant function needs to lie in $\mathcal{H}$ (Assumption B). In this setting, the correct expression of the CME formula is

$$\mu_{Y|X=x} = \mu_Y + (C_X^{\dagger} C_{XY})^{*} \, (\varphi(x) - \mu_X) \qquad \text{for } \mathbb{P}_X\text{-a.e. } x \in \mathcal{X}, \tag{1.3}$$

where A^* denotes the adjoint and $A^{\dagger}$ the Moore–Penrose pseudoinverse of a linear operator A. As a first sanity check, note that this formula indeed yields $\mu_{Y|X=x} = \mu_Y$ when X and Y are independent. Similarly, as shown in section 5, for *uncentered* kernel (cross-)covariance operators ${}^{u}C_X$ and ${}^{u}C_{XY}$ as defined later in (2.5), the correct formulation of the CME is

$$\mu_{Y|X=x} = ({}^{u}C_X^{\dagger} \, {}^{u}C_{XY})^{*} \, \varphi(x) \qquad \text{for } \mathbb{P}_X\text{-a.e. } x \in \mathcal{X}. \tag{1.4}$$

(2) Furthermore, the assumption $\mathbb{E}[g(Y)|X = \cdot] \in \mathcal{H}$, $g \in \mathcal{G}$, is hard to check in most applications. To the best of our knowledge, the only verifiable condition that implies this assumption is given by [11, Proposition 4]. However, this condition is itself difficult to check.[2] We will present weaker assumptions (Assumption B*) for the applicability of CMEs that hold whenever the kernel k is characteristic.[3] Characteristic kernels are

[1] Fukumizu, Song, and Gretton [14] themselves write "Note, however, that the assumptions [...] may not hold in general; we can easily give counterexamples for the latter in the case of Gaussian kernels." More precisely, for a Gaussian kernel k on, say, $[0, 1]$ and independent random variables X and Y, $\mathbb{E}[g(Y)|X = \cdot]$ is a constant function for each $g \in \mathcal{G}$, which does not lie in the RKHS corresponding to k (unless it happens to be the zero function) by [31, Corollary 5] or [30, Corollary 4.44].

[2] The original condition of [10, Proposition 4] was verifiable in certain situations, but the proposition itself turned out to be incorrect. The corrected condition in the erratum [11] seems to be much harder to check—at least, no explicit case is given in which it is easier to verify than $\mathbb{E}[g(Y)|X = \cdot]$ being in $\mathcal{H}$ for each $g \in \mathcal{G}$.

[3] A kernel k is called *characteristic* [13] if the kernel mean embedding is injective as a function from $\{\mathbb{Q} \mid \mathbb{Q}$ is a prob. meas. on $\mathcal{X}$ with $\int_{\mathcal{X}} \|\varphi(x)\|_{\mathcal{H}} \, \mathrm{d}\mathbb{Q}(x) < \infty\}$ into $\mathcal{H}$; naturally, the KME cannot be injective as a function from the space of random variables on $\mathcal{X}$ to $\mathcal{H}$, since random variables with the same law embed to the same point of $\mathcal{H}$.

well studied (see, e.g., [29]) and therefore provide a verifiable condition as desired.

(3) The applicability of (1.2) requires the additional assumptions that C_X is injective and that $\varphi(x)$ lies in the range of C_X, which is also hard to verify in practice.[4] We show that both assumptions can be avoided completely by replacing $C_{YX}C_X^{-1}$ in (1.2) by $(C_X^{\dagger}C_{XY})^*$ in (1.3) and $({}^uC_X^{\dagger}\,{}^uC_{XY})^*$ in (1.4), which turn out to be globally defined and bounded operators under rather weak assumptions (Assumptions C and ^{u}C).

(4) The experienced reader will also observe that, modulo the replacement of $C_{YX}C_X^{-1}$ by $(C_X^{\dagger}C_{XY})^*$, (1.3) is identical to the familiar Sherman–Morrison–Woodbury/Schur complement formula for conditional Gaussian distributions, a connection upon which we will elaborate in detail in section 7. We call particular attention to the fact that the random variable $(\psi(Y), \varphi(X))$, which has no reason to be normally distributed, behaves very much like a Gaussian random variable in terms of its conditional mean.

Remark 1.1. Note that we stated (1.3) and (1.4) only for $\mathbb{P}_X$-a.e. $x \in \mathcal{X}$. This is the best that one can generally hope for, since the regular conditional probability $\mathbb{P}_{Y|X=x}$ is uniquely determined only for $\mathbb{P}_X$-a.e. $x \in \mathcal{X}$ [17, Theorem 5.3]. The work on CMEs so far completely ignores the fact that conditioning (especially on events of the form $X = x$) is not trivial, requires certain assumptions, and, in general, yields results only for $\mathbb{P}_X$-a.e. $x \in \mathcal{X}$. In particular, the condition on $\mathbb{E}[g(Y)|X = \cdot]$ to lie in $\mathcal{H}$ is ill posed, since these functions are uniquely defined only $\mathbb{P}_X$-a.e., which in certain situations may be practically nowhere, and the same reasoning applies to the above-mentioned condition given by [11, Proposition 4]. The existence and almost sure uniqueness of the regular conditional probability distribution $\mathbb{P}_{Y|X=x}$ will be addressed in a precise manner in section 2.

Remark 1.2. The focus of this paper is the validity of the nonregularized *population* formulation of the CME in terms of the covariance structure of the KME of the data-generating distribution $\mathbb{P}_{XY}$. The construction of valid CME formulae based on *empirical sample data* (i.e., finitely many draws from $\mathbb{P}_{XY}$) is vital in practice but is also much harder to analyze. We give some remarks on this setting in section SM2 of the supplementary material.

The rest of the paper is structured as follows. Section 2 establishes the notation and problem setting and motivates some of the assumptions that are made. Section 3 discusses several critical assumptions for the applicability of the theory of CMEs and the relations among them. Section 4 proceeds to build a rigorous theory of CMEs using centered covariance operators, with the main results being Theorems 4.3 and 4.4, whereas section 5 does the same for uncentered covariance operators, with the main results being Theorems 5.3 and 5.4. Section 6 reviews the established theory for the conditioning of Gaussian measures on Hilbert spaces, and this is then used in section 7 to rigorously connect the theory of CMEs to the conditioning of Gaussian measures, with the main result being Theorem 7.1. We give some closing remarks in section 8. The supplementary material contains various auxiliary technical results (section SM1) and discusses the possible extension of our results to empirical estimation of CMEs (section SM2).

[4] Note that, typically, $\dim \mathcal{H} = \infty$, in which case the compact operator C_X cannot possibly be surjective. To verify that $\varphi(x) \in \operatorname{ran} C_X$, one would need to compute a singular value decomposition $C_X = \sum_{n\in\mathbb{N}} \sigma_n h_n \otimes h_n$ of C_X and check the Picard condition $\sum_{n\in\mathbb{N}} \sigma_n^{-2} \langle \varphi(x), h_n\rangle_{\mathcal{H}}^2 < \infty$.

2. Setup and notation.

Throughout this paper, when considering Hilbert space-valued random variables $U \in \mathcal{L}^2(\Omega, \Sigma, \mathbb{P}; \mathcal{G})$ and $V \in \mathcal{L}^2(\Omega, \Sigma, \mathbb{P}; \mathcal{H})$ defined over a probability space $(\Omega, \Sigma, \mathbb{P})$, the expected value $\mathbb{E}[U] \coloneqq \int_\Omega U(\omega) \, \mathrm{d}\mathbb{P}(\omega)$ is meant in the sense of a Bochner integral [5, section II.2], as are the uncentered and centered cross-covariance operators

$$^u\mathbb{C}\mathrm{ov}[U, V] \coloneqq \mathbb{E}[U \otimes V] \qquad \text{and} \qquad \mathbb{C}\mathrm{ov}[U, V] \coloneqq \mathbb{E}[(U - \mathbb{E}[U]) \otimes (V - \mathbb{E}[V])]$$

from $\mathcal{H}$ into $\mathcal{G}$, where, for $h \in \mathcal{H}$ and $g \in \mathcal{G}$, the outer product $g \otimes h \colon \mathcal{H} \to \mathcal{G}$ is the rank-one linear operator $(g \otimes h)(h') \coloneqq \langle h, h' \rangle_{\mathcal{G}} \, g$. Naturally, we write $^u\mathbb{C}\mathrm{ov}[U]$ and $\mathbb{C}\mathrm{ov}[U]$ for the covariance operators $^u\mathbb{C}\mathrm{ov}[U, U]$ and $\mathbb{C}\mathrm{ov}[U, U]$, respectively, and all of the above reduces to the usual definitions in the scalar-valued case. Both the centered and uncentered covariance operators of a square-integrable random variable are self-adjoint and nonnegative, and—in the separable Hilbert case that is our exclusive focus—also trace-class (see [2, 22] for the centered case; the uncentered case follows from [15, Corollary 2.1]).

Our treatment of CMEs will operate under the following assumptions and notation.

Assumption 2.1.

(a) $(\Omega, \Sigma, \mathbb{P})$ is a probability space, $\mathcal{X}$ is a measurable space, and $\mathcal{Y}$ is a Borel space.[5]

(b) $k \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $\ell \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ are symmetric and positive definite kernels, such that $k(x, \cdot)$ and $\ell(y, \cdot)$ are Borel-measurable functions for each $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.

(c) $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ and $(\mathcal{G}, \langle \cdot, \cdot \rangle_{\mathcal{G}})$ are the corresponding RKHSs, which we assume to be separable. Indeed, according to [18], if the base sets $\mathcal{X}$ and $\mathcal{Y}$ are separable absolute Borel spaces or analytic subsets of Polish spaces, then separability of $\mathcal{H}$ and $\mathcal{G}$ follows from the measurability of their respective kernels and feature maps.

(d) The corresponding canonical feature maps are $\varphi \colon \mathcal{X} \to \mathcal{H}$, $\varphi(x) \coloneqq k(x, \cdot)$, and $\psi \colon \mathcal{Y} \to \mathcal{G}$, $\psi(y) \coloneqq \ell(y, \cdot)$, respectively. Note that they satisfy the "reproducing properties" $\langle h, \varphi(x) \rangle_{\mathcal{H}} = h(x)$, $\langle g, \psi(y) \rangle_{\mathcal{G}} = g(y)$ for $x \in \mathcal{X}, y \in \mathcal{Y}, h \in \mathcal{H}, g \in \mathcal{G}$ and that φ and ψ are Borel measurable in view of [30, Lemma 4.25].

(e) $X \colon \Omega \to \mathcal{X}$ and $Y \colon \Omega \to \mathcal{Y}$ are random variables with distributions $\mathbb{P}_X$ and $\mathbb{P}_Y$ and joint distribution $\mathbb{P}_{XY}$. Assumption 2.1(a) and [17, Theorem 5.3] ensure the existence of a $\mathbb{P}_X$-a.e.-unique regular version of the conditional probability distribution $\mathbb{P}_{Y|X=x}$; the choice of a representative of $\mathbb{P}_{Y|X=x}$ has no impact on our results. We assume that

$$\mathbb{E}\big[\|\varphi(X)\|_{\mathcal{H}}^2 + \|\psi(Y)\|_{\mathcal{G}}^2\big] < \infty, \tag{2.1}$$

which also implies that $\mathcal{X}_Y \coloneqq \{x \in \mathcal{X} \mid \mathbb{E}\big[\|\psi(Y)\|_{\mathcal{G}}^2 | X = x\big] < \infty\}$ has full $\mathbb{P}_X$ measure.[6] Hence, $\mathcal{H} \subseteq \mathcal{L}^2(\mathbb{P}_X)$, $\mathcal{G} \subseteq \mathcal{L}^2(\mathbb{P}_Y)$, and $\mathcal{G} \subseteq \mathcal{L}^2(\mathbb{P}_{Y|X=x})$ for $x \in \mathcal{X}_Y$ since, by the reproducing property and the Cauchy–Schwarz inequality,

$$\begin{aligned} \|h\|_{\mathcal{L}^2(\mathbb{P}_X)}^2 &= \int_{\mathcal{X}} |h(x)|^2 \, \mathrm{d}\mathbb{P}_X(x) = \int_{\mathcal{X}} |\langle h, \varphi(x) \rangle_{\mathcal{H}}|^2 \, \mathrm{d}\mathbb{P}_X(x) \\ &\leq \int_{\mathcal{X}} \|h\|_{\mathcal{H}}^2 \|\varphi(x)\|_{\mathcal{H}}^2 \, \mathrm{d}\mathbb{P}_X(x) = \mathbb{E}[\|\varphi(X)\|_{\mathcal{H}}^2] \, \|h\|_{\mathcal{H}}^2 \end{aligned} \tag{2.2}$$

[5] A space $\mathcal{Y}$ is called a *Borel space* if it is Borel isomorphic to a Borel subset of $[0, 1]$. In particular, $\mathcal{Y}$ is a Borel space if it is Polish, i.e., if it is separable and completely metrizable; see [17, Chapter 1].

