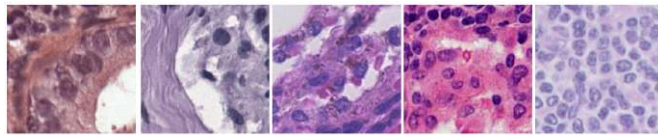


IMPROVEMENT, CLASSIFICATION
AND INTERPRETATION OF CANCER
HISTOLOGICAL IMAGES USING
PROBABILISTIC MODELS



Presented to obtain the degree of

Doctor of Philosophy

by

Fernando Pérez Bueno

supervised by

Rafael Molina Soriano and Valery Naranjo Olmedo



**UNIVERSIDAD
DE GRANADA**

**Programa de Doctorado en Tecnologías de la Información
y las Comunicaciones**

Editor: Universidad de Granada. Tesis Doctorales
Autor: Fernando Pérez Bueno
ISBN: 978-84-1117-645-3
URI: <https://hdl.handle.net/10481/79642>

Agradecimientos (Acknowledgments)

Antes incluso de empezar a escribir este documento, comencé a recibir enhorabuenas y felicitaciones ante la simple mención de leer la tesis y finalizar el doctorado. Quisiera compartir cada una de ellas con quienes han participado en mayor o menor medida de esta tesis, que ha sido un trabajo en equipo, y significado un desarrollo profesional y personal que hubieran sido imposibles sin todos y cada uno de ellos.

Comienzo agradeciendo su tiempo y paciencia a mi director Rafael Molina, que tiene la habilidad de hacerse respetar y transmitir su conocimiento sin perder la familiaridad y el buen humor.

Junto con él, es imprescindible agradecer a todo el grupo de investigación VIP (Visual Information Processing). Especialmente a Javier y Miguel Vega, que han trabajado tanto como Rafa y como yo en esta tesis. Y por supuesto también a Nicolás, Santiago, Pablo, Juanga, Arne y Miguel López (Miguelito) que me han ayudado siempre que ha hecho falta y me han allanado el camino. Incluyo también a Gabriel y Luz, trabajar con vosotros ha sido un placer, y a las últimas incorporaciones, Alba, Fran y Shuowen.

También quiero agradecer a mi co-directora Valery y al equipo de la UPV, ha sido muy interesante colaborar con vosotros aunque la pandemia nos haya privado de parte del contacto.

I would also like to say “Tusen Takk” to Kjersti Engan, Emiel Janssen, and the whole team in the University of Stavanger (Including Neel and the visiting Clarifys) for their warm welcome and support during my stay in Norway. It was a wonderful experience both in academics and climbing rocks.

Thanks as well to Prof. Katsaggelos. I’m so sorry that I could not visit Northwestern University as planned and I hope to do it in the near future.

Gracias también a mis compañeros del CITIC, especialmente al DB3 goloso por hacer del espacio de trabajo un lugar maravilloso y por las aventuras juntos. A mis amigos de la Sala de Armas Pedro del Monte, que aportan un complemento magnífico para el mens sana in corpore sano. A mis compañeros de Teleco, sin los que no sería ingeniero. Y también a Aurora Hermoso, por la afortunada casualidad.

Y por último, pero no por ello menos importante, a Lucía, que también ha trabajado en esta tesis, por estar a mi lado y ser estupenda. A mis padres y mi hermano que me han dado todo lo que he necesitado para llegar hasta aquí y a toda mi familia por su apoyo.

Abstract

Histopathological images are commonly used for the diagnosis of cancer and other diseases. These images are also used by Computer Assisted Diagnosis (CAD) systems, which have shown a promising performance on the diagnosis of cancer and other diseases.

However, the images obtained in different laboratories show acquisition differences that hamper the performance of AI-based CAD systems. In particular, color variation is often considered the most relevant issue when working with images from different centers. To accurately reduce color variation it is important to consider the acquisition procedure of histopathological image and, specifically, the staining protocol with two or more stains. Blind Color Deconvolution (BCD) use an observation model that takes the staining protocol into account and separate the stains mixed in the observed image, separating also color from structural information.

This thesis studies Bayesian modeling and inference, and their application to BCD techniques. With the proposed approach, it is possible to combine prior knowledge, the observation model, and the data evidence, obtaining robust, high quality posterior distributions that can be used to reduce the effect of color variation on CAD systems.

We propose three different Bayesian models for BCD of histopathological images: A Total Variation (TV) framework, Super Gaussian Sparse priors, and Bayesian K -Singular Value Decomposition (BKSVD), and apply them to stain separation, color normalization, stain augmentation, and cancer classification.

Two additional contributions are included in this thesis: the improvement of multi-spectral satellite images and network anomaly detection. Both benefit too from the use of Bayesian modeling and inference.

Furthermore, we also present three additional works in which we have collaborated but are not part of the compendium of publications presented to obtain the Ph.D. degree: a tutorial paper on the processing of histological images, a paper on blood detection using BCD, and a paper on Deep Variational BCD.

The project of this thesis was awarded the 3 Minute Thesis (3MT) prize at the University of Granada, and represented the university at the international Coimbra group competition in 2021.

Resumen

Las imágenes histopatológicas son una herramienta ampliamente utilizada para el diagnóstico del cáncer y otras enfermedades. También son utilizadas en sistemas de diagnóstico asistido por computador (CAD por sus siglas en inglés). Estos sistemas han obtenido resultados muy prometedores en el diagnóstico automático del cáncer y otras enfermedades.

Sin embargo, las imágenes obtenidas en distintos laboratorios presentan diferencias, debido al proceso de adquisición, que dificultan el uso de técnicas de inteligencia artificial. De estas diferencias, la variación de color se suele considerar el mayor problema cuando se trabaja con imágenes de distintos centros. Una solución precisa del problema de la variación de color requiere tener en cuenta el proceso de adquisición de las imágenes histopatológicas y, en concreto, el proceso de tinción con dos o más tinciones. Las técnicas de deconvolución ciega de color (BCD, por sus siglas en inglés Blind Color Deconvolution) utilizan un modelo de observación que tiene en cuenta este proceso y permite separar las tinciones que aparecen mezcladas en la imagen observada, separando además el color de la estructura de las tinciones.

En esta tesis se estudian la modelización e inferencia Bayesianas y su aplicación a BCD. Con la aproximación propuesta, es posible combinar conocimiento a priori, el modelo de observación y la evidencia que proporcionan los datos, y obtener así distribuciones a posteriori robustas y de gran calidad que pueden utilizarse para reducir la variación de color.

Se han propuesto tres modelos Bayesianos diferentes para la deconvolución de color de imágenes histológicas: El primero utiliza un marco de trabajo basado en la función de Variación Total (TV), el segundo utiliza distribuciones a priori de la familia super Gaussiana y por último un modelo basado en descomposición Bayesiana en K -valores singulares. Los modelos propuestos se han aplicado a la separación de tinciones, la normalización de color, el aumento de datos y la clasificación de varios tipos de cáncer.

Se incluyen en esta tesis dos contribuciones adicionales: la mejora de imágenes multiespectrales tomadas por satélite y la detección de anomalías en redes de ordenadores. Ambas se benefician también del uso de la modelización e inferencia Bayesianas.

Además, se presentan tres publicaciones en las que se ha colaborado pero que no se consideran parte del compendio presentado para obtener el título de doctor: una revisión sobre el procesamiento de imágenes histológicas, un trabajo de detección de sangre usando BCD y otro sobre BCD variacional profundo.

Esta tesis obtuvo el primer premio de la universidad de Granada en el concurso Tesis en 3 Minutos (3MT), y representó a la universidad en el concurso internacional del grupo Coimbra en 2021.