[6] Otherwise, $\mathbb{E}[\|\psi(Y)\|_{\mathcal{G}}^2] = \mathbb{E}[\mathbb{E}[\|\psi(Y)\|_{\mathcal{G}}^2 | X]]$ could not be finite.

for all $h \in \mathcal{H}$, and similarly for $g \in \mathcal{G}$ and $\mathbb{P}_Y$, $\mathbb{P}_{Y|X=x}$, $x \in \mathcal{X}_Y$. It follows from (2.2) that the inclusions $\iota_{\varphi,\mathbb{P}_X} \colon \mathcal{H} \hookrightarrow \mathcal{L}^2(\mathbb{P}_X)$, $\iota_{\psi,\mathbb{P}_Y} \colon \mathcal{G} \hookrightarrow \mathcal{L}^2(\mathbb{P}_Y)$ are bounded linear operators, and so is $\iota_{\psi,\mathbb{P}_{Y|X=x}} \colon \mathcal{G} \hookrightarrow \mathcal{L}^2(\mathbb{P}_{Y|X=x})$ for $x \in \mathcal{X}_Y$.

(f) We further assume that, for all $h \in \mathcal{H}$, $h = 0$ $\mathbb{P}_X$-a.e. in $\mathcal{X}$ if and only if $h = 0$, i.e., almost everywhere equality separates points in $\mathcal{H}$. This assumption clearly holds if k is continuous and the topological support of $\mathbb{P}_X$ is all of $\mathcal{X}$.[7] It ensures that we can view $\mathcal{H}$ as a subspace of $L^2(\mathbb{P}_X)$ and write $f \in \mathcal{H}$ for functions $f \in L^2(\mathbb{P}_X)$ whenever there exists $h \in \mathcal{H}$ (which, by this assumption, is unique) such that $f = h$ $\mathbb{P}_X$-a.e.

(g) Several derivations will rely on the Bochner space $L^2(\mathbb{P}_X;\mathcal{F})$, which is isometrically isomorphic to the Hilbert tensor product space $L^2(\mathbb{P}_X) \otimes \mathcal{F}$. Here, $\mathcal{F}$ denotes another Hilbert space, which in our case will be equal to either $\mathbb{R}$ or $\mathcal{G}$. Motivated by the discussion in section 1 and the fact that $\mathbb{C}\mathrm{ov}[f(X), f(X)] = \mathbb{V}[f(X)] = 0$ if and only if f is $\mathbb{P}_X$-a.e. constant, we consider the quotient space $L^2_{\mathcal{C}}(\mathbb{P}_X;\mathcal{F}) := L^2(\mathbb{P}_X;\mathcal{F})/\mathcal{C}$,[8]

$$\mathcal{C} := \{ f \in L^2(\mathbb{P}_X;\mathcal{F}) \mid \exists c \in \mathcal{F} \colon f(x) = c \text{ for } \mathbb{P}_X\text{-a.e. } x \in \mathcal{X} \},$$
$$\langle [f_1], [f_2] \rangle_{L^2_{\mathcal{C}}(\mathbb{P}_X;\mathcal{F})} := \langle f_1 - \mathbb{E}[f_1(X)], f_2 - \mathbb{E}[f_2(X)] \rangle_{L^2(\mathbb{P}_X;\mathcal{F})}.$$

Note that, in the case $\mathcal{F} = \mathbb{R}$, we obtain $\langle [f_1], [f_2] \rangle_{L^2_{\mathcal{C}}(\mathbb{P}_X;\mathbb{R})} = \mathbb{C}\mathrm{ov}[f_1(X), f_2(X)]$, in which case we will abbreviate the space $L^2(\mathbb{P}_X;\mathbb{R})$ by $L^2(\mathbb{P}_X)$ or simply L^2 and the space $L^2_{\mathcal{C}}(\mathbb{P}_X;\mathbb{R})$ by $L^2_{\mathcal{C}}(\mathbb{P}_X)$ or simply $L^2_{\mathcal{C}}$. For any closed subspace $U \subseteq L^2(\mathbb{P}_X)$ we can view $U \otimes \mathcal{F}$ as a subspace of $L^2(\mathbb{P}_X;\mathcal{F})$ by the above isometry and identify $(U \otimes \mathcal{F})_{\mathcal{C}} := (U \otimes \mathcal{F})/((U \otimes \mathcal{F}) \cap \mathcal{C})$ with a subspace of $L^2_{\mathcal{C}}(\mathbb{P}_X;\mathcal{F})$. Note that, in the particular case $U \subseteq \mathcal{H}$ (with U closed in $L^2(\mathbb{P}_X)$), the construction of $U \otimes \mathcal{F}$ and $(U \otimes \mathcal{F})_{\mathcal{C}}$ treats U as a subspace of $L^2(\mathbb{P}_X)$ and ignores the existence of the RKHS norm $\|\cdot\|_{\mathcal{H}}$.

(h) We use overlines and superscripts to denote topological closures, so that, for example, $\overline{\mathcal{H}_{\mathcal{C}}}^{L^2_{\mathcal{C}}}$ is the closure of $\mathcal{H}_{\mathcal{C}}$ with respect to the norm $\|\cdot\|_{L^2_{\mathcal{C}}}$, and $\overline{\mathcal{H}}^{L^2}$ is the closure of $\mathcal{H}$ with respect to the norm $\|\cdot\|_{L^2}$.

(i) Since φ and ψ are Borel measurable, $Z := (\psi(Y), \varphi(X))$ is a well-defined $\mathcal{G} \oplus \mathcal{H}$-valued random variable; (2.1) ensures that Z has finite second moment, and hence its mean $\mathbb{E}[Z]$ and covariance operator $\mathbb{C}\mathrm{ov}[Z]$ are well defined, Sazonov's theorem implies that $\mathbb{C}\mathrm{ov}[Z]$ has finite trace, and we obtain the following block structures:

$$\mu := \mathbb{E}\left[\begin{pmatrix} \psi(Y) \\ \varphi(X) \end{pmatrix} \right] = \begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix}, \qquad C := \mathbb{C}\mathrm{ov}\left[\begin{pmatrix} \psi(Y) \\ \varphi(X) \end{pmatrix} \right] = \begin{pmatrix} C_Y & C_{YX} \\ C_{XY} & C_X \end{pmatrix}, \tag{2.3}$$

[7] If k is continuous, then so is every $h \in \mathcal{H}$ [21, Theorem 2.3]. So, if $h \in \mathcal{H}$ and $|h(x)| = \varepsilon > 0$ for some $x \in \mathcal{X}$, then $|h| > \varepsilon/2$ on some open neighborhood of x. Thus, if $\operatorname{supp}(\mathbb{P}_X) = \mathcal{X}$, then $h = 0$ $\mathbb{P}_X$-a.e. cannot hold.

[8] By the variational characterization of the expected value $\mathbb{E}[Z]$ of a random variable $Z \in L^2(\mathbb{P};\mathcal{F})$, $\mathbb{E}[Z] = \arg\min_{m \in \mathcal{F}} \mathbb{E}[\|Z - \mathbb{E}[Z]\|_{\mathcal{F}}^2]$, the norm $\|\cdot\|_{L^2_{\mathcal{C}}(\mathbb{P}_X;\mathcal{F})}$ coincides with the norm $\|[f]\| = \inf_{m \in \mathcal{C}} \|f - m\|_{L^2(\mathbb{P}_X;\mathcal{F})}$ induced on $L^2_{\mathcal{C}}(\mathbb{P}_X;\mathcal{F})$ by the norm $\|\cdot\|_{L^2(\mathbb{P}_X;\mathcal{F})}$.

where the components

$$\begin{aligned} &\mu_Y := \mathbb{E}[\psi(Y)], && C_Y := \mathbb{C}\mathrm{ov}[\psi(Y)], && C_{YX} := \mathbb{C}\mathrm{ov}[\psi(Y), \varphi(X)], \\ &\mu_X := \mathbb{E}[\varphi(X)], && C_{XY} := \mathbb{C}\mathrm{ov}[\varphi(X), \psi(Y)], && C_X := \mathbb{C}\mathrm{ov}[\varphi(X)] \end{aligned}$$

are called the *kernel mean embeddings* (KMEs) and *kernel (cross-)covariance operators*, respectively. Note that $C_{XY}^* = C_{YX}$ and that the reproducing properties translate to the KMEs and covariance operators as follows: for arbitrary $h, h' \in \mathcal{H}$ and $g \in \mathcal{G}$,

$$\begin{aligned} \langle h, \mu_X \rangle_{\mathcal{H}} &= \mathbb{E}[h(X)], \\ \langle h, C_X h' \rangle_{\mathcal{H}} &= \mathbb{C}\mathrm{ov}[h(X), h'(X)], \\ \langle h, C_{XY} g \rangle_{\mathcal{H}} &= \mathbb{C}\mathrm{ov}[h(X), g(Y)], \end{aligned}$$

and so on. We are further interested in the *conditional kernel mean embedding* and the *conditional kernel covariance operator* given by

$$\mu_{Y|X=x} = \mathbb{E}[\psi(Y)|X=x], \qquad C_{Y|X=x} = \mathbb{C}\mathrm{ov}[\psi(Y)|X=x], \qquad x \in \mathcal{X}_Y. \tag{2.4}$$

We set $\mu_{Y|X=x} := 0$ on the $\mathbb{P}_X$-null set $\mathcal{X} \setminus \mathcal{X}_Y$. Similarly, $Z = (\psi(Y), \varphi(X))$ has the uncentered kernel covariance structure

$${}^u C := {}^u\mathbb{C}\mathrm{ov}\left[\begin{pmatrix} \psi(Y) \\ \varphi(X) \end{pmatrix}\right] = \begin{pmatrix} {}^u C_Y & {}^u C_{YX} \\ {}^u C_{XY} & {}^u C_X \end{pmatrix}, \tag{2.5}$$

where ${}^u C_Y := {}^u\mathbb{C}\mathrm{ov}[\psi(Y)]$, etc. Note that for $f_1, f_2 \in L^2(\mathbb{P}_X)$, ${}^u\mathbb{C}\mathrm{ov}(f_1(X), f_2(X)) = \langle f_1, f_2 \rangle_{L^2(\mathbb{P}_X)}$, and similarly for functions of Y.

(j) For $g \in \mathcal{G}$ we let $f_g(x) := \mathbb{E}[g(Y)|X=x]$. More precisely,

$$f_g(x) := \begin{cases} \mathbb{E}[g(Y)|X=x] & \text{for } x \in \mathcal{X}_Y, \\ 0 & \text{otherwise.} \end{cases}$$

These functions f_g will be of particular importance since, for $g = \psi(y)$, $y \in \mathcal{Y}$, and $x \in \mathcal{X}$, we obtain $f_{\psi(y)}(x) = \mu_{Y|X=x}(y)$, our main object of interest (note that $\mu_{Y|X=x} \in \mathcal{G}$ for each $x \in \mathcal{X}$, and so its pointwise evaluation at $y \in \mathcal{Y}$ is meaningful). By (2.1), (2.2), and the law of total expectation, $f_g \in L^2(\mathbb{P}_X)$ for every $g \in \mathcal{G}$, since

$$\begin{aligned} \|f_g\|_{L^2(\mathbb{P}_X)} &= \mathbb{E}[f_g(X)^2] = \mathbb{E}\left[\mathbb{E}[g(Y)|X]^2\right] \leq \mathbb{E}\left[\mathbb{E}[g(Y)^2|X]\right] \\ &= \mathbb{E}[g(Y)^2] = \|g\|_{\mathcal{L}^2(\mathbb{P}_Y)} < \infty. \end{aligned}$$

Further, another application of the law of total expectation yields

$$\mathbb{E}[f_g(X)] = \mathbb{E}[g(Y)], \qquad \mathbb{E}[f_{\psi(y)}(X)] = \mu_Y(y). \tag{2.6}$$

(k) For a linear operator A between Hilbert spaces, $A^\dagger$ denotes its Moore–Penrose pseudo-inverse, i.e., the unique extension of $A|_{(\ker A)^\perp}^{-1} \colon \operatorname{ran} A \to (\ker A)^\perp$ to a linear operator $A^\dagger$ defined on $\operatorname{dom} A^\dagger := (\operatorname{ran} A) \oplus (\operatorname{ran} A)^\perp$ subject to the criterion that $\ker A^\dagger = (\operatorname{ran} A)^\perp$. In general, $\operatorname{dom} A^\dagger$ is a dense but proper subspace, and $A^\dagger$ is an unbounded operator; global definition and boundedness occur precisely when $\operatorname{ran} A$ is closed; see, e.g., [7, section 2.1].