Índice general

Resumen extendido en castellano	ix
1 Introduction	1
1.1 Objectives	3
1.2 Methodology	5
1.3 Results	5
2 Reference-based Blind Color Deconvolution Using a Total Variation Prior	9
2.1 JCR Publication Details	9
2.2 Main Contributions	9
3 Reference-based Blind Color Deconvolution Using General Super Gaussian Priors	33
3.1 JCR Publication Details	33
3.2 Main Contributions	33
3.3 Related Conference Papers	35
4 Dictionary Learning for Blind Color Deconvolution	69
4.1 JCR Publication Details	69
4.2 Main Contributions	70
5 Pansharpening of Multispectral Images Using Probabilistic Models	101
5.1 JCR Publication Details	101
5.2 Main Contributions	101
6 Network Security Anomaly Detection Using Probabilistic Models	131
6.1 JCR Publication Details	131
6.2 Main Contributions	131

7 Other works (published, submitted, in preparation)	155
7.1 WSI Acquisition and Processing. A Review.	157
7.2 Deep Variational Stain Separation Using BCD	159
7.3 Robust BCD and Blood Detection	161
7.4 3 Minute Thesis (3MT) Competition	163
8 Concluding Remarks	165
Bibliography	169

Resumen extendido en castellano

Introducción

El análisis de imágenes histopatológicas realizado por patólogos y/o sistemas de diagnóstico asistido por computador (CAD por sus siglas en inglés) es una parte importante del proceso de diagnóstico del cáncer y otras enfermedades.

Los casos de cáncer, así como el cribado orientado a su detección, muestran una tendencia creciente en los últimos años, lo que aumenta el número de imágenes que tienen que ser procesadas y analizadas. Esto hace que el uso de sistemas CAD sea de vital importancia para reducir la carga de trabajo de los patólogos, aumentar la precisión de los diagnósticos y reducir el tiempo que requieren. Aunque estos sistemas han obtenido resultados muy prometedores en varias áreas, su desarrollo viene acompañado de nuevos retos. Por ejemplo, su rendimiento se reduce significativamente cuando se utilizan imágenes de hospitales que no aparecían en el conjunto de entrenamiento.

Las variaciones de color intra- and inter-hospitalarias causadas por diferencias en la tinción de las imágenes suelen considerarse una de las principales causas de este problema. El proceso de tinción se utiliza para resaltar las estructuras biológicas con colores diferentes, permitiendo que sean identificadas por los patólogos. Sin embargo, el aspecto final de las imágenes depende de muchos factores (temperatura, agentes químicos y software del escáner, entre otros) lo que imposibilita la estandarización de las imágenes durante su adquisición.

El procesado de las imágenes para eliminar la variación de color mejora el rendimiento de los sistemas CAD, y es de vital importancia para el desarrollo de sistemas que puedan usarse en distintos hospitales. Aparte de entrenar los CAD con una mayor cantidad de datos de diferentes hospitales, se han propuesto distintas aproximaciones para resolver el problema de la variación de color: normalización de color, aumento de color y separación de tinciones.

La normalización de color trata de transformar las imágenes observadas obteniendo imágenes que simulan haber sido teñidas bajo las mismas condiciones. El aumento de color crea nuevas imágenes con más variación de color, con el objetivo de reducir el error de generalización de los sistemas CAD en datos no vistos

anteriormente. Como se explica más adelante en esta tesis, estas aproximaciones requieren un paso previo que garantice la fidelidad a la estructura de la imagen original.

Esta tesis se centra en la separación de tinciones. Dado que las tinciones se fijan a elementos específicos del tejido, la variación de color dificulta la capacidad de los sistemas CAD para identificar correctamente la estructura y condición de los mismos. Mediante las técnicas de separación de tinciones es posible identificar y separar cada una de las tinciones en la imagen. Esta forma de expresar la información, tiene un mayor sentido biológico que la mezcla de tinciones en la imagen RGB observada. Además, se ha probado que la separación de tinciones es útil para el diagnóstico automático, como un paso previo a la normalización de color, y para realizar aumento de color.

Para afrontar el problema de la separación de tinciones se usan frecuentemente técnicas de deconvolución ciega de color (BCD, por sus siglas en inglés *Blind Color Deconvolution*). Estas técnicas, tras una transformación apropiada de la imagen RGB, estiman los vectores de color y la estructura (concentraciones) correspondientes a cada tinción. Para ello, se hace uso de la ley de Beer-Lambert que, en el espacio de la densidad óptica (OD), establece una relación lineal entre intensidad observada y la concentración de cada tinción.

Para una imagen observada $I \in \mathbb{R}^{Q \times 3}$ donde cada columna corresponde a un canal RGB, y Q es el número de píxeles, la OD para cada canal, c , se obtiene como $\mathbf{y}_c = -\log_{10}(\mathbf{i}_c/\mathbf{i}_c^0)$, donde \mathbf{i}_c^0 representa la luz incidente (típicamente 255 para imágenes RGB) y las operaciones de división y log se ejecutan para cada píxel. Así, la OD de una imagen histológica $Y \in \mathbb{R}^{Q \times 3}$ teñida con N_s tinciones, puede separarse en una matriz de color $\mathbf{M} \in \mathbb{R}^{3 \times N_s}$ y una matriz de concentraciones $\mathbf{C} \in \mathbb{R}^{Q \times N_s}$ como

$$\mathbf{Y}^T = \mathbf{M}\mathbf{C}^T + \mathbf{N}^T, \quad (1)$$

donde $\mathbf{N} \in \mathbb{R}^{Q \times 3}$ es una matriz de ruido.

La matriz de color \mathbf{M} tendrá tantas columnas como tinciones en la imagen, donde cada columna representa las propiedades de color de una de las tinciones. La matriz de concentraciones \mathbf{C} contiene también una columna para cada tinción, y tantas filas como píxeles en la imagen. Cada columna $\mathbf{c}_q = [c_{1,q}, \dots, c_{N_s,q}]^T$ representa la contribución de cada tinción al píxel q -ésimo en \mathbf{Y} y cada columna de \mathbf{C} representa las concentraciones de cada tinción.

Nótese que para estimar la separación de tinciones latente, es necesario gestionar la incertidumbre que surge del ruido presente en los datos y del número finito de observaciones. Los modelos probabilísticos proporcionan herramientas para cuantificar y gestionar dicha incertidumbre. Mediante el modelado y la inferencia Bayesiana, se pueden combinar el conocimiento a priori sobre la matriz de

color y las concentraciones, el modelo de observación, y la evidencia de los datos observados al microscopio.

Objetivos y estructura de la tesis

El objetivo principal de esta tesis es el uso de modelización e inferencia Bayesianas para la mejora de imágenes histopatológicas y su uso en sistemas CAD. En particular, se centra en desarrollar y aplicar modelos probabilísticos para abordar la variación de color mediante BCD. Además, se incluye un segundo objetivo que aplica la modelización Bayesiana a otras áreas de interés. Para alcanzar estos objetivos, definimos un conjunto de objetivos específicos que se detallan a continuación.

1. Mejorar las imágenes histopatológicas para su uso en sistemas de diagnóstico asistido por computador.
 - (a) Desarrollar modelos probabilísticos para deconvolución ciega de color de imágenes histológicas.
 - (b) Aplicar BCD para mejorar el rendimiento de los sistemas CAD.
 - (c) Proponer nuevas aproximaciones para resolver la variación de color.
2. Utilizar los modelos probabilísticos en otras áreas de interés.
 - (a) Imágenes multiespectrales tomadas por satélite.
 - (b) Detección de anomalías en redes de ordenadores.