Remark 2.2. Measurability of $k(x,\cdot)$ and $\ell(y,\cdot)$ together with the separability of $\mathcal{H}$ and $\mathcal{G}$ guarantees the measurability of φ and ψ [30, Lemma 4.25]. Separability of $\mathcal{H}$ and $\mathcal{G}$ is also needed for Gaussian conditioning (see [19] and section 6), for the existence of a countable orthonormal basis of $\mathcal{H}$, and to ensure that weak (Pettis) and strong (Bochner) measurability of Hilbert-valued random variables coincide.

3. The crucial assumptions for CMEs.

This section discusses various versions of the assumption $f_g \in \mathcal{H}$ under which we are going to prove various versions of the CME formula (note that, by Assumption 2.1(f), their formulations are unambiguous).

Assumption A. For all $g \in \mathcal{G}$, $f_g \in \mathcal{H}$.

Assumption B. For all $g \in \mathcal{G}$ there exist a function $h_g \in \mathcal{H}$ and a constant $c_g \in \mathbb{R}$ such that $h_g = f_g - c_g$ $\mathbb{P}_X$-a.e. in $\mathcal{X}$.

Assumption C. For all $g \in \mathcal{G}$ there exists a function $h_g \in \mathcal{H}$ such that

$$\mathbb{C}\mathrm{ov}[h_g(X) - f_g(X), h(X)] = 0 \qquad \text{for all } h \in \mathcal{H}.$$

In this case we denote $c_g := \mathbb{E}[f_g(X) - h_g(X)]$ (in conformity with Assumption B).

Assumption ^{u}C. For all $g \in \mathcal{G}$ there exists a function $h_g \in \mathcal{H}$ such that

$${}^u\mathbb{C}\mathrm{ov}[h_g(X) - f_g(X), h(X)] = \langle h_g - f_g, h \rangle_{L^2(\mathbb{P}_X)} = 0 \qquad \text{for all } h \in \mathcal{H}.$$

Remark 3.1. Note that A $\implies$ B $\implies$ C, that A $\implies$ ^{u}C, that C $\implies$ B if $\mathcal{H}_{\mathcal{C}} \subseteq L^2_{\mathcal{C}}(\mathbb{P}_X)$ is dense, and that C $\implies$ A and ^{u}C $\implies$ A if $\mathcal{H} \subseteq L^2(\mathbb{P}_X)$ is dense.

Unlike Assumption A, Assumptions B and C do not require the unfavorable property $\mathbb{1}_{\mathcal{X}} \in \mathcal{H}$ for independent random variables X and Y. Instead, this case reduces to the trivial condition $0 \in \mathcal{H}$. At the same time, the proofs of the key properties of CMEs are not affected by replacing Assumption A with Assumption B as long as we work with centered operators (see Theorems 4.1 and 4.3 below). Therefore, it is surprising that this modification has not been considered earlier, even though the issues with independent random variables have been observed before [14]. One reason might be that, instead of centered operators, researchers started using uncentered ones, for which such a modification is not feasible.

Assumption C, on the other hand, is not strong enough for proving the main formula for CMEs (the last statement of Theorem 4.3). Clearly, this cannot be expected: If $\mathcal{X}$ and $\mathcal{G}$ are reasonably large, but $\mathcal{H}$ is not rich enough, e.g., $\mathcal{H} = \{0\}$ or $\mathcal{H} = \operatorname{span}\{\mathbb{1}_{\mathcal{X}}\}$, then no map from $\mathcal{H}$ to $\mathcal{G}$ can cover sufficiently many KMEs, in particular the embeddings of the conditional probability $\mathbb{P}_{Y|X=x}$ for various x (while Assumption C is trivially fulfilled for these choices of $\mathcal{H}$). The weakness of Assumption C lies in the fact that it only requires the vanishing of the orthogonal projection of $[h_g] - [f_g]$ onto $\mathcal{H}_{\mathcal{C}}$. Only if $\mathcal{H}_{\mathcal{C}}$ is rich enough (e.g., if it is dense in $L^2_{\mathcal{C}}$) can this condition have useful implications. A similar reasoning applies to Assumption ^{u}C.

While it is nice to have a weaker form of Assumption A, Assumptions A, B, and C remain hard to check in practice. Another condition, provided by [11, Proposition 4], is also hard to verify in most applications (see footnote 2). Since characteristic kernels are well studied in the literature, Lemma SM1.4 gives hope for a verifiable condition for the applicability of CMEs: it states that $\mathcal{H}_{\mathcal{C}}$ is dense in $L^2_{\mathcal{C}}(\mathbb{P}_X)$ whenever the kernel k is characteristic. So, if the denseness of $\mathcal{H}_{\mathcal{C}}$ in $L^2_{\mathcal{C}}(\mathbb{P}_X)$ were sufficient for performing CMEs, then the condition that k be characteristic would be sufficient as well, thus providing a favorable criterion for the applicability of formula (1.3). A similar argumentation applies to the condition that $\mathcal{H}$ is dense in $L^2(\mathbb{P}_X)$ and the condition that k is L^2-universal.[9] Unfortunately, neither condition implies Assumption B. Therefore, we will consider the following slightly weaker versions of Assumptions A and B, under which CMEs can be performed if one allows for certain finite-rank approximations of the (cross-)covariance operators.

Assumption A*. For all $g \in \mathcal{G}$, $f_g \in \overline{\mathcal{H}}^{L^2}$.

Assumption B*. For all $g \in \mathcal{G}$ there exist a function $h_g \in \overline{\mathcal{H}}^{L^2}$ and a constant $c_g \in \mathbb{R}$ such that $h_g = f_g - c_g$ $\mathbb{P}_X$-a.e. in $\mathcal{X}$.

Note that Assumptions C and ^{u}C have no weaker versions, since they would become trivial if $h_g \in \mathcal{H}_{\mathcal{C}}$ were replaced by $h_g \in \overline{\mathcal{H}_{\mathcal{C}}}^{L^2_{\mathcal{C}}}$ and $h_g \in \mathcal{H}$ by $h_g \in \overline{\mathcal{H}}^{L^2}$, respectively.

Remark 3.2. In terms of the spaces $L^2_{\mathcal{C}}$ and $\mathcal{H}_{\mathcal{C}}$, Assumptions A–B* can be reformulated as follows: For all $g \in \mathcal{G}$,

(A) $f_g \in \mathcal{H}$;
(B) $[f_g] \in \mathcal{H}_{\mathcal{C}}$;
(C) the orthogonal projection $P_{\overline{\mathcal{H}_{\mathcal{C}}}^{L^2_{\mathcal{C}}}}[f_g]$ of $[f_g]$ onto $\overline{\mathcal{H}_{\mathcal{C}}}^{L^2_{\mathcal{C}}}$ lies in $\mathcal{H}_{\mathcal{C}}$;
(^{u}C) the orthogonal projection $P_{\overline{\mathcal{H}}^{L^2}} f_g$ of f_g onto $\overline{\mathcal{H}}^{L^2}$ lies in $\mathcal{H}$;
(A*) $f_g \in \overline{\mathcal{H}}^{L^2}$;
(B*) $[f_g] \in \overline{\mathcal{H}_{\mathcal{C}}}^{L^2_{\mathcal{C}}}$.

In summary, we consider the hierarchy of assumptions illustrated in Figure 3.1. The main contributions of this paper are rigorous proofs of three versions of the CME formula under various assumptions:

- Theorem 4.3 uses Assumption B and centered operators.
- Theorem 4.4 uses Assumption B* and finite-rank approximations of centered operators.
- Theorem 5.3 uses Assumption A and uncentered operators.
- Theorem 5.4 uses Assumption A* and finite-rank approximations of uncentered operators.

Note that the theorems for uncentered covariance operators require stronger assumptions than their centered counterparts and that Theorem 5.4 provides weaker statements than its centered analogue Theorem 4.4 (we show only the convergence in $L^2(\mathbb{P}_X; \mathcal{G})$, which does not guarantee convergence for $\mathbb{P}_X$-a.e. $x \in \mathcal{X}$).

4. Theory for centered operators.

In this section we will formulate and prove two versions of the CME formula (1.3)—the original one under Assumption B and a weaker version involving finite-rank approximations $C_X^{(n)}, C_{XY}^{(n)}$ of the (cross-)covariance operators under Assumption B*. The following theorem demonstrates the importance of Assumption C (which

[9] A kernel k on $\mathcal{X}$ is called L^2-*universal* [28] if it is Borel measurable and bounded and if $\mathcal{H}$ is dense in $L^2(\mathbb{Q})$ for any probability measure $\mathbb{Q}$ on $\mathcal{X}$. Any L^2-universal kernel is characteristic [28].

$$\begin{tikzcd}
\boxed{\mathcal{H} = L^2(\mathbb{P}_X)} \arrow[r, Rightarrow] \arrow[d, Rightarrow] & \boxed{\mathcal{H}_\mathcal{C} = L^2_\mathcal{C}(\mathbb{P}_X)} \arrow[d, Rightarrow] & \\
\boxed{\text{A: } f_g \in \mathcal{H} \text{ for } g \in \mathcal{G}} \arrow[r, Rightarrow] \arrow[dd, Rightarrow] \arrow[drr, Rightarrow, bend right=10] & \boxed{\text{B: } [f_g] \in \mathcal{H}_\mathcal{C} \text{ for } g \in \mathcal{G}} \arrow[r, Rightarrow] \arrow[dd, Rightarrow] & \boxed{\text{C: } P_{\overline{\mathcal{H}_\mathcal{C}}^{L^2_\mathcal{C}}}[f_g] \in \mathcal{H}_\mathcal{C} \text{ for } g \in \mathcal{G}} \\
 & & \boxed{{}^u\text{C: } P_{\overline{\mathcal{H}}^{L^2}} f_g \in \mathcal{H} \text{ for } g \in \mathcal{G}} \\
\boxed{\text{A}^*\text{: } f_g \in \overline{\mathcal{H}}^{L^2} \text{ for } g \in \mathcal{G}} \arrow[r, Rightarrow] & \boxed{\boxed{\text{B}^*\text{: } [f_g] \in \overline{\mathcal{H}_\mathcal{C}}^{L^2_\mathcal{C}} \text{ for } g \in \mathcal{G}}} & \\
\boxed{\mathcal{H} \text{ dense in } L^2(\mathbb{P}_X)} \arrow[r, Rightarrow] \arrow[u, Rightarrow] & \boxed{\mathcal{H}_\mathcal{C} \text{ dense in } L^2_\mathcal{C}(\mathbb{P}_X)} \arrow[u, Rightarrow] & \\
\boxed{k \text{ is } L^2\text{-universal}} \arrow[r, Rightarrow] \arrow[u, Rightarrow] & \boxed{k \text{ is characteristic}} \arrow[u, Rightarrow] &
\end{tikzcd}$$

Figure 3.1. *A hierarchy of CME-related assumptions. Sufficient conditions for validity of the CME formula are indicated by solid boxes, while the insufficient Assumptions* C *and* ^{u}C*, indicated by dashed boxes, have several strong theoretical implications and Assumption* C *has a beautiful connection to Gaussian conditioning (Theorem* 7.3*). Assumption* B* *is the most favorable one, since it is verifiable in practice and, by Lemma* SM1.4*, in particular is fulfilled if the kernel is universal or even just characteristic (marked in green). The shaded boxes correspond to Theorems* 4.3*,* 4.4*,* 5.3*, and* 5.4*.*

follows from Assumption B). It implies that the range of C_{XY} is contained in that of C_X, making the operator $C_X^\dagger C_{XY}$ well defined. By Theorem SM1.1 it is even a bounded operator, which is a nontrivial result requiring the application of the closed graph theorem.[10]

Similar considerations cannot be performed, in general, under Assumption B* alone: it can no longer be expected that $\operatorname{ran} C_{XY} \subseteq \operatorname{ran} C_X$, which is why we must introduce the above-mentioned finite-rank approximations in order to guarantee that $\operatorname{ran} C_{XY}^{(n)} \subseteq \operatorname{ran} C_X^{(n)}$.

In summary, Assumption B allows for the simple CME formula (1.3) by Theorem 4.1, while under Assumption B* we must make a detour using certain approximations. Note that this distinction is very similar to the theory of Gaussian conditioning in Hilbert spaces introduced by [19] and recapped in section 6 below, a connection that will be elaborated upon in detail in section 7.

Theorem 4.1. *Under Assumption* 2.1*, the following statements are equivalent:*

(i) *Assumption* C *holds.*

(ii) *For each $g \in \mathcal{G}$ there exists $h_g \in \mathcal{H}$ such that $C_X h_g = C_{XY} g$.*

(iii) $\operatorname{ran} C_{XY} \subseteq \operatorname{ran} C_X$.

Proof. Note that (iii) is just a reformulation of (ii), so we only have to prove (i) $\iff$ (ii).

[10] Furthermore, by Lemma SM1.3, this operator is actually Hilbert–Schmidt when thought of as an operator taking values in the appropriate L^2 space rather than in the RKHS.

Let $g \in \mathcal{G}$ and $h, h_g \in \mathcal{H}$. By Lemma SM1.6, $\mathbb{C}\mathrm{ov}[h(X), f_g(X)] = \langle h, C_{XY} g \rangle_{\mathcal{H}}$, and so

$$\begin{aligned}\mathbb{C}\mathrm{ov}[h(X), h_g(X)] = \mathbb{C}\mathrm{ov}[h(X), f_g(X)]\ \forall h \in \mathcal{H} &\iff \langle h, C_X h_g \rangle_{\mathcal{H}} = \langle h, C_{XY} g \rangle_{\mathcal{H}}\ \forall h \in \mathcal{H} \\ &\iff C_X h_g = C_{XY} g. \qquad \blacksquare\end{aligned}$$

Note that Assumption C implies that $[h_g] \in \mathcal{H}_{\mathcal{C}}$ is the orthogonal projection of $[f_g] \in L^2_{\mathcal{C}}$ onto $\mathcal{H}_{\mathcal{C}}$ with respect to $\langle \cdot, \cdot \rangle_{L^2_{\mathcal{C}}}$ (see the reformulation of Assumption C in Remark 3.2). Therefore, there might be some ambiguity in the choice of $h_g \in \mathcal{H}$ if $\mathcal{H}$ contains constant functions. However, there is a particular choice of h_g that always works.

Proposition 4.2. *Under Assumption 2.1, if Assumption B or C holds, then h_g may be chosen as*

$$h_g = C_X^{\dagger} C_{XY} g. \tag{4.1}$$

More precisely, if Assumption C holds, then $\mathbb{C}\mathrm{ov}[(C_X^{\dagger} C_{XY} g)(X) - f_g(X), h(X)] = 0$ for all $h \in \mathcal{H}$ and $g \in \mathcal{G}$, and if Assumption B holds, or even just $f_g \in \mathcal{H}_{\mathcal{C}}$ for some $g \in \mathcal{G}$, then there exists a constant $c_g \in \mathbb{R}$ such that $\mathbb{P}_X$-a.e. $f_g = c_g + C_X^{\dagger} C_{XY} g$.

Proof. By Theorem 4.1, (4.1) is well defined. Under Assumption C, for all $g \in \mathcal{G}$ and $h \in \mathcal{H}$, and appealing to Theorem 4.1 and Lemma SM1.6,

$$\mathbb{C}\mathrm{ov}[h(X), (C_X^{\dagger} C_{XY} g)(X)] = \langle h, C_X C_X^{\dagger} C_{XY} g \rangle_{\mathcal{H}} = \langle h, C_{XY} g \rangle_{\mathcal{H}} = \mathbb{C}\mathrm{ov}[h(X), f_g(X)].$$

If $f_g \in \mathcal{H}_{\mathcal{C}}$ for some $g \in \mathcal{G}$, then there exist a function $h'_g \in \mathcal{H}$ and a constant $c'_g \in \mathbb{R}$ such that, $\mathbb{P}_X$-a.e. in $\mathcal{X}$, $h'_g = f_g - c'_g$. Theorem 4.1 implies that $C_X h'_g = C_{XY} g$, and so Lemma SM1.5 implies that $h'_g - C_X^{\dagger} C_{XY} g$ is constant $\mathbb{P}_X$-a.e. Hence, $f_g - C_X^{\dagger} C_{XY} g$ is constant $\mathbb{P}_X$-a.e. $\blacksquare$

We now give our first main result, the rigorous statement of the CME formula for centered (cross-)covariance operators. In fact, we give two results: a "weak" result (4.2) under Assumption C in which the CME, as a function on $\mathcal{X}$, holds only when tested against elements of $\mathcal{H}$ in the $L^2(\mathbb{P}_X)$ inner product, and a "strong" almost-sure equality in $\mathcal{G}$ (4.3) under Assumption B.

Theorem 4.3 (centered CME). *Under Assumptions 2.1 and C, $C_X^{\dagger} C_{XY} \colon \mathcal{G} \to \mathcal{H}$ is a bounded (see footnote 10) operator and, for all $y \in \mathcal{Y}$ and $h \in \mathcal{H}$,*

$$\langle h, \mu_{Y|X=\,\cdot\,}(y) \rangle_{L^2(\mathbb{P}_X)} = \Big\langle h, \big(\mu_Y + (C_X^{\dagger} C_{XY})^* (\varphi(\cdot) - \mu_X)\big)(y) \Big\rangle_{L^2(\mathbb{P}_X)}. \tag{4.2}$$

Suppose in addition that any of the following four conditions holds:
(i) *the kernel k is characteristic;*
(ii) *$\mathcal{H}_{\mathcal{C}}$ is dense in $L^2_{\mathcal{C}}(\mathbb{P}_X)$;*
(iii) *Assumption B holds;*
(iv) *$f_{\psi(y)} \in \mathcal{H}_{\mathcal{C}}$ for each $y \in \mathcal{Y}$.*
Then, for $\mathbb{P}_X$-a.e. $x \in \mathcal{X}$,

$$\mu_{Y|X=x} = \mu_Y + (C_X^{\dagger} C_{XY})^* (\varphi(x) - \mu_X). \tag{4.3}$$

Proof. Theorems SM1.1 and 4.1 imply that $C_X^{\dagger} C_{XY}$ is well defined and bounded (see footnote 10) and that, for each $g \in \mathcal{G}$, we may choose the function $h_g \in \mathcal{H}$ in Assumptions B and C to be $h_g = C_X^{\dagger} C_{XY} g$ (by Proposition 4.2). Now (2.6), Lemma SM1.7, and the definition of c_g (see Assumption C) yield that, for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$,

$$h_{\psi(y)}(x) + c_{\psi(y)} = \big(\mu_Y + (C_X^{\dagger} C_{XY})^{*}(\varphi(x) - \mu_X)\big)(y). \tag{4.4}$$

This yields (4.2) for each $h \in \mathcal{H}$ via

$$\begin{aligned}
&\langle h, \big(\mu_{Y|X=\,\cdot} - \mu_Y - (C_X^{\dagger} C_{XY})^{*}\,(\varphi(\,\cdot\,) - \mu_X)\big)(y)\rangle_{L^2(\mathbb{P}_X)} \\
&\quad = \langle h, f_{\psi(y)} - h_{\psi(y)} - c_{\psi(y)}\rangle_{L^2(\mathbb{P}_X)} \\
&\quad = \underbrace{\mathbb{C}\mathrm{ov}[h(X), (f_{\psi(y)} - h_{\psi(y)})(X)]}_{=0} + \mathbb{E}[h(X)]\big(\underbrace{\mathbb{E}[(f_{\psi(y)} - h_{\psi(y)})(X)] - c_{\psi}(y)}_{=0}\big) = 0.
\end{aligned}$$

If (i) or (ii) holds (note that, by Lemma SM1.4, (i) $\Longrightarrow$ (ii)), then (4.3) follows directly. If (iii) or (iv) holds (with $f_{\psi(y)} = h_{\psi(y)} + c_{\psi(y)}$, $h_{\psi(y)} \in \mathcal{H}$, $c_{\psi(y)} \in \mathbb{R}$), then (4.3) can be obtained from

$$\begin{aligned}
\mu_{Y|X=x}(y) = \mathbb{E}[\ell(y,Y)|X=x] = f_{\psi(y)}(x) &\overset{(*)}{=} h_{\psi(y)}(x) + c_{\psi(y)} \\
&= \big(\mu_Y + (C_X^{\dagger} C_{XY})^{*}\,(\varphi(x) - \mu_X)\big)(y),
\end{aligned}$$

where all equalities hold for $\mathbb{P}_X$-a.e. $x \in \mathcal{X}$ and the last equality follows from (4.4) (note that we might be arguing with two different choices of $h_{\psi(y)}$, which we may assume to agree by Proposition 4.2). ■

Note that step $(*)$ in the proof of Theorem 4.3 genuinely requires condition (iv) (which follows from Assumption B), and Assumption C alone does not suffice. Again we see that $\mathcal{H}$ needs to be rich enough. The reason that we get (4.2) in terms of the inner product of $L^2(\mathbb{P}_X)$, and not its weaker version in $L^2_{\mathcal{C}}(\mathbb{P}_X)$, is that we took care of the shifting constant $c_g := \mathbb{E}[f_g(X) - h_g(X)]$.

Motivated by the theory of Gaussian conditioning in Hilbert spaces [19] presented in section 6 and Theorem 6.2 in particular, we hope to generalize CMEs to the case where $\operatorname{ran} C_{XY} \subseteq \operatorname{ran} C_X$ (i.e., by Theorem 4.1, Assumption C) does not necessarily hold. As mentioned above, this will require us to work with certain finite-rank approximations of the operators C_X and C_{XY}. We are still going to need some assumption that guarantees that $\mathcal{H}$ is rich enough to be able to perform the conditioning process in the RKHSs. For this purpose Assumption B will be replaced by its weaker version, Assumption B*.

Theorem 4.4 (centered CME under finite-rank approximation). *Let Assumption 2.1 hold. Further, let $(h_n)_{n\in\mathbb{N}}$ be a complete orthonormal system of $\mathcal{H}$ that is an eigenbasis of C_X, let $\mathcal{H}^{(n)} := \operatorname{span}\{h_1, \dots, h_n\}$, let $\mathcal{F} := \mathcal{G} \oplus \mathcal{H}$, let $P^{(n)} \colon \mathcal{F} \to \mathcal{F}$ be the orthogonal projection onto $\mathcal{G} \oplus \mathcal{H}^{(n)}$, and let*

$$C := \begin{pmatrix} C_Y & C_{YX} \\ C_{XY} & C_X \end{pmatrix}, \qquad C^{(n)} := P^{(n)} C P^{(n)} = \begin{pmatrix} C_Y & C_{YX}^{(n)} \\ C_{XY}^{(n)} & C_X^{(n)} \end{pmatrix}.$$

Then $\operatorname{ran} C_{XY}^{(n)} \subseteq \operatorname{ran} C_X^{(n)}$ *and therefore* $h_g^{(n)} := C_X^{(n)\dagger} C_{XY}^{(n)} g \in \mathcal{H}$ *is well defined for each* $g \in \mathcal{G}$. *For each* $y \in \mathcal{Y}$ *and* $h \in \mathcal{H}$,

$$\langle h, \mu_{Y|X=\,\cdot\,}(y)\rangle_{L^2(\mathbb{P}_X)} = \lim_{n\to\infty} \langle h, \mu^{(n)}(\,\cdot\,, y)\rangle_{L^2(\mathbb{P}_X)}, \tag{4.5}$$

where, for $x \in \mathcal{X}$ *and* $y \in \mathcal{Y}$,

$$\mu^{(n)}(x,y) := \big(\mu_Y + (C_X^{(n)\dagger} C_{XY}^{(n)})^* \,(\varphi(x) - \mu_X)\big)(y).$$

Suppose in addition that any of the following four conditions holds:

(i) *the kernel* k *is characteristic;*
(ii) $\mathcal{H}_\mathcal{C}$ *is dense in* $L^2_\mathcal{C}(\mathbb{P}_X)$;
(iii) *Assumption* B* *holds;*
(iv) $f_{\psi(y)} \in \overline{\mathcal{H}_\mathcal{C}}^{L^2_\mathcal{C}}$ *for each* $y \in \mathcal{Y}$.