Esta tesis está estructurada en tres bloques. En los dos primeros se abordan los objetivos presentados: Modelización e inferencia Bayesianas para BCD, otras áreas de aplicación de la modelización probabilística. El tercero incluye otros trabajos, bien aceptados en revistas con índice de impacto, enviados o en preparación que no forman parte de las cinco publicaciones que conforman el compendio de la tesis. Los tres bloques se describen a continuación.

1. **Modelización e inferencia Bayesianas para BCD.** Incluye los capítulos 2, 3 y 4. Los dos primeros capítulos presentan modelos BCD Bayesianos basados en referencias. Aunque los auténticos vectores de color de la imagen observada son desconocidos, no lo es el protocolo de tinción (ej. H&E). Por tanto, podemos asumir que las tinciones serán similares a una referencia dada. Los modelos BCD basados en referencias utilizan una a priori de similitud sobre los vectores de color, junto con diferentes distribuciones a priori sobre las concentraciones. En concreto se presentan los modelos basados en

la a priori de Variación Total (TV) y la familia de distribuciones Super-Gaussiana (SG). Estos modelos se han aplicado a separación de tinciones, normalización de color y clasificación de distintos tipos de cáncer.

El capítulo 4 aborda el problema de BCD como un problema de aprendizaje de diccionarios. Los modelos basados en referencias funcionan muy bien cuando las imágenes observadas están cercanas a la referencia. Sin embargo, carecen de la flexibilidad necesaria para adaptarse a cambios grandes de color, como los que podrían aparecer entre imágenes de distintos laboratorios. Afrontar el problema de encontrar la matriz de color como un problema de aprendizaje de diccionarios resuelve este problema, permitiéndonos encontrar la matriz de color que mejor representa las tinciones en la imagen. Este modelo se ha aplicado a separación de tinciones, normalización de color, aumento de datos y clasificación de imágenes de cáncer.

Los objetivos 1.a, 1.b y 1.c se trabajan en este bloque.

2. **Otras áreas de aplicación de la modelización probabilística.** Incluye los capítulos 5 y 6, donde se aplican los modelos probabilísticos a otras áreas de interés. En particular, las distribuciones SG se aplican al pansharpening de imágenes de satélite (Capítulo 5, objetivo 2.a) y el análisis de componentes principales probabilístico (PPCA por sus siglas en inglés) se utiliza para detección de anomalías en redes de ordenadores (Capítulo 6, objetivo 2.b)
3. **Otros trabajos.** Este tercer bloque, presentado en el capítulo 7, incluye otros trabajos de interés en los que el doctorando ha tenido un papel relevante. Se incluyen, un trabajo en revista JCR, (i) una revisión sobre técnicas de procesado de WSIs, un trabajo enviado a una revista, (ii) el desarrollo de modelos Bayesianos profundos para BCD), y un trabajo en preparación, (iii) una aplicación de las técnicas BCD para la detección de sangre en WSIs. También se incluye la participación en el concurso *Tesis en 3 Minutos* (3MT) del grupo Coimbra, en el que el doctorando obtuvo el primer premio en la final institucional de la universidad de Granada en 2021. Estos trabajos están relacionados con los objetivos 1.a, 1.b y 1.c.

Conclusiones

La principal conclusión de esta tesis doctoral es que la modelización e inferencia Bayesianas pueden utilizarse para mejorar las imágenes histológicas de cáncer, haciéndolas más fáciles de clasificar e interpretar con sistemas CAD. Hemos explorado la aplicación de las técnicas Bayesianas a la deconvolución ciega de color para separar las imágenes observadas en los elementos latentes que las componen

(es decir, el color de las tinciones y su concentración en cada píxel). La modelización e inferencia Bayesianas también pueden aplicarse en otros ámbitos como el pansharping y la detección de anomalías. A continuación se incluye un desglose en conclusiones específicas:

- El uso de conocimiento a priori sobre las imágenes ayuda a obtener una mejor separación del color y las concentraciones. Los modelos probabilísticos y la inferencia Bayesiana proporcionan las herramientas para utilizar el conocimiento a priori y gestionar la incertidumbre del problema.
- La mejora de las imágenes puede tener significados diferentes en función de la tarea a realizar. Para el análisis visual se desea una mayor fidelidad al tejido original, mientras que para su uso en sistemas CAD se busca obtener mejores características de clasificación a partir de imágenes en las que se eliminan el ruido y los elementos residuales.
- La rareza¹ (en el sentido de que un número elevado de elementos tienen un valor cercano a cero y solo existen unos pocos separados y distintos de cero) es una característica deseada para la separación latente de las tinciones en las imágenes. Hemos explorado la rareza en las concentraciones de tinción con tres enfoques diferentes: utilizando una a priori TV directamente sobre las concentraciones, utilizando la familia de distribuciones a priori SG en las concentraciones filtradas mediante paso alto para remarcar los bordes de la imagen, y con una a priori jerárquica (equivalente a una distribución Laplaciana) en las concentraciones de cada píxel que promueve la asignación de los mismos a una sola tinción.
- Las técnicas BCD basadas en referencias de color presentan un enfoque robusto que no se ve afectado por los artefactos de la imagen, pero carece de la flexibilidad necesaria para adaptarse a distribuciones de color alejadas de la referencia. Por otro lado, el aprendizaje de diccionarios para BCD es capaz de estimar una matriz de color que representa mejor la tinción diferencial de la imagen, pero que puede estar influida por elementos inesperados en las imágenes.
- Las técnicas Bayesianas para BCD son computacionalmente costosas pero superan a los enfoques no probabilísticos para la separación de tinciones. Sin embargo, debido al reducido número de tinciones en las imágenes histológicas, su aplicación a las imágenes de gran tamaño puede acelerarse mediante el muestreo de píxeles para la estimación de los parámetros del modelo BKSVD.

¹La traducción del término en inglés *sparsity* no está muy extendida y no queda claro en la literatura cual es el término más adecuado. Los más habituales son rareza o dispersión.

- Las técnicas BCD tienen un gran potencial para la mejora e interpretación de las imágenes histológicas. La separación de tinciones puede utilizarse para la normalización y el aumento del color, y directamente para su uso en sistemas CAD. Utilizar los sistemas CAD con las imágenes de concentración de una sola tinción en lugar de la imagen observada o normalizada RGB puede mejorar la precisión del diagnóstico. Este enfoque imita el análisis realizado por los patólogos, ya que diferencian las tinciones en la imagen y no el color que presentan.
- Algunas de las lecciones aprendidas al trabajar con imágenes histológicas pueden extenderse a otras áreas. El uso de modelos probabilísticos para separar la información en sus componentes latentes puede ayudar a resaltar la información previamente confundida o disfrazada en las variables observadas. En concreto:
 - La estimación de imágenes multiespectrales a partir de imágenes multiespectrales de baja resolución e imágenes pancromáticas de alta resolución, puede mejorarse haciendo uso de la familia de distribuciones SG, separando la contribución de la imagen pancromática a cada canal de la imagen multiespectral de alta resolución.
 - El PCA probabilístico (PPCA) proporciona un espacio latente que puede utilizarse para la detección robusta de anomalías. Además, el PPCA establece un puente entre los modelos clásicos de detección de anomalías en redes y los enfoques generativos recientes, como los VAE, que pueden utilizarse para comprender mejor estos últimos.
- Los trabajos incluidos en el capítulo 7 muestran que es posible e interesante continuar la investigación en el procesamiento de imágenes histológicas y la modelización probabilística. El trabajo de revista en la sección 7.1, revisa el estado del arte y señala líneas y retos por afrontar. El trabajo recientemente enviado de la sección 7.2, es según nuestro conocimiento, la primera aproximación Bayesiana a BCD usando redes neuronales profundas, y abre el camino a nuevos trabajos de aprendizaje profundo. El trabajo en preparación de la sección 7.3, también abre una nueva línea de investigación, presentando el uso de BCD para la detección de artefactos en imágenes histológicas. Finalmente, el premio 3MT presentado en la sección 7.4, muestra que la investigación realizada en esta tesis es de interés para el público general.

CHAPTER 1

Introduction

The analysis of histopathological images by pathologists and/or Computer-Aided Diagnosis (CAD) systems plays a critical role in cancer diagnosis and treatment decision. These images are tissue sections, stained using a combination of stains to reveal their underlying structures and conditions, and then observed with a digital microscope and stored as Whole Slide Images (WSIs) [1]. To provide an insight for the unfamiliar reader, figure 1.1 depicts a glass slide and the histopathological image observed on the screen.



Figure 1.1: **Left:** Glass slide ready to be observed or scanned on the microscope. **Right:** Histopathological image observed on a screen.

The growing trend of cancer cases, as well as the associated screening for its detection, in recent years [2], increases the number of slides that need to be processed and analysed. The use of CAD systems is of paramount importance to reduce pathologists' workload, to increase diagnostic accuracy and to reduce turnaround times [3]. Their development, however, comes with its own challenges. Although data-driven CAD systems work well in several areas of diagnosis, their performance degrades significantly when tested on images from hospitals not included in the training set [4].

Intra- and inter-hospital color variation [5, 4] caused by differences in the staining of the images is often considered one of the major causes of the loss of performance. Staining is used to highlight biological structures with differential

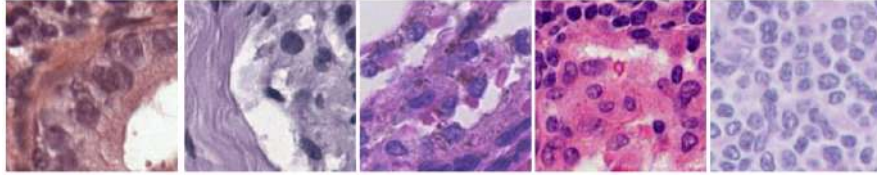


Figure 1.2: Color variation depicted with histological image patches from five different laboratories.

colors that ultimately help pathologists to identify them. The images that we include in figure 1.2 as an example, have been stained in five different laboratories using the most common staining protocol (i.e. Hematoxylin and Eosin (H&E)), where hematoxylin stains the cell nuclei in blue and eosin stains the cytoplasm and connective tissue in pink. The final appearance of the images, however, is affected by variations in temperature, chemicals, and scanning software, among others, making almost impossible to standardize the appearance of the images during its acquisition.

Preprocessing the images to eliminate color variation improves the performance of CAD systems [6, 4] and is crucial to obtain transferable systems that can be used in different hospitals [7]. Beyond training CAD systems with more and multi-hospital data, several approaches have been proposed to tackle color variation: color normalization (CN), color augmentation (CA), and *stain separation*.

CN aims at transforming the observed images into new ones as if all of them had been tainted using the same protocol and stains. CA aims at hallucinating new images with augmented color variation, with the objective of reducing the generalization error of CAD systems on unseen data. See [5]. As we will later explain in this dissertation, these approaches require a preliminary step to guarantee fidelity to the structure of the original image.

This thesis focuses on *stain separation*. Given that stains bind to specific elements on the tissue, color variation jeopardizes the ability of the CAD system to correctly identify the structures of the tissues and their conditions. By using *stain separation* techniques it is possible to identify and separate each stain in the image [5]. This is more biologically meaningful than the mixture of stains in the observed RGB image [8]. In addition, *stain separation* has proven to be useful for automated diagnosis [9, 10, 7], as a preliminary step for CN [11, 12], and to CA [7, 4].

To tackle the *stain separation* problem, Blind Color Deconvolution (BCD) techniques are frequently used. After an appropriated RGB image transformation, they estimate both the stain color-vectors and the corresponding stain structure (concentrations). To do so, they make use of the Beer-Lambert law that, in the

Optical Density (OD) space, establishes a linear relationship between the intensity observed and the concentration of each stain [13].

For an observed histological image $I \in \mathbb{R}^{Q \times 3}$ where each column corresponds to a RGB channel and Q is the number of pixels, the OD for each channel, c , is obtained as $\mathbf{y}_c = -\log_{10}(\mathbf{i}_c/\mathbf{i}_c^0)$, where \mathbf{i}_c^0 denotes the incident light (typically 255 for RGB images) and the log and division operations are performed pixel-wise. Then the OD for a histopathological image $Y \in \mathbb{R}^{Q \times 3}$ stained with N_s stains can be separated into a color-vector matrix $\mathbf{M} \in \mathbb{R}^{3 \times N_s}$ and a concentration matrix $\mathbf{C} \in \mathbb{R}^{Q \times N_s}$ as

$$\mathbf{Y}^T = \mathbf{M}\mathbf{C}^T + \mathbf{N}^T, \quad (1.1)$$

where $\mathbf{N} \in \mathbb{R}^{Q \times 3}$ is a noise matrix.

The color-vector matrix \mathbf{M} will contain as many columns as stains in the image, where each column represents the color properties of one of the stains. The concentration matrix \mathbf{C} contains also one column for each stain and as many rows as pixels in the image. Each column $\mathbf{c}_q = [c_{1,q}, \dots, c_{ns,q}]^T$ represents the contribution of each stain to the q -th pixel in \mathbf{Y} and each row of \mathbf{C} represents each stain concentrations.

Notice that estimating the latent *stain separation* from the observed multi-stained image requires to deal with the uncertainty that arises both through noise on the observations, as well as through the finite size of data. Probabilistic models provide a consistent framework for the quantification and manipulation of such uncertainty [14]. Using Bayesian modeling and inference it is possible to combine prior knowledge, the observation model, and the evidence from the data observed on the microscope. In particular, they can be applied to perform Blind Color Deconvolution (BCD), by setting prior distributions on the unknown colors and concentrations of the stains [15]. An example is depicted in figure 1.3.

1.1 Objectives

The main objective in this thesis is to use Bayesian modeling and inference to improve histopathological images for their use in CAD systems. Specifically, this Ph.D. thesis focuses on the application of probabilistic modeling to BCD techniques. In addition, we include a second objective where we apply Bayesian modeling and inference to other areas of interest. To achieve those objectives, we define a set of specific objectives that are summarized as follows.

1. **To improve histopathological images for their use on Computer Aided Diagnosis systems.**

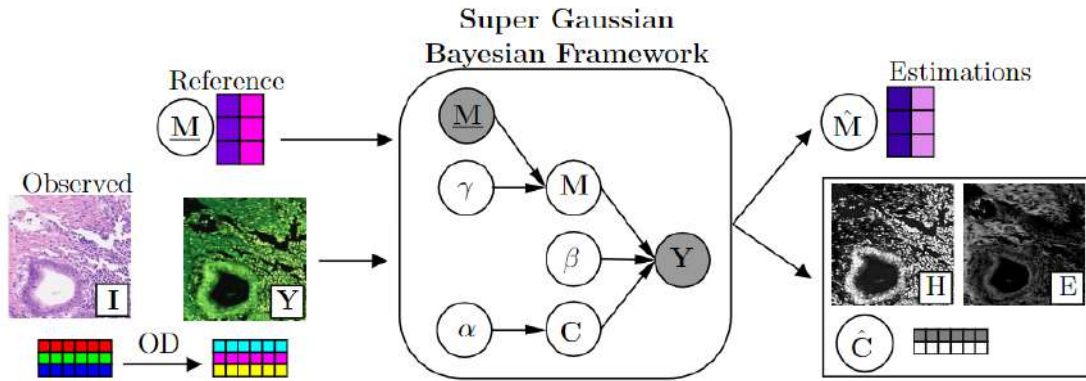


Figure 1.3: Pipeline of the BCD framework in [12], depicted here to illustrate the BCD procedure. First, the H&E image is converted to the OD space and given to the Bayesian framework. Then, applying Bayesian modeling and inference, posterior estimations for the color-vector matrix M and C are obtained.