Then, as $n \to \infty$,

$$\big\|\mu^{(n)}(X,\,\cdot\,) - \mu_{Y|X}\big\|_{L^2(\mathbb{P};\mathcal{G})} \to 0, \qquad \big\|\mu_{Y|X=x} - \mu^{(n)}(x,\,\cdot\,)\big\|_{\mathcal{G}} \to 0 \textit{ for } \mathbb{P}_X\textit{-a.e. } x \in \mathcal{X}. \tag{4.6}$$

Proof. Note that, since C is a trace-class operator, so is $C^{(n)}$. Furthermore, by [2, Theorem 1], $C_{XY}^{(n)} = (C_X^{(n)})^{1/2} V C_Y^{1/2}$ for some bounded operator $V\colon \mathcal{G} \to \mathcal{H}$. Since $C_X^{(n)}$ has finite rank, this implies that $\operatorname{ran} C_{XY}^{(n)} \subseteq \operatorname{ran} C_X^{(n)}$. Similarly to the proof of Theorem 4.3, we define $c_g^{(n)} := \mathbb{E}[(f_g - h_g^{(n)})(X)]$ for $g \in \mathcal{G}, n \in \mathbb{N}$ and obtain by (2.6) and Lemma SM1.7 for $x \in \mathcal{X}$, $y \in \mathcal{Y}$, and $n \in \mathbb{N}$ that

$$h_{\psi(y)}^{(n)}(x) + c_{\psi(y)}^{(n)} = \mu^{(n)}(x,y). \tag{4.7}$$

Identity (4.5) can be obtained similarly to (4.2) except that we also need to show that $\mathbb{C}\mathrm{ov}[h(X), f_g(X)] = \lim_{n\to\infty} \mathbb{C}\mathrm{ov}[h(X), h_g^{(n)}(X)]$ for all $h \in \mathcal{H}$, as proved in Lemma SM1.9(a).

To establish (4.6), we first note that, by Lemma SM1.9(b), for all $g \in \mathcal{G}$, $[h_g^{(n)}]$ is the $L^2_\mathcal{C}$-orthogonal projection of $[f_g]$ onto $\mathcal{H}_\mathcal{C}^{(n)}$. Now let $y \in \mathcal{Y}$ and $U := \bigcup_{n\in\mathbb{N}} \mathcal{H}_\mathcal{C}^{(n)}$. Note that, by Lemma SM1.4, (i) $\implies$ (ii) $\implies$ (iii) $\implies$ (iv), so let us assume (iv). Since $\overline{U}^{\mathcal{H}_\mathcal{C}} = \mathcal{H}_\mathcal{C}$ and $[f_{\psi(y)}] \in \overline{\mathcal{H}_\mathcal{C}}^{L^2_\mathcal{C}}$ by assumption, and since (2.2) implies that $\|\cdot\|_\mathcal{H}$ is a stronger norm than $\|\cdot\|_{L^2}$, we also have $[f_{\psi(y)}] \in \overline{U}^{L^2_\mathcal{C}}$ and Lemma SM1.8 implies

$$\big\|[h_{\psi(y)}^{(n)}] - [f_{\psi(y)}]\big\|_{L^2_\mathcal{C}} \xrightarrow[n\to\infty]{} 0. \tag{4.8}$$

For $x \in \mathcal{X}$ and $n \in \mathbb{N}$ let $\mathfrak{m}^{(n)}(x) := h_{\psi(\,\cdot\,)}^{(n)}(x) = (C_X^{(n)\dagger} C_{XY}^{(n)})^* \varphi(x) \in \mathcal{G}$ and $\mathfrak{m}(x) := f_{\psi(\,\cdot\,)}(x) = \mu_{Y|X=x} \in \mathcal{G}$. Then $\mathfrak{m}^{(n)}, \mathfrak{m} \in L^2(\mathbb{P}_X;\mathcal{G})$ by (2.1), since

$$\|\mathfrak{m}^{(n)}\|^2_{L^2(\mathbb{P}_X;\mathcal{G})} = \mathbb{E}\big[\big\|(C_X^{(n)\dagger} C_{XY}^{(n)})^* \varphi(X)\big\|^2_\mathcal{G}\big] \le \big\|(C_X^{(n)\dagger} C_{XY}^{(n)})^*\big\|\, \mathbb{E}\big[\|\varphi(X)\|^2_\mathcal{H}\big] < \infty,$$
$$\|\mathfrak{m}\|^2_{L^2(\mathbb{P}_X;\mathcal{G})} = \mathbb{E}\big[\|\mathbb{E}[\psi(Y)|X]\|^2_\mathcal{G}\big] \le \mathbb{E}\big[\mathbb{E}[\|\psi(Y)\|^2_\mathcal{G}|X]\big] = \mathbb{E}\big[\|\psi(Y)\|^2_\mathcal{G}\big] < \infty.$$

So far, we have shown that, for each $y \in \mathcal{Y}$,

- $([\mathfrak{m}^{(n)}](\,\cdot\,))(y)$ is the $L^2_{\mathcal{C}}(\mathbb{P}_X)$-orthogonal projection of $([\mathfrak{m}](\,\cdot\,))(y)$ onto $\mathcal{H}_{\mathcal{C}}^{(n)}$;
- $([\mathfrak{m}^{(n)}](\,\cdot\,))(y) \to ([\mathfrak{m}](\,\cdot\,))(y)$ in $L^2_{\mathcal{C}}(\mathbb{P}_X)$ as $n \to \infty$.

Hence, by Lemma SM1.10(a) and (b),

$$\left\|(\mathfrak{m}^{(n)}(X) - \mathbb{E}[\mathfrak{m}^{(n)}(X)]) - (\mathfrak{m}(X) - \mathbb{E}[\mathfrak{m}(X)])\right\|_{L^2(\mathbb{P};\mathcal{G})} \xrightarrow[n\to\infty]{} 0. \tag{4.9}$$

Therefore, by (4.7) and the definition of $c_g^{(n)}$, $\mu^{(n)}(X,\,\cdot\,)$ converges to $\mu_{Y|X} = f_{\psi(\,\cdot\,)}(X) = \mathfrak{m}(X)$ in $L^p(\mathbb{P};\mathcal{G})$ for $p = 2$ and, since $\mathbb{P}$ is a finite measure, also for $p = 1$. By Lemma SM1.11, $(\mu^{(n)}(X,\cdot))_{n\in\mathbb{N}}$ is a martingale, and so [5, Theorem V.2.8] implies that this convergence even holds a.e., i.e., $\mu^{(n)}(x,\,\cdot\,)$ converges in $\mathcal{G}$ to $\mu_{Y|X=x}$ for $\mathbb{P}_X$-a.e. $x \in \mathcal{X}$. ■

5. Theory for uncentered operators.

Beginning with the work of [25, 26], uncentered (cross-)covariance operators became more commonly used than centered ones. This section shows how results similar to those of section 4 can be obtained for uncentered operators. Roughly speaking, the same conclusions can be made as in Theorem 4.3 but under Assumption A in place of B, while only weaker statements than in Theorem 4.4 can be obtained in Theorem 5.4 (no $\mathbb{P}_X$-a.e. convergence; see below) and again under the stronger Assumption A* in place of B*. This observation suggests that centered operators are superior to uncentered ones in terms of generality. So far, the theoretical justification for CME using uncentered operators relies on [14, Theorems 1 and 2], which require rather strong assumptions. Our improvement can be summarized as follows:

- Since we use ${}^uC_X^\dagger$ instead of ${}^uC_X^{-1}$ our theory can cope with noninjective operators ${}^uC_X$. This is only a minor advance, since ${}^uC_X$ is injective under rather mild conditions on X and k (see [14, Footnote 3]).
- In contrast to [14, Theorem 2], we do not require the assumption that $\varphi(x)$ lies in the range of ${}^uC_X$. The reason for this is that the operator $({}^uC_X^\dagger\, {}^uC_{XY})^*$ in (4.3) is globally defined, whereas ${}^uC_{YX}\, {}^uC_X^{-1}$ is not. This is an important improvement since the assumption that $\varphi(x) \in \operatorname{ran} {}^uC_X$ is typically hard to verify (see Footnote 4).
- We state a version of the CME formula under Assumption A*, which is a verifiable condition since it follows from the kernel k being L^2-universal.
- As explained in Remark 1.1, the condition in [14, Theorem 2] on $\mathbb{E}[g(Y)|X = \cdot\,]$ to lie in $\mathcal{H}$ for each $g \in \mathcal{G}$ is ill posed, since these functions are uniquely defined only $\mathbb{P}_X$-a.e. However, in our case, Assumption 2.1(f) ensures that Assumptions A and A* are unambiguous.

As mentioned above, using centered operators instead of uncentered ones yields the important advantage of requiring only the weaker Assumption B in place of A or Assumption B* in place of A*, respectively. Further, Theorem 5.4 provides weaker statements than its centered analogue, Theorem 4.4: we show only convergence in $L^2(\mathbb{P}_X;\mathcal{G})$, which does not guarantee convergence for $\mathbb{P}_X$-a.e. $x \in \mathcal{X}$.

Theorem 5.1. *Under Assumption 2.1, the following statements are equivalent:*

(i) *Assumption ^{u}C holds;*

(ii) *for each $g \in \mathcal{G}$ there exists $h_g \in \mathcal{H}$ such that ${}^uC_X h_g = {}^uC_{XY} g$;*

(iii) $\operatorname{ran} {}^uC_{XY} \subseteq \operatorname{ran} {}^uC_X$.

Proof. The proof is identical to that of Theorem 4.1 (apart from using uncentered covariance operators in place of centered ones). ■

Similar to Proposition 4.2, the element $h_g \in \mathcal{H}$ in Assumption ^{u}C can always be chosen as $h_g = {}^uC_X^\dagger\, {}^uC_{XY} g$.

Proposition 5.2. *Let Assumption 2.1 hold. Under Assumption ^{u}C, h_g may be chosen as*

$$h_g = {}^uC_X^\dagger\, {}^uC_{XY} g. \tag{5.1}$$

More precisely, ${}^u\mathbb{C}\mathrm{ov}[({}^uC_X^\dagger\, {}^uC_{XY} g)(X) - f_g(X), h(X)] = 0$ for all $h \in \mathcal{H}$ and $g \in \mathcal{G}$. If $f_g \in \mathcal{H}$ for some $g \in \mathcal{G}$, then the identity $f_g = {}^uC_X^\dagger\, {}^uC_{XY} g$ holds $\mathbb{P}_X$-a.e.

Proof. By Theorem 5.1, (5.1) is well defined. If Assumption ^{u}C holds, then, by Theorem 5.1 and Lemma SM1.6, for all $g \in \mathcal{G}$ and $h \in \mathcal{H}$,

$$\begin{aligned}
{}^u\mathbb{C}\mathrm{ov}[h(X), ({}^uC_X^\dagger\, {}^uC_{XY} g)(X)] &= \langle h, {}^uC_X\, {}^uC_X^\dagger\, {}^uC_{XY} g\rangle_{\mathcal{H}} \\
&= \langle h, {}^uC_{XY} g\rangle_{\mathcal{H}} = {}^u\mathbb{C}\mathrm{ov}[h(X), f_g(X)].
\end{aligned}$$

If $f_g \in \mathcal{H}$ holds for some $g \in \mathcal{G}$, then Lemma SM1.6 implies that ${}^uC_X f_g = {}^uC_{XY} g$ for all $g \in \mathcal{G}$ and the claim follows from Lemma SM1.5. ■

Let us now formulate and prove the analogues of Theorems 4.3 and 4.4 for uncentered operators.