- (a) **To develop state-of-the-art methods for blind color deconvolution of histopathological images using a probabilistic approach.** So far, the application of probabilistic modeling for BCD has not been explored in deep. Bayesian techniques have proven to achieve a high performance on other areas of image processing. Therefore, we truly believe that Bayesian BCD will perform better than state-of-the-art methods.
 - (b) **Apply the BCD results to improve CAD performance.** The ultimate goal of histological image processing is to facilitate and improve their analysis. We will evaluate the effect of BCD techniques on cancer classification. We plan to evaluate the use of stain-separated classification and color normalization using classifiers and to compare the performance achieved by different BCD techniques.
 - (c) **To propose new approaches to solve color variation.** Data augmentation often used to solve data shortage on complex classification models. Using BCD is possible to produce stain augmentation, a histology-specific type of data augmentation. We seek to propose a new approach for stain augmentation using Bayesian BCD.
2. **Use Bayesian modeling and inference in other areas of interest.** Notice that ultimately, we are using probabilistic models to estimate a separation of the observed information in their latent components. We plan to use the Bayesian framework in other areas that might be benefited from such separation. In particular we explore:

- (a) Pansharpening of multispectral satellite images.
- (b) Network anomaly detection.

1.2 Methodology

The development of this thesis requires a methodology that consider both theory and practice. We follow the guidelines of the scientific method and include the following steps:

1. **Observation:** We first study the literature regarding BCD and other pre-processing techniques for histopathological images.
2. **Data collection:** The techniques to develop in this thesis require to use real-world data to assess the algorithms. We consider public databases whenever possible.
3. **Hypothesis formulation:** We address the problems presented in the objectives by proposing new methods that can improve the state-of-the-art methods.
4. **Experimentation:** We design and perform a rigorous experimentation of the methods proposed in step 3 on the collected data from step 2. We use the computation resources of the Visual Information Processing research group of the University of Granada. The metrics to use must be chosen according to the task, ensuring an appropriate evaluation of the results.
5. **Hypothesis contrast:** We compare, analyze and validate the results obtained in the experimentation with those of the state-of-the-art techniques in the literature.
6. **Hypothesis proof or refusal:** The hypothesis is accepted, rejected or modified in consequence with the gathered results. If necessary, the previous steps are repeated.
7. **Thesis extraction:** We formalize the conclusions during the research process and justify the developed methods through the experimentation. All the proposals and results are synthesized in this dissertation.

1.3 Results

According to the objectives and methodology of this thesis, here we present the results that have been obtained. We have developed and applied different Bayesian

models to histopathological images, satellite images, and network anomaly detection. Three main blocks can be distinguished in this thesis. The first two, we tackle the objectives of this thesis: Bayesian modeling and inference for BCD, other applications of probabilistic modeling. The third one include other works that are published in indexed journals, submitted, or in preparation that are not part of the compendium of five publications presented to obtain the Ph.D. degree. Their contents are described below:

- **Bayesian Modeling and Inference for Blind Color Deconvolution.**

This includes Chapters 2, 3, and 4. The first two chapters extend the Bayesian BCD approach presented in [15]. The actual color-vector matrix of the stains in the images is assumed to be unknown. The staining protocol, however, is known (e.g. H&E). Therefore, we can assume that the stain will show a certain similarity with a given reference. Reference-based BCD models rely on a similarity prior on the color-vector matrix, together with different priors on the concentration of the stains such as the Total Variation (TV) prior or General Super Gaussians (SG) priors. These models were applied to *stain separation*, color normalization, and cancer classification.

Chapter 4 approaches the BCD problem as a dictionary learning problem. The quality of reference-based models results improve when the images are close to the reference. However, they lack the ability to adapt to larger color changes as those that might arise between different laboratories, requiring to set a different reference for each of them. Tackling the problem of finding the color-vector matrix as a dictionary learning problem solves this issue, allowing us to find the color-vector matrix that better represents the different stains in the image. This model was applied to *stain separation*, color normalization, stain augmentation, and cancer classification.

Objectives 1.a, 1.b and 1.c are addressed in this block.

- **Other applications of probabilistic modeling.** This includes Chapters 5 and 6. This block explores the application of probabilistic models to other areas of interest addressing objective 2. In particular, the General SG priors are applied for the pansharpening of satellite images (Chapter 5, objective 2.a) and Probabilistic Principal Component Analysis (PPCA) is used for network anomaly detection (Chapter 6, objective 2.b).

- **Other works.** This third block, presented in Chapter 7, includes other works of interest in which the Ph.D. candidate had a relevant role in their elaboration. It includes: A JCR journal paper, (i) a survey on WSI acquisition and preprocessing techniques, a submitted paper, (ii) the development of a deep variational Bayesian models for BCD, and a work in preparation,

(iii) an application of BCD techniques for blood detection on WSI images. This block also includes the participation of the Ph.D. candidate on the 3 Minute Thesis (3MT) competition hold by the Coimbra group, in which this thesis won the first prize in the institutional finals of the University of Granada in 2021.

Since these works are not part of the compendium of publications presented to obtain the Ph.D. degree, they will only be mentioned with their relevant contributions. The works included in this chapter are related to objectives 1.a, 1.b and 1.c.

Next, we provide a general overview of each chapter. The Journal Citation Report (JCR) publication details and main contributions will be highlighted at the beginning of the corresponding chapter. Finally, the main joint conclusions will be drawn in Chapter 8. Notice that the focus is on the application of probabilistic modeling to the field of histopathological images with three journal papers, and how the two additional journal papers included, on pansharpening of satellite images and network anomaly detection, share a common element: the underlying Bayesian modeling and inference to separate the observed information into latent features of interest.

Chapter 2: In this work, we use a similarity prior on the color-vector matrix, a Beer-Lambert based observation model, and propose the use of a TV prior on the stain concentration for BCD of histological images. The TV prior is used to reduce noise on the image while preserving sharp edges. We also explore the dichotomy between two conflicting objectives often pursued in histopathological image analysis: closeness to the original tissue and high classification performance. The proposed approach was evaluated on real images of different tissues and prostate cancer classification using shallow and deep classifiers.

Chapter 3: In this work, we apply Bayesian modeling and inference based on the use of general SG sparse priors on the stain concentrations and the previously proposed similarity prior on the color-vector matrix for BCD of histopathological images. While the inference procedure is more complex than in the previous chapter, SG priors include a large class of sparse image priors which represent well-sharp image characteristics. The experimental validation was extended with additional databases including images from different laboratories, applying the BCD results to obtain CN, comparing the classification performance of BCD-separated stain concentration versus CN RGB images, and analyzing the dependency of the method on the similarity prior on the color vectors.