Theorem 5.3 (uncentered CME). *Under Assumptions 2.1 and ^{u}C, the linear operator ${}^uC_X^\dagger\, {}^uC_{XY} \colon \mathcal{G} \to \mathcal{H}$ is bounded (see footnote 10) and, for all $y \in \mathcal{Y}$ and $h \in \mathcal{H}$,*

$$\langle h, \mu_{Y|X=\,\cdot\,}(y)\rangle_{L^2(\mathbb{P}_X)} = \left\langle h, \big(({}^uC_X^\dagger\, {}^uC_{XY})^* \varphi(\,\cdot\,)\big)(y)\right\rangle_{L^2(\mathbb{P}_X)}. \tag{5.2}$$

Suppose in addition that any of the following four conditions holds:

(i) *the kernel k is L^2-universal;*
(ii) *$\mathcal{H}$ is dense in $L^2(\mathbb{P}_X)$;*
(iii) *Assumption A holds;*
(iv) *$f_{\psi(y)} \in \mathcal{H}$ for each $y \in \mathcal{Y}$.*

Then, for $\mathbb{P}_X$-a.e. $x \in \mathcal{X}$,

$$\mu_{Y|X=x} = ({}^uC_X^\dagger\, {}^uC_{XY})^*\, \varphi(x). \tag{5.3}$$

Proof. First note that, by Theorems SM1.1 and 5.1, ${}^uC_X^\dagger\, {}^uC_{XY}$ is well defined and bounded (see footnote 10) and that for each $g \in \mathcal{G}$ we may choose the function $h_g \in \mathcal{H}$ in Assumption ^{u}C as $h_g = {}^uC_X^\dagger\, {}^uC_{XY} g$ by Proposition 5.2. By Lemma SM1.7 we obtain, for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$,

$$h_{\psi(y)}(x) = \big(({}^uC_X^\dagger\, {}^uC_{XY})^* \varphi(x)\big)(y).$$

This yields (5.2) via

$$\begin{aligned}
\left\langle h, \big(\mu_{Y|X=\,\cdot\,} - ({}^uC_X^\dagger\, {}^uC_{XY})^*\, \varphi(\,\cdot\,)\big)(y)\right\rangle_{L^2(\mathbb{P}_X)} &= \langle h, f_{\psi(y)} - h_{\psi(y)}\rangle_{L^2(\mathbb{P}_X)} \\
&= {}^u\mathbb{C}\mathrm{ov}[h(X), (f_{\psi(y)} - h_{\psi(y)})(X)] = 0,
\end{aligned}$$

which implies (5.3) under any of the four conditions stated in the theorem (possibly using Proposition 5.2). ■

Theorem 5.4 (uncentered CME under finite-rank approximation). *Let Assumption 2.1 hold. Further, let $(h_n)_{n\in\mathbb{N}}$ be a complete orthonormal system of $\mathcal{H}$ that is an eigenbasis of C_X, let $\mathcal{H}^{(n)} := \operatorname{span}\{h_1,\dots,h_n\}$, let $\mathcal{F} := \mathcal{G}\oplus\mathcal{H}$, let $P^{(n)}\colon \mathcal{F}\to\mathcal{F}$ be the orthogonal projection onto $\mathcal{G}\oplus\mathcal{H}^{(n)}$, and let*

$$ {}^{u}C := \begin{pmatrix} {}^{u}C_Y & {}^{u}C_{YX} \\ {}^{u}C_{XY} & {}^{u}C_X \end{pmatrix}, \qquad {}^{u}C^{(n)} := P^{(n)}\,{}^{u}C P^{(n)} = \begin{pmatrix} {}^{u}C_Y & {}^{u}C_{YX}^{(n)} \\ {}^{u}C_{XY}^{(n)} & {}^{u}C_X^{(n)} \end{pmatrix}. $$

Then $\operatorname{ran}{}^{u}C_{XY}^{(n)} \subseteq \operatorname{ran}{}^{u}C_X^{(n)}$ and therefore ${}^{u}h_g^{(n)} := {}^{u}C_X^{(n)\dagger}\,{}^{u}C_{XY}^{(n)}g \in \mathcal{H}$ is well defined for each $g\in\mathcal{G}$. For each $y\in\mathcal{Y}$ and $h\in\mathcal{H}$,

$$ \langle h, \mu_{Y|X=\,\cdot\,}(y)\rangle_{L^2(\mathbb{P}_X)} = \lim_{n\to\infty} \langle h, \mu^{(n)}(\,\cdot\,,y)\rangle_{L^2(\mathbb{P}_X)}, \tag{5.4} $$

where, for $x\in\mathcal{X}$ and $y\in\mathcal{Y}$,

$$ {}^{u}\mu^{(n)}(x,y) := \big(({}^{u}C_X^{(n)\dagger}\,{}^{u}C_{XY}^{(n)})^{*}\varphi(x)\big)(y). \tag{5.5} $$

Suppose in addition that any of the following four conditions holds:

(i) *the kernel k is L^2-universal;*
(ii) *$\mathcal{H}$ is dense in $L^2(\mathbb{P}_X)$;*
(iii) *Assumption A* holds;*
(iv) *$f_{\psi(y)} \in \overline{\mathcal{H}}^{L^2}$ for each $y\in\mathcal{Y}$.*

Then

$$ \big\| {}^{u}\mu^{(n)}(X,\,\cdot\,) - \mu_{Y|X} \big\|_{L^2(\mathbb{P};\mathcal{G})} \xrightarrow[n\to\infty]{} 0. \tag{5.6} $$

Proof. The proof is analogous to that of Theorem 4.4 up to (4.9), using uncentered operators instead of centered ones and the statements (c), (d) instead of (a), (b) of Lemmas SM1.9 and SM1.10. However, we cannot draw the final conclusion of convergence almost everywhere since we do not have the martingale property, which is provided by Lemma SM1.11 for the centered case. Note that our proof relies on [2, Theorem 1], which, strictly speaking, only treats the centered case, but its uncentered version can be proven similarly. ■

Corollary 5.5. *Under the assumptions of Theorem 5.3 (including either of the additional ones),*

$$ \mu_Y = ({}^{u}C_X^{\dagger}\,{}^{u}C_{XY})^{*}\mu_X. $$

Under the assumptions of Theorem 5.4 (including either of the additional ones),

$$ \big\| \mu_Y - ({}^{u}C_X^{(n)\dagger}\,{}^{u}C_{XY}^{(n)})^{*}\mu_X \big\|_{\mathcal{G}} \xrightarrow[n\to\infty]{} 0. $$

Proof. As stated in Theorem 5.3, ${}^{u}C_X^{\dagger}\,{}^{u}C_{XY}$ is a well-defined and bounded (see footnote 10) linear operator. Hence, by the law of total expectation and Theorem 5.3,

$$ \mu_Y = \mathbb{E}[\mu_{Y|X}] = \mathbb{E}\big[({}^{u}C_X^{\dagger}\,{}^{u}C_{XY})^{*}\varphi(X)\big] = ({}^{u}C_X^{\dagger}\,{}^{u}C_{XY})^{*}\mathbb{E}[\varphi(X)] = ({}^{u}C_X^{\dagger}\,{}^{u}C_{XY})^{*}\mu_X, $$

proving the first claim. The second one follows from Jensen's inequality and Theorem 5.4 via

$$\begin{aligned}\big\|\mu_Y - ({}^u C_X^{(n)\dagger}\, {}^u C_{XY}^{(n)})^* \mu_X\big\|_{\mathcal{G}}^2 &= \big\|\mathbb{E}[\mu_{Y|X}] - ({}^u C_X^{(n)\dagger}\, {}^u C_{XY}^{(n)})^*\, \mathbb{E}[\varphi(X)]\big\|_{\mathcal{G}}^2 \\ &= \big\|\mathbb{E}\big[\mu_{Y|X} - ({}^u C_X^{(n)\dagger}\, {}^u C_{XY}^{(n)})^* \varphi(X)\big]\big\|_{\mathcal{G}}^2 \\ &\le \big\|\mu_{Y|X} - ({}^u C_X^{(n)\dagger}\, {}^u C_{XY}^{(n)})^* \varphi(X)\big\|_{L^2(\mathbb{P}_X;\mathcal{G})}^2 \xrightarrow[n\to\infty]{} 0. \qquad \blacksquare\end{aligned}$$

6. Gaussian conditioning in Hilbert spaces.

This section gives a review of conditioning theory for Gaussian random variables in separable Hilbert spaces, summarizing the work of [19]. Our only somewhat novel contribution here is the explicit characterization of the essential operator $\widehat{Q}_{C,\mathcal{H}}$ in terms of the Moore–Penrose pseudoinverse, which appears as an exercise for the reader in [1, Remark 2.3].

In the following let $\mathcal{F} = \mathcal{G} \oplus \mathcal{H}$ be the sum of two separable Hilbert spaces $\mathcal{G}$ and $\mathcal{H}$, and let (U, V) be an $\mathcal{F}$-valued jointly Gaussian random variable with mean $\mu \in \mathcal{F}$ and covariance operator $C\colon \mathcal{F} \to \mathcal{F}$ given by the following block structures:

$$\begin{pmatrix} U \\ V \end{pmatrix} \sim \mathcal{N}(\mu, C), \qquad \mu = \begin{pmatrix} \mu_U \\ \mu_V \end{pmatrix}, \qquad C = \begin{pmatrix} C_U & C_{UV} \\ C_{VU} & C_V \end{pmatrix} \ge 0$$

with $\mu_U \in \mathcal{G}$, etc. We denote by $L(\mathcal{F})$ the Banach algebra of bounded linear operators on $\mathcal{F}$ and by $L_+(\mathcal{F}) = \{A \in L(\mathcal{F}) \mid A \ge 0\}$ the set of positive operators, i.e., those self-adjoint operators A for which $\langle x, Ax\rangle \ge 0$ for all $x \in \mathcal{F}$. The theory of Gaussian conditioning relies on the concept of so-called *oblique projections*.

Definition 6.1. *Let $\mathcal{F} = \mathcal{G} \oplus \mathcal{H}$ be a direct sum of two Hilbert spaces $\mathcal{G}$ and $\mathcal{H}$, and let $C \in L_+(\mathcal{F})$ be a positive operator. The set of (C-symmetric)* oblique projections *onto $\mathcal{H}$ is given by*

$$\mathcal{P}(C, \mathcal{H}) = \{Q \in L(\mathcal{F}) \mid Q^2 = Q,\ \operatorname{ran} Q = \mathcal{H},\ CQ = Q^*C\}.$$

The pair $(C, \mathcal{H})$ is said to be compatible *if $\mathcal{P}(C, \mathcal{H})$ is nonempty.*

The first two conditions, $Q^2 = Q$ and $\operatorname{ran} Q = \mathcal{H}$, imply that Q has the block structure

$$Q = \begin{pmatrix} 0 & 0 \\ \widehat{Q} & \mathrm{Id}_{\mathcal{H}} \end{pmatrix}, \qquad \widehat{Q}\colon \mathcal{G} \to \mathcal{H}. \tag{6.1}$$

Then the condition $CQ = Q^*C$ is equivalent to $C_V\widehat{Q} = C_{VU}$ (which follows from a straightforward blockwise multiplication; see Lemma 6.3) and implies in particular $\operatorname{ran} C_{VU} \subseteq \operatorname{ran} C_V$. The other way round, as we will see later on, the condition $\operatorname{ran} C_{VU} \subseteq \operatorname{ran} C_V$ guarantees the existence of an oblique projection $Q \in \mathcal{P}(C, \mathcal{H})$ and will provide a crucial link between the theory of Gaussian conditioning and CMEs in section 7.

The results on conditioning Gaussian measures can then be summarized as follows.

Theorem 6.2 (see [19, Theorem 3.3, Corollary 3.4]). *If $(C, \mathcal{H})$ is compatible, then conditioning U on $V = v \in \mathcal{H}$ results in a Gaussian random variable on $\mathcal{G}$ with mean $\mu_{U|V=v}$ and covariance operator $C_{U|V=v}$ given by*

$$\begin{cases} \mu_{U|V=v} = \mu_U + \widehat{Q}^*(v - \mu_V), \\ C_{U|V=v} = C_U - C_{UV}\widehat{Q} \end{cases} \tag{6.2}$$

for any oblique projection $Q \in \mathcal{P}(C, \mathcal{H})$ given in the form (6.1). Also, in this case, $\mathcal{P}(C, \mathcal{H})$ contains a unique element

$$Q_{C,\mathcal{H}} = \begin{pmatrix} 0 & 0 \\ \widehat{Q}_{C,\mathcal{H}} & \mathrm{Id}_{\mathcal{H}} \end{pmatrix}$$

that fulfills the properties (6.4) defined below.

If $(C, \mathcal{H})$ is incompatible, then conditioning U on $V = v \in \mathcal{H}$ still yields a Gaussian random variable on $\mathcal{G}$, but the corresponding formulae for the conditional mean $\mu_{U|V=v}$ and covariance operator $C_{U|V=v}$ are given by a limiting process using finite-rank approximations of C in the following way. Let $(h_n)_{n\in\mathbb{N}}$ be a complete orthonormal system of $\mathcal{H}$, let $P^{(n)} \colon \mathcal{F} \to \mathcal{F}$ denote the orthogonal projection onto $\mathcal{G} \oplus \operatorname{span}\{h_1, \dots, h_n\}$, and let $C^{(n)} = P^{(n)} C P^{(n)}$. Then $(C^{(n)}, \mathcal{H})$ is compatible for each $n \in \mathbb{N}$ and, for $\mathbb{P}_V$-a.e. $v \in \mathcal{H}$ (with $\mathbb{P}_V$ denoting the distribution of V),

$$\begin{cases} \mu_{U|V=v} = \mu_U + \lim_{n\to\infty} \widehat{Q}^{*}_{C^{(n)},\mathcal{H}}(v - \mu_V), \\ C_{U|V=v} = C_U - \lim_{n\to\infty} C_{UV} \widehat{Q}_{C^{(n)},\mathcal{H}}, \end{cases} \tag{6.3}$$

where the second limit is in the trace norm.

In the following we will revisit some theory on oblique projections which will be necessary to establish the connection between Gaussian conditioning and CMEs. We will also characterize the special oblique projection $Q_{C,\mathcal{H}} \in \mathcal{P}(C, \mathcal{H})$ by means of the Moore–Penrose pseudoinverse.

Lemma 6.3. *If $\widehat{Q} \colon \mathcal{G} \to \mathcal{H}$ is a bounded linear operator such that $C_V \widehat{Q} = C_{VU}$, then*

$$Q = \begin{pmatrix} 0 & 0 \\ \widehat{Q} & \mathrm{Id}_{\mathcal{H}} \end{pmatrix} \in \mathcal{P}(C, \mathcal{H}).$$

In particular, the pair $(C, \mathcal{H})$ is compatible.

Proof. The properties $Q^2 = Q$ and $\operatorname{ran} Q = \mathcal{H}$ are clear from the definition of Q, and a straightforward blockwise multiplication shows that $CQ = Q^* C$. ■

Proposition 6.4. *In the setup of Definition 6.1, if $(C, \mathcal{H})$ is compatible, then there exists a unique bounded operator $\widehat{Q}_{C,\mathcal{H}} \colon \mathcal{G} \to \mathcal{H}$ such that*

$$C_V \widehat{Q}_{C,\mathcal{H}} = C_{VU}, \qquad \ker \widehat{Q}_{C,\mathcal{H}} = \ker C_{VU}, \qquad \operatorname{ran} \widehat{Q}_{C,\mathcal{H}} \subseteq \overline{\operatorname{ran} C_V}. \tag{6.4}$$

By Lemma 6.3 the first property implies that

$$Q_{C,\mathcal{H}} = \begin{pmatrix} 0 & 0 \\ \widehat{Q}_{C,\mathcal{H}} & \mathrm{Id}_{\mathcal{H}} \end{pmatrix} \in \mathcal{P}(C, \mathcal{H}).$$

Proof. See [6, Theorem 1] or [8, Theorem 2.1] for the existence and uniqueness of $\widehat{Q}_{C,\mathcal{H}}$ and [4] or [19] for its connection to oblique projections. ■

If one follows the original construction of [6, Theorem 1] or [8, Theorem 2.1], it is easy to see how this unique element can be characterized in terms of the Moore–Penrose pseudoinverse $C_V^{\dagger}$ of C_V.

Theorem 6.5. *If* $\operatorname{ran} C_{VU} \subseteq \operatorname{ran} C_V$*, then* $\widehat{Q} = C_V^\dagger C_{VU} \colon \mathcal{G} \to \mathcal{H}$ *is a well-defined bounded operator which uniquely fulfills the conditions* (6.4)*.*

Proof. This is a direct application of Theorem SM1.1. ■

Theorem 6.6. *In the setup of Definition 6.1, the following statements are equivalent:*
(i) $(C, \mathcal{H})$ *is compatible;*
(ii) $\operatorname{ran} C_{VU} \subseteq \operatorname{ran} C_V$*.*
If either of these conditions holds, then the unique element $Q_{C,\mathcal{H}} \in \mathcal{P}(C, \mathcal{H})$ *in Proposition 6.4 is given by*

$$\widehat{Q}_{C,\mathcal{H}} = C_V^\dagger C_{VU}. \tag{6.5}$$

Proof. If $(C, \mathcal{H})$ is compatible, there exists an element $\widehat{Q}_{C,\mathcal{H}} \colon \mathcal{G} \to \mathcal{H}$ with $C_V \widehat{Q}_{C,\mathcal{H}} = C_{VU}$ by Proposition 6.4, which implies (ii). If $\operatorname{ran} C_{VU} \subseteq \operatorname{ran} C_V$, then Theorem 6.5 and Lemma 6.3 imply (i). Theorem 6.5 and the uniqueness of $\widehat{Q}_{C,\mathcal{H}}$ in Proposition 6.4 imply (6.5). ■

Remark 6.7. Lemma 6.3 and the equivalence part of Theorem 6.6 were already proved by [4]; we state them for the sake of readability. The second part of Theorem 6.6(ii) characterizes the operator $\widehat{Q}_{C,\mathcal{H}}$ in terms of the Moore–Penrose pseudoinverse, without an assumption of closed range, as anticipated by [1, Remark 2.3].

There are covariance operators C for which the above conditions do not hold.

Example 6.8. Let $\mathcal{H} = \mathcal{G}$ be any (separable) infinite-dimensional Hilbert space with complete orthonormal basis $(e_j)_{j\in\mathbb{N}}$. Let

$$\begin{aligned} C_U &:= \sum_{j\in\mathbb{N}} j^{-2} e_j \otimes e_j, \\ C_V &:= \sum_{j\in\mathbb{N}} j^{-4} e_j \otimes e_j, \\ C_{VU} = C_{UV} &:= C_U^{1/2} \operatorname{Id}_{\mathcal{H}} C_V^{1/2} = \sum_{j\in\mathbb{N}} j^{-3} e_j \otimes e_j. \end{aligned}$$

By [2, Theorem 2],

$$C := \begin{pmatrix} C_U & C_{UV} \\ C_{VU} & C_V \end{pmatrix}$$

is a legitimate positive definite covariance operator on $\mathcal{F} = \mathcal{G} \oplus \mathcal{H}$. However,

$$\operatorname{ran} C_{VU} = \left\{ \sum_{j\in\mathbb{N}} \alpha_j e_j \,\middle|\, (j^3\alpha_j)_{j\in\mathbb{N}} \in \ell^2 \right\} \not\subseteq \left\{ \sum_{j\in\mathbb{N}} \alpha_j e_j \,\middle|\, (j^4\alpha_j)_{j\in\mathbb{N}} \in \ell^2 \right\} = \operatorname{ran} C_V.$$

7. Connection between CME and Gaussian conditioning.

If we compare the theories of CMEs and Gaussian conditioning in Hilbert spaces, we make the following observations:

- Formula (4.3) for CME and formula (6.2) for Gaussian conditioning look very similar (in view of Theorem 6.6).

- The assumptions under which the conditioning process is "easy"—namely, Assumption C (as long as Assumption B* holds as well) and the compatibility of $(C, \mathcal{H})$—are equivalent to the conditions that $\operatorname{ran} C_{XY} \subseteq \operatorname{ran} C_X$ and $\operatorname{ran} C_{VU} \subseteq \operatorname{ran} C_V$, respectively (Theorems 4.1 and 6.6).

This motivates us to connect these two theories by working in the setup of section 2 and introducing new jointly Gaussian random variables U and V that take values in the RKHSs $\mathcal{G}$ and $\mathcal{H}$, respectively, where the means μ_U and μ_V and (cross-)covariance operators C_U, C_{UV}, C_{VU}, and C_V are chosen to coincide with the kernel mean embeddings μ_Y and μ_X and the kernel (cross-)covariance operators C_Y, C_{YX}, C_{XY}, and C_X, respectively:

$$(7.1)\quad \begin{pmatrix} U \\ V \end{pmatrix} \sim \mathcal{N}(\mu, C), \quad \mu = \begin{pmatrix} \mu_U \\ \mu_V \end{pmatrix} := \begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix}, \quad C = \begin{pmatrix} C_U & C_{UV} \\ C_{VU} & C_V \end{pmatrix} := \begin{pmatrix} C_Y & C_{YX} \\ C_{XY} & C_X \end{pmatrix}.$$

By [2, Theorem 1] and since Assumption 2.1(e) implies that C is a trace-class covariance operator, the Gaussian random variable (U, V) is well defined in $\mathcal{G} \oplus \mathcal{H}$. Note that the random variables $W = (U, V)$ and $Z = (\psi(Y), \varphi(X))$ do not coincide even though they have the same mean and covariance operator, since the latter will not generally be Gaussian. Surprisingly, their conditional means agree, as long as we condition on $V = v = \varphi(x)$ and $X = x$, respectively. This is obvious when one compares (4.3) with (6.2) (and (4.6) with (6.3) using Theorem 6.5). A natural question is whether a similar equality holds for the conditional covariance operator $C_{Y|X=x}$. However, the covariance operator $C_{U|V=v}$ obtained from Gaussian conditioning is independent of v, a special property of Gaussian measures that cannot be expected of the conditional kernel covariance operator $C_{Y|X=x}$. Instead, $C_{U|V=v}$ equals the mean of $C_{Y|X=x}$ when averaged over all possible outcomes $x \in \mathcal{X}$.[11] These insights are summarized in the following proposition and illustrated in Figure 7.1.

Note that the distributions of $\varphi(X)$ and V might have different (and even disjoint!) supports, and so one must be particularly careful with "almost every" statements in this context.

Theorem 7.1. *Let Assumptions 2.1 and B* hold, let (U, V) be the random variable defined by (7.1), and let $\mathbb{P}_{\varphi(X)}$ and $\mathbb{P}_V$ denote the probability distributions of $\varphi(X)$ and V, respectively. Then, for $\mathbb{P}_V$-a.e. $v \in \mathcal{H}$,*

$$C_{U|V=v} = \mathbb{E}[C_{Y|X}] = \int_{\mathcal{X}} C_{Y|X=x} \, \mathrm{d}\mathbb{P}_X(x).$$

Further, there exist $N_1, N_2 \subseteq \Omega$ with $\mathbb{P}_{\varphi(X)}(N_1) = 0$ and $\mathbb{P}_V(N_2) = 0$, such that, for every $v = \varphi(x) \notin N_1 \cup N_2$, $\mu_{U|V=v} = \mu_{Y|X=x}$.

Proof. By Lemma SM1.12, $\mathbb{E}[C_{Y|X}]$ is well defined. The identity $\mu_{U|V=v} = \mu_{Y|X=x}$ for the means follows directly from Theorems 4.4, 6.2, and 6.5. For the covariance identity, using the notation of Theorem 4.4, note that $\|[h_{\psi(y)}^{(n)}] - [f_{\psi(y)}]\|_{L^2_{\mathbb{C}}} \xrightarrow[n\to\infty]{} 0$ by (4.8). Therefore, for

[11] This observation has already been made by [10, Proposition 5] under stronger assumptions and by [12, Proposition 3] in a weaker form.

$$\begin{tikzcd}
\mathcal{X},\mathcal{Y} & \mathcal{H},\mathcal{G} & \text{Gaussian on } \mathcal{G}\oplus\mathcal{H} \\
\left\{\begin{matrix} x\in\mathcal{X} \\ X\sim\mathbb{P}_X \\ Y\sim\mathbb{P}_Y \\ (X,Y)\sim\mathbb{P}_{XY}\end{matrix}\right\} \arrow[r, "\text{embed}", "\psi,\varphi"'] \arrow[d, "\text{conditioning on}", "x=X"'] & \left\{\begin{matrix} \varphi(x) \\ \psi(Y),\varphi(X) \\ \mu_Y, C_Y, C_{YX} \\ \mu_X, C_{XY}, C_X \end{matrix}\right\} \arrow[r, leftrightarrow] \arrow[d, "\text{conditional mean}", "\text{embedding}"'] & \begin{pmatrix} U \\ V \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}\mu_Y \\ \mu_X\end{pmatrix}, \begin{pmatrix} C_Y & C_{YX} \\ C_{XY} & C_X \end{pmatrix}\right) \arrow[d, "\text{conditioning on}", "V=v=\varphi(x)"'] \\
(Y|X=x)\sim\mathbb{P}_{Y|X=x} \arrow[r, "\text{embed}", "\psi,\varphi"'] & \mu_{Y|X=x}, C_{Y|X=x} \arrow[r, leftrightarrow] & (U|V=v)\sim\mathcal{N}(\mu_{U|V=v}, C_{U|V=v})
\end{tikzcd}$$

Figure 7.1. *A normally distributed $\mathcal{G}\oplus\mathcal{H}$-valued normal random variable (U,V) can be defined with the same mean and covariance structure as $(\psi(Y),\varphi(X))$. While the latter will typically fail to be normally distributed, surprisingly, the conditional means of the two random variables happen to agree! Since $C_{U|V=v}$ does not depend on the realization v, a specific property of Gaussian random variables that cannot be expected from $C_{Y|X=x}$, a similar agreement for the conditional covariance operators cannot be obtained. Instead, the identity provided by Theorem 7.1 holds, which is open to interpretation.*

$y, y' \in \mathcal{Y}$, $g = \psi(y)$, and $g' = \psi(y')$,

$$\begin{aligned}
\mathbb{C}\mathrm{ov}\left[f_g(X), f_{g'}(X)\right] &= \lim_{n\to\infty} \mathbb{C}\mathrm{ov}\left[h_g^{(n)}(X), h_{g'}^{(n)}(X)\right] \\
&= \lim_{n\to\infty} \langle C_X h_g^{(n)}, h_{g'}^{(n)}\rangle_{\mathcal{H}} \\
&= \lim_{n\to\infty} \langle C_{XY}^{(n)} g, C_X^{(n)\dagger} C_{XY}^{(n)} g'\rangle_{\mathcal{H}} \\
&= \lim_{n\to\infty} \langle g, C_{YX}^{(n)} C_X^{(n)\dagger} C_{XY}^{(n)} g'\rangle_{\mathcal{G}} \\
&= \lim_{n\to\infty} \langle g, C_{UV} C_V^{(n)\dagger} C_{VU}^{(n)} g'\rangle_{\mathcal{G}}.
\end{aligned}$$

By the law of total covariance and (6.3), (6.5) this implies that, for $g = \psi(y)$ and $g' = \psi(y')$,

$$\begin{aligned}
\langle g, \mathbb{E}[C_{Y|X}] g'\rangle_{\mathcal{G}} &= \mathbb{E}\big[\mathbb{C}\mathrm{ov}[g(Y), g'(Y)|X]\big] \\
&= \mathbb{C}\mathrm{ov}[g(Y), g'(Y)] - \mathbb{C}\mathrm{ov}\big[f_g(X), f_{g'}(X)\big] \\
&= \langle g, C_U g'\rangle_{\mathcal{G}} - \lim_{n\to\infty} \langle g, C_{UV} C_V^{(n)\dagger} C_{VU}^{(n)} g'\rangle_{\mathcal{G}} \\
&= \langle g, C_{U|V=v}\, g'\rangle_{\mathcal{G}}
\end{aligned}$$

for $\mathbb{P}_V$-a.e. $v \in \mathcal{H}$. Since $\operatorname{span}\{\psi(y) \mid y \in \mathcal{Y}\}$ is dense in $\mathcal{G}$, this finishes the proof. ■

Remark 7.2. Theorem 7.1 implies in particular that the posterior mean $\mu_{U|V=v}$ of the U-component of a jointly Gaussian random variable (U,V) in an RKHS $\mathcal{G}\oplus\mathcal{H}$ is not just some element in $\mathcal{G}$, but in fact the KME of some probability distribution on $\mathcal{Y}$, as long as we condition on an event of the form $V = v = \varphi(x)$ outside the null events N_1 and N_2. Note, though, that these null sets could be geometrically quite large.