Chapter 4: In this work, we introduce Bayesian dictionary learning for BCD of histopathological images using Bayesian K -Singular Value Decomposition (BKSVD) to estimate the color-vector matrix. The idea here is that the stains are added to give differential color to the structures on the image. A spar-

sity constraint equivalent to a zero-mean Laplace distribution is set on the stains concentration for each pixel, promoting the unsupervised estimation of a color-vector matrix that sparsely represents the staining on the image. The method was tested on stain separation, CN, CA, and classification performance using large histological datasets with intra- and inter-laboratory variations.

Chapter 5: In this work, we explore Bayesian modeling and inference using the SG prior model for the pansharpening of multispectral (MS) images. The pansharpening technique fuses a low spatial resolution MS image and a high spatial resolution panchromatic one to obtain a high-resolution MS image. In this case, the panchromatic image is modeled as a convex combination of the high resolution MS channels and the SG distributions are used as priors for the MS-high resolution image. Therefore, it is possible to separate the contribution of the panchromatic image to each channel of the MS high resolution image. The method was tested on real and synthetic images from three different satellites.

Chapter 6: In this work, anomaly detection techniques based on PCA are revisited from a probabilistic point of view. The Probabilistic PCA (PPCA) provides a separation of the data into its latent component and a generative modeling that is at the basis of the definition of Variational AutoEncoders (VAE). Relating PCA-based anomaly detection models to generative approaches, our objective is to allow well-known lessons from PCAs to be applied to generative models. The mathematical model was evaluated using a synthetic dataset created to better understand the analysis, and a real-traffic dataset for network anomaly detection.

Chapter 7: This chapter comprises additional works in which the Ph.D. candidate had a relevant role in their elaboration. It includes: A JCR journal paper, (i) a survey on WSI acquisition and preprocessing techniques, a submitted paper, (ii) the development of a deep variational Bayesian models for BCD, and a work in preparation, (iii) an application of BCD techniques for blood detection on WSI images. This block also includes the participation of the Ph.D. candidate on the 3 Minute Thesis (3MT) competition hold by the Coimbra group, in which this thesis won the first prize in the institutional finals of the Univesity of Granada in 2021.

CHAPTER 2

Reference-based Blind Color Deconvolution Using a Total Variation Prior

2.1 JCR Publication Details

Authors: Fernando Pérez-Bueno, Miguel López-Pérez, Miguel Vega, Javier Mateos, Valery Naranjo, Rafael Molina, Aggelos K. Katsaggelos

Title: A TV-based Image Processing Framework for Blind Color Deconvolution and Classification of Histological Images

Reference: Digital Signal Processing, Volume 101, 2020, 102727,

Status: Published

DOI: <https://doi.org/10.1016/j.dsp.2020.102727>

Quality indices:

- Impact Factor (JCR 2020): 3.381
 - Rank: 125/276 (Q2) in Engineering, Electrical and Electronic
- Journal Citation Indicator (JCR 2020): 0.9
 - Rank: 98/319 (Q2) in Engineering, Electrical and Electronic

2.2 Main Contributions

- We use Bayesian probabilistic models for the deconvolution of histological images. We propose the use of the TV prior on the stain concentrations which removes noise and preserve sharp edges, in combination with a reference color-vector matrix prior and a observation model that follows the Beer-Lambert law. All model parameters and latent variables are automatically estimated.

- We discuss that visual inspection and automatic classification may be conflicting goals, as the better reconstruction of the images does not always lead to the extraction of better features for an improved classification.
- The proposed approach was successfully evaluated on two real histopathological image datasets: one for stain separation and one for prostate cancer classification. It achieved the most accurate stain separation and improved the performance of several classifiers tested.

A TV-based Image Processing Framework for Blind Color Deconvolution and Classification of Histological Images

Fernando Pérez-Bueno^{a,1,*}, Miguel López-Pérez^a, Miguel Vega^b, Javier Mateos^a, Valery Naranjo^c, Rafael Molina^a, Aggelos K. Katsaggelos^d

^a*Dpto. Ciencias de la Computación e Inteligencia Artificial, Universidad de Granada, Spain*

^b*Dpto. de Lenguajes y Sistemas Informáticos, Universidad de Granada, Spain*

^c*Dpto. de Comunicaciones, Universidad Politécnica de Valencia, Spain*

^d*Dept. of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL, USA*

Abstract

In digital histopathological image analysis, two conflicting objectives are often pursued: closeness to the original tissue and high classification performance. The former objective tries to recover images (stains) that are as close as possible to the ones obtained by staining the tissue with a single dye. The latter objective requires images that allow the extraction of better features for an improved classification, even if their appearance is not close to single stained tissues. In this paper we propose a framework that achieves both objectives depending on the number of stains used to mathematically decompose the scanned image. The proposed framework uses a total variation prior for each stain together with the similarity to a given reference color-vector matrix. Variational inference and an evidence lower bound are utilized to automatically estimate all the latent variables and model parameters. The proposed methodology is tested on real images and compared to classical and state-of-the-art methods for histopathological blind image color deconvolution and prostate cancer classification.

Keywords: Blind Color Deconvolution, histopathological images, Variational Bayes, Prostate Cancer

*Corresponding author

Email addresses: fpb@ugr.es (Fernando Pérez-Bueno), mlopez@decsai.ugr.es (Miguel López-Pérez), mvega@ugr.es (Miguel Vega), jmd@decsai.ugr.es (Javier Mateos), [vnaranajo@dcom.upv.es](mailto:vnanarajo@dcom.upv.es) (Valery Naranjo), rms@decsai.ugr.es (Rafael Molina), aggk@eecs.northwestern.edu (Aggelos K. Katsaggelos)

¹This work was sponsored in part by Ministerio de Ciencia e Innovación under Contract BES-2017-081584 and project DPI2016-77869-C2-2-R.

1. Introduction

Histopathological tissues are usually stained with a combination of stains that bind to specific proteins on the tissue. Hematoxylin and Eosin (H&E) is one of the most commonly used combination of stains. Hematoxylin stains cell nuclei while eosin stains cytoplasm and extracellular matrix components [1]. In digital brightfield microscopy, stained slides are then scanned to obtain high resolution whole-slide images (WSI). WSI analysis requires a lot of time and effort and computer-aided diagnosis (CAD) systems have become a valuable ally for pathologists. These systems frequently make use of the information provided by the different stains separately [2]. The separation of the stains in a WSI is known as Color Deconvolution (CD) and aims at estimating each stain concentration at each pixel location. Usually, the color spectral properties of each stain are also unknown since they vary from image to image. Color variations have a wide range of origins: different scanners, stain manufactures, or staining procedures, among others that create inter- and intra-laboratory differences. A study on color variation sources can be found in [3]. Blind CD techniques estimate image specific stain color-vectors together with stain concentrations.

CD is usually considered as a branch of color normalization. Tosta *et al.* [3] classified normalization methods into histogram matching, color transfer, and spectral matching. Normalization does not always require CD. Histogram matching methods do not use it, which leads to information loss as stains are assumed to be equally distributed. Color transfer usually separates histological regions identified by a segmentation step or between dyes. Although they usually involve deconvolution steps, it is not their main objective but a way to apply an statistical based color correction. Spectral matching techniques require to identify image specific spectral properties through CD. One of the first CD methods was proposed by Ruifrok *et al.* [4]. They obtained a set of globally standard color-vectors for hematoxylin, eosin and 3,3'-Diaminobenzidine (DAB), by measuring the relative absorption of each stain in single-stained images. The proposed set of stain color-vectors was calibrated for processing and digitization at the authors' laboratory. While these color vectors have been widely used, they do not take into account inter-slide variability. Empirical determination of the color-vector using single-stained tissue was used in [5, 6]. Aside from techniques that require the user to select pixels corresponding to each stain [7], several methods have been proposed to tackle inter-slide variability. In [2] Non-negative Matrix Factorization (NMF) is used to solve the problem formulated as a blind source separation one. This line of research was further developed in [8] and [9] by adding regularization and sparsity terms to take into account that a type of stain is only bounds to certain biological structures. Singular Value Decomposition (SVD), proposed in [10] to separate H&E images, was extended by McCann *et al.* [11] by taking into account the interaction between eosin and hematoxylin. The use of Non-Negative Least Squares (NNLS) to improve the performance of NMF is proposed in [12] to obtain a faster and less memory demanding method. Clustering techniques were explored in [13] where the stain vectors are estimated by projecting the input color image onto the Maxwellian chromaticity plane to form clusters, each one corresponding to one stained tissue type. In [14], to estimate the stain color-vector matrix, the image is segmented into background and pixels belonging to each stain using supervised relevant vector machines.