As mentioned above, there is another analogy between CMEs and Gaussian conditioning, namely, the assumption under which the formula for the conditional mean is particularly nice, i.e., does not require finite-rank approximations of the (cross-)covariance operators.

Theorem 7.3. *Under Assumption 2.1 and with the random variable* (U, V) *defined by* (7.1), *Assumption* C *is equivalent to the compatibility of* $(C, \mathcal{H})$.

Proof. By Theorems 4.1 and 6.6, both conditions are equivalent to $\operatorname{ran} C_{XY} \subseteq \operatorname{ran} C_X$. ■

8. Closing remarks.

This article has demonstrated rigorous foundations for the method of conditional mean embedding in reproducing kernel Hilbert spaces. Mild and verifiable sufficient conditions have been provided for the centered and uncentered variants of the CME formula to yield an element $\mu_{Y|X=x}$ that is indeed the kernel mean embedding of the conditional distribution $\mathbb{P}_{Y|X=x}$ on $\mathcal{Y}$. The CME formula required a correction in the centered case but, modulo this correction, it is more generally applicable than its uncentered counterpart and provides stronger statements: Theorem 4.4 proves convergence in $L^2(\mathbb{P}; \mathcal{G})$ as well as $\mathbb{P}_X$-a.e. convergence, while its analogue Theorem 5.4 yields only convergence in $L^2(\mathbb{P}; \mathcal{G})$. The reason is that $({}^u\mu^{(n)}(X, \cdot))_{n\in\mathbb{N}}$ defined by (5.5), in contrast to $(\mu^{(n)}(X, \cdot))_{n\in\mathbb{N}}$, may fail to be a martingale (cf. Lemma SM1.11), and we cannot apply [5, Theorem V.2.8]. Therefore, we advocate for the centered version of the CME formula as the preferred formulation in practice. We have also demonstrated the precise relationship between CMEs and well-established formulae for the conditioning of Gaussian random variables in Hilbert spaces.

Some natural directions for further research suggest themselves:

First, in practice, the KMEs and kernel (cross-)covariance operators will often be estimated using sampled data, and so empirical versions of the CME, along with convergence guarantees, are of great practical importance. Various empirical CMEs have already been considered and applied in the literature [9, 14, 16, 20], but their approximation accuracy is not at all trivial to analyze, conditions for validity along the lines of our Assumptions A–^{u}C are not yet known, and a detailed treatment would be too long to consider in this work, which has deliberately focused on the population CME. Section SM2 gives an overview of the technical obstacles that must be overcome in the empirical setting, existing results in the area, and work yet to do.

Second, when using CMEs for inference, a remaining step might be to undo the kernel mean embedding, i.e., to recover the conditional distribution $\mathbb{P}_{Y|X=x}$ on $\mathcal{Y}$ from its embedding $\mu_{Y|X=x} \in \mathcal{G}$, or its density with respect to a reference measure on $\mathcal{Y}$. This is a particular instance of a nonparametric inverse problem, and a principled solution, based upon Tikhonov regularization, has been proposed in the context of the *kernel conditional density operator* (KCDO) by [23]. The relationship between this KCDO approach and the sufficient conditions for CME that have been considered in this article remains to be precisely formulated; given the intimate relationship between Tikhonov regularization and the Moore–Penrose pseudoinverse, this should be a fruitful avenue of research.

Acknowledgments. The authors wish to thank S. Klus, H. C. Lie, M. Mollenhauer, and B. Sprungk for helpful and collegial discussions.

REFERENCES

[1] M. L. Arias, G. Corach, and M. C. Gonzalez, *Generalized inverses and Douglas equations*, Proc. Amer. Math. Soc., 136 (2008), pp. 3177–3183, https://doi.org/10.1090/S0002-9939-08-09298-8.

[2] C. R. Baker, *Joint measures and cross-covariance operators*, Trans. Amer. Math. Soc., 186 (1973), pp. 273–289, https://doi.org/10.2307/1996566.

[3] A. Berlinet and C. Thomas-Agnan, *Reproducing Kernel Hilbert Spaces in Probability and Statistics*, Springer, Boston, 2004, https://doi.org/10.1007/978-1-4419-9096-9.

[4] G. Corach, A. Maestripieri, and D. Stojanoff, *Oblique projections and Schur complements*, Acta Sci. Math. (Szeged), 67 (2001), pp. 337–356.

[5] J. Diestel and J. J. Uhl, *Vector Measures*, Math. Surveys 15, AMS, Providence, RI, 1977, https://doi.org/10.1090/surv/015.

[6] R. G. Douglas, *On majorization, factorization, and range inclusion of operators on Hilbert space*, Proc. Amer. Math. Soc., 17 (1966), pp. 413–415, https://doi.org/10.2307/2035178.

[7] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Math. Appl. 375, Kluwer Academic, Dordrecht, 1996.

[8] P. A. Fillmore and J. P. Williams, *On operator ranges*, Adv. Math., 7 (1971), pp. 254–281, https://doi.org/10.1016/S0001-8708(71)80006-3.

[9] K. Fukumizu, *Nonparametric Bayesian inference with kernel mean embedding*, in Modern Methodology and Applications in Spatial-Temporal Modeling, Springer Japan, Tokyo, 2015, pp. 1–24, https://doi.org/10.1007/978-4-431-55339-7_1.

[10] K. Fukumizu, F. R. Bach, and M. I. Jordan, *Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces*, J. Mach. Learn. Res., 5 (2004), pp. 73–99, http://www.jmlr.org/papers/volume5/fukumizu04a/fukumizu04a.pdf.

[11] K. Fukumizu, F. R. Bach, and M. I. Jordan, *Erratum: Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces*, J. Mach. Learn. Res., (2004), http://www.jmlr.org/papers/volume5/fukumizu04a/fukumizu04a-erratum.pdf.

[12] K. Fukumizu, F. R. Bach, and M. I. Jordan, *Kernel dimension reduction in regression*, Ann. Statist., 37 (2009), pp. 1871–1905, https://doi.org/10.1214/08-AOS637.

[13] K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf, *Kernel measures of conditional dependence*, in Advances in Neural Information Processing Systems 20, Curran Associates, 2008, pp. 489–496, https://papers.nips.cc/paper/3340-kernel-measures-of-conditional-dependence.pdf.

[14] K. Fukumizu, L. Song, and A. Gretton, *Kernel Bayes' rule: Bayesian inference with positive definite kernels*, J. Mach. Learn. Res., 14 (2013), pp. 3753–3783, http://jmlr.org/papers/volume14/fukumizu13a/fukumizu13a.pdf.

[15] I. C. Gohberg and M. G. Kreĭn, *Introduction to the Theory of Linear Nonselfadjoint Operators*, Transl. Math. Monogr. 18, AMS, Providence, RI, 1969; translated by A. Feinstein.

[16] S. Grünewälder, G. Lever, L. Baldassarre, S. Patterson, A. Gretton, and M. Pontil, *Conditional mean embeddings as regressors*, in Proceedings of the 29th International Conference on Machine Learning, Omnipress, Madison, WI, 2012, pp. 1823–1830, https://icml.cc/2012/papers/898.pdf.

[17] O. Kallenberg, *Foundations of Modern Probability*, Springer, New York, 2006, https://doi.org/10.1007/978-1-4757-4015-8.

[18] H. Owhadi and C. Scovel, *Separability of reproducing kernel spaces*, Proc. Amer. Math. Soc., 145 (2017), pp. 2131–2138, https://doi.org/10.1090/proc/13354.

[19] H. Owhadi and C. Scovel, *Conditioning Gaussian measure on Hilbert space*, J. Math. Stat. Anal., 1 (2018), https://arxiv.org/abs/1506.04208.

[20] J. Park and K. Muandet, *A Measure-Theoretic Approach to Kernel Conditional Mean Embeddings*, preprint, https://arxiv.org/abs/2002.03689, 2020.

[21] S. Saitoh and Y. Sawano, *Theory of Reproducing Kernels and Applications*, Dev. Math. 44, Springer, Singapore, 2016, https://doi.org/10.1007/978-981-10-0530-5.

[22] V. V. Sazonov, *On characteristic functionals*, Teor. Veroyatnost. i Primenen., 3 (1958), pp. 201–205.

[23] I. Schuster, M. Mollenhauer, S. Klus, and K. Muandet, *Kernel conditional density operators*, in Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, Palermo, Sicily, Italy, 2020, Proceedings of Machine Learning Research, 2020, https://arxiv.org/abs/1905.11255.

[24] A. J. Smola, A. Gretton, L. Song, and B. Schölkopf, *A Hilbert space embedding for distributions*, in Proceedings of the 18th International Conference on Algorithmic Learning Theory, Springer, Berlin, Heidelberg, 2007, pp. 13–31, https://doi.org/10.1007/978-3-540-75225-7_5.

[25] L. Song, B. Boots, S. M. Siddiqi, G. Gordon, and A. Smola, *Hilbert space embeddings of hidden Markov models*, in Proceedings of the 27th International Conference on Machine Learning, ICML2010, ACM, New York, 2010, pp. 991–998, https://dl.acm.org/citation.cfm?id=3104322.3104448.

[26] L. Song, A. Gretton, and C. Guestrin, *Nonparametric tree graphical models*, in Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, Proc. Mach. Learn. Res. 9, Y. W. Teh and M. Titterington, eds., Proceedings of Machine Learning Research, 2010, pp. 765–772, http://proceedings.mlr.press/v9/song10a/song10a.pdf.

[27] L. Song, J. Huang, A. Smola, and K. Fukumizu, *Hilbert space embeddings of conditional distributions with applications to dynamical systems*, in Proceedings of the 26th Annual International Conference on Machine Learning, ACM, New York, 2009, pp. 961–968, https://doi.org/10.1145/1553374.1553497.

[28] B. K. Sriperumbudur, K. Fukumizu, and G. R. G. Lanckriet, *Universality, characteristic kernels and RKHS embedding of measures*, J. Mach. Learn. Res., 12 (2011), pp. 2389–2410, http://www.jmlr.org/papers/volume12/sriperumbudur11a/sriperumbudur11a.pdf.

[29] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. G. Lanckriet, *Hilbert space embeddings and metrics on probability measures*, J. Mach. Learn. Res., 11 (2010), pp. 1517–1561, http://www.jmlr.org/papers/volume11/sriperumbudur10a/sriperumbudur10a.pdf.

[30] I. Steinwart and A. Christmann, *Support Vector Machines*, Inf. Sci. Stat., Springer, New York, 2008, https://doi.org/10.1007/978-0-387-77242-4.

[31] I. Steinwart, D. Hush, and C. Scovel, *An explicit description of the reproducing kernel Hilbert spaces of Gaussian RBF kernels*, IEEE Trans. Inform. Theory, 52 (2006), pp. 4635–4643, https://doi.org/10.1109/tit.2006.881713.