The mean color of the pixels in each class is utilized as the stain color vector. Alsubaie *et al.* [15, 16], following [17], applied Independent Component Analysis (ICA) in the wavelet domain where the independence condition among sources is relaxed. Astola *et al.* [18] states that the method in [10] obtains better results applied in the linearly inverted RGB-space and not in the (logarithmically inverted) absorbency space. In [19] a loss function based on the authors’ experience is optimized to obtain the image stain color-vectors. For further information on classical and state-of-the-art methods, the interested reader might check the reviews published in [20, 3].

In this paper, we present a framework for blind color deconvolution and classification of histological images. Depending on the number of stains used to mathematically model the observed image, the framework can be utilized to either recover the original H&E stains or to produce an H&E separation that boosts the performance of image classifiers. Within the framework, the proposed Bayesian blind CD problem algorithm, extends our previous work in [21] and [22]. In [21], a prior on the color-vectors, favouring similarity to a reference stain color-vectors, as well as a smoothness Simultaneous Autoregressive (SAR) prior model on each stain concentrations was used. As the SAR prior tends to oversmooth the edges of the image structures, in [22], we proposed the use of a Total Variation (TV) prior on each stain. The TV prior reduces the noise in the images while preserving sharp edges [23]. All model parameters were experimentally determined. In this paper, we extend the work in [22] by applying the Variational Bayes inference [24] and an evidence lower bound to automatically estimate all the latent variables and model parameters for blind color deconvolution and classification purposes. The proposed framework has been tested on additional real images for blind color deconvolution, where the fidelity to a ground-truth stain separation is assessed, and, for the first time, on classification tasks.

The rest of the paper is organized as follows: in section 2 the problem of color deconvolution is mathematically formulated. Following the Bayesian modelling and inference, in section 3 we propose a fully Bayesian algorithm for the estimation of the concentrations and the color-vector matrix as well as all the model parameters. In section 4, the proposed framework is evaluated on H&E stained images and its performance is compared with other classical and state-of-the-art CD methods in two different scenarios: color deconvolution and prostate cancer classification. Finally, section 5 concludes the paper.

2. Problem Formulation

Digital brightfield microscopes usually store a stained histological specimen’s WSI as an RGB color image of size $M \times N$, represented by the $MN \times 3$ matrix, $\mathbf{I} = [\mathbf{i}_R \ \mathbf{i}_G \ \mathbf{i}_B]$. Each color plane is stacked into a $MN \times 1$ column vector $\mathbf{i}_c = (i_{1c}, \dots, i_{MNc})^T$, $c \in \{R, G, B\}$. Each value i_{ic} represents the transmitted light on color band $c \in \{R, G, B\}$ for the pixel i of the slide.

CAD systems, on the other side, usually work with images in the *Optical Density* (OD) space. In this space, the intensity is linear with the amount of each stain absorbed by a sample. The OD of an image channel, $\mathbf{y}_c \in \mathbb{R}^{MN \times 1}$, is defined as $\mathbf{y}_c = -\log_{10}(\mathbf{i}_c/\mathbf{i}_c^0)$, where \mathbf{i}_c^0 denotes the incident light, and the division operation and $\log_{10}(\cdot)$ function are computed

element-wise. The observed OD image $\mathbf{Y} \in \mathbb{R}^{MN \times 3}$ has three OD channels, i.e.,

$$\mathbf{Y} = [\mathbf{y}_R \ \mathbf{y}_G \ \mathbf{y}_B] = \begin{bmatrix} \mathbf{y}_{1,:}^T \\ \vdots \\ \mathbf{y}_{MN,:}^T \end{bmatrix} = \begin{bmatrix} y_{1R} & y_{1G} & y_{1B} \\ \vdots & \vdots & \vdots \\ y_{MNR} & y_{MNG} & y_{MNB} \end{bmatrix}. \quad (1)$$

The Beer-Lambert law, for a slide stained with n_s stains, establishes that

$$\mathbf{Y}^T = \mathbf{M}\mathbf{C}^T + \mathbf{N}^T, \quad (2)$$

where $\mathbf{M} \in \mathbb{R}^{3 \times n_s}$ is the color-vector matrix,

$$\mathbf{M} = [\mathbf{m}_1 \ \dots \ \mathbf{m}_{n_s}] = \begin{bmatrix} \mathbf{m}_R^T \\ \mathbf{m}_G^T \\ \mathbf{m}_B^T \end{bmatrix} = \begin{bmatrix} m_{R1} & \dots & m_{Rn_s} \\ m_{G1} & \dots & m_{Gn_s} \\ m_{B1} & \dots & m_{Bn_s} \end{bmatrix} \in \mathbb{R}^{3 \times n_s}, \quad (3)$$

with each column \mathbf{m}_s in matrix \mathbf{M} being a unit ℓ_2 -norm stain color-vector containing the relative RGB color composition of the corresponding stain in the OD space, $\mathbf{C} \in \mathbb{R}^{MN \times n_s}$ is the stain concentration matrix,

$$\mathbf{C} = \begin{bmatrix} c_{11} & \dots & c_{1n_s} \\ \vdots & \ddots & \vdots \\ c_{MN1} & \dots & c_{MNn_s} \end{bmatrix} = \begin{bmatrix} \mathbf{c}_{1,:}^T \\ \vdots \\ \mathbf{c}_{MN,:}^T \end{bmatrix} = [\mathbf{c}_1 \ \dots \ \mathbf{c}_{n_s}], \quad (4)$$

with the s -th column $\mathbf{c}_s = (c_{1s}, \dots, c_{MN s})^T$, $s \in \{1, \dots, n_s\}$, representing the concentrations of the s -th stain, and the i -th row $\mathbf{c}_{i,:}^T = (c_{i1}, \dots, c_{in_s})$, $i = 1, \dots, MN$, representing the contribution of each stain to the i -th \mathbf{Y} pixel value, $\mathbf{y}_{i,:}$, and \mathbf{N} is a random matrix of size $MN \times 3$ with i.i.d. zero mean Gaussian components with variance β^{-1} , representing the noise introduced by the image capture system.

In the following section we use Bayesian modeling and inference to estimate both \mathbf{C} and \mathbf{M} , from \mathbf{Y}

3. Bayesian Modelling and Inference

Bayesian methods start with a prior distribution on the unknowns. In this paper we adopt the TV prior, which smooths the image noise while preserving its edges, for each one of the independent stain concentration vectors \mathbf{c}_s , that is,

$$p(\mathbf{C}|\boldsymbol{\alpha}) = \prod_{s=1}^{n_s} p(\mathbf{c}_s|\alpha_s) \propto \prod_{s=1}^{n_s} \exp[-\alpha_s \text{TV}(\mathbf{c}_s)], \quad (5)$$

with $\alpha_s > 0$ controlling the image smoothness. The TV function is defined for any \mathbf{c}_s , $s \in \{1, \dots, n_s\}$, as

$$\text{TV}(\mathbf{c}_s) = \sum_{i=1}^{MN} \sqrt{(\Delta_i^h(\mathbf{c}_s))^2 + (\Delta_i^v(\mathbf{c}_s))^2}, \quad (6)$$

where the operators $\Delta_i^h(\mathbf{c}_s)$ and $\Delta_i^v(\mathbf{c}_s)$ correspond to the horizontal and vertical first order differences of \mathbf{c}_s at pixel i , respectively.

The color-vector matrix $\mathbf{M} = [\mathbf{m}_1, \dots, \mathbf{m}_{n_s}]$ varies, as previously discussed, from image to image. However, images from the same laboratory usually have similar colors and we can benefit from this prior knowledge. Ruifrok *et al.* [4] proposed a procedure to obtain a laboratory dependant standard color-vectors. Although those vectors are not exact for every single image, they are representative and widely used. To take into account these considerations, we incorporate similarity to a reference color-vector matrix $\underline{\mathbf{M}} = [\underline{\mathbf{m}}_1, \dots, \underline{\mathbf{m}}_{n_s}]$ into the color-vector matrix prior model as

$$p(\mathbf{M}|\boldsymbol{\gamma}) = \prod_{s=1}^{n_s} p(\mathbf{m}_s|\gamma_s) \propto \prod_{s=1}^{n_s} \gamma_s^{\frac{3}{2}} \exp\left(-\frac{1}{2}\gamma_s\|\mathbf{m}_s - \underline{\mathbf{m}}_s\|^2\right), \quad (7)$$

where γ_s , $s = 1, \dots, n_s$, controls our confidence on the accuracy of $\underline{\mathbf{m}}_s$.

Finally, from the degradation model in (2), we have

$$p(\mathbf{Y}|\mathbf{M}, \mathbf{C}, \beta) = \prod_{i=1}^{MN} p(\mathbf{y}_{i,:}|\mathbf{M}, \mathbf{c}_{i,:}, \beta) = \prod_{i=1}^{MN} \mathcal{N}(\mathbf{y}_{i,:}|\mathbf{M}\mathbf{c}_{i,:}, \beta^{-1}\mathbf{I}_{3\times 3}). \quad (8)$$

With all these ingredients, we define the joint probability distribution as

$$p(\mathbf{Y}, \mathbf{C}, \mathbf{M}, \beta, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = p(\mathbf{Y}|\mathbf{C}, \mathbf{M}, \beta) p(\mathbf{C}|\boldsymbol{\alpha}) p(\mathbf{M}|\boldsymbol{\gamma}) p(\beta) p(\boldsymbol{\alpha}) p(\boldsymbol{\gamma}), \quad (9)$$

where $p(\boldsymbol{\gamma})$, $p(\boldsymbol{\alpha})$ and $p(\beta)$ are improper distributions of the form $p(w) \propto \text{const.}$

Following the Bayesian paradigm, inference will be based on the posterior distribution $p(\Theta|\mathbf{Y})$ with $\Theta = \{\mathbf{C}, \mathbf{M}, \beta, \boldsymbol{\alpha}, \boldsymbol{\gamma}\} = \{\mathbf{c}_1, \dots, \mathbf{c}_{n_s}, \mathbf{m}_1, \dots, \mathbf{m}_{n_s}, \beta, \alpha_1, \dots, \alpha_{n_s}, \gamma_1, \dots, \gamma_{n_s}\}$, the set of all unknowns.

Since the above posterior cannot be obtained in closed form, several approaches have been proposed to approximate it. In this paper we use the mean-field variational Bayesian model [25] to approximate $p(\Theta|\mathbf{Y})$ by the distribution $q(\Theta)$ of the form

$$q(\Theta) = q(\beta) \prod_{s=1}^{n_s} q(\mathbf{m}_s)q(\mathbf{c}_s)q(\alpha_s)q(\gamma_s), \quad (10)$$

where $q(\beta)$, $q(\alpha_s)$, $q(\gamma_s)$, $s = 1, \dots, n_s$, are assumed to be degenerate distributions. The optimal $q(\Theta)$ minimizes the Kullback-Leibler divergence [26] defined as

$$\mathbf{KL}(q(\Theta) \parallel p(\Theta|\mathbf{Y})) = \int q(\Theta) \log \frac{q(\Theta)}{p(\Theta|\mathbf{Y})} d\Theta \quad (11)$$

$$= \int q(\Theta) \log \frac{q(\Theta)}{p(\Theta, \mathbf{Y})} d\Theta + \log p(\mathbf{Y}). \quad (12)$$

The Kullback-Leibler divergence is always non negative and equal to zero if and only if $q(\Theta) = p(\Theta|\mathbf{Y})$.

Even with this factorization, the TV prior for \mathbf{C} hampers the evaluation of this divergence. To solve this problem, we define for α_s , \mathbf{c}_s , and any N -dimensions vector $\mathbf{u}_s \in (R^+)^{MN}$, $s = 1, \dots, n_s$, the functional

$$\mathcal{M}_s(\mathbf{c}_s, \mathbf{u}_s | \alpha_s) = \exp \left[-\frac{\alpha_s}{2} \sum_{i=1}^{MN} \frac{(\Delta_i^h(\mathbf{c}_s))^2 + (\Delta_i^v(\mathbf{c}_s))^2 + u_{is}}{\sqrt{u_{is}}} \right]. \quad (13)$$

Now, using the inequality for $w \geq 0$ and $z > 0$, $\sqrt{wz} \leq \frac{w+z}{2} \Rightarrow \sqrt{w} \leq \frac{w+z}{2\sqrt{z}}$, we can write

$$\exp[-\alpha_s \text{TV}(\mathbf{c}_s)] \geq \mathcal{M}_s(\mathbf{c}_s, \mathbf{u}_s | \alpha_s), \quad s = 1, \dots, n_s. \quad (14)$$

We, then, define

$$\mathcal{M}(\mathbf{C}, \mathbf{U} | \boldsymbol{\alpha}) = \prod_s \mathcal{M}_s(\mathbf{c}_s, \mathbf{u}_s | \alpha_s), \quad (15)$$

where $\mathbf{U} = [\mathbf{u}_1 \dots \mathbf{u}_{n_s}]$ and $F(\Theta, \mathbf{U}, \mathbf{Y}) = p(\mathbf{Y} | \mathbf{M}, \mathbf{C}, \beta) \mathcal{M}(\mathbf{C}, \mathbf{U}, \boldsymbol{\alpha}) p(\mathbf{M}, \boldsymbol{\gamma}) p(\beta) p(\boldsymbol{\alpha}) p(\boldsymbol{\gamma})$ to obtain the inequality

$$\log p(\Theta, \mathbf{Y}) \geq \log F(\Theta, \mathbf{U}, \mathbf{Y}). \quad (16)$$

We have then found a lower bound, $F(\Theta, \mathbf{U}, \mathbf{Y})$, for the joint probability $p(\Theta, \mathbf{Y})$ defined in (9). Utilizing this lower bound in (12), we minimize $\mathbf{KL}(q(\Theta) || F(\Theta, \mathbf{U}, \mathbf{Y}))$ instead of $\mathbf{KL}(q(\Theta) || p(\Theta | \mathbf{Y}))$.

As shown in [25], the mean field variational distribution approximation establishes that for each unknown $\theta \in \Theta$, $q(\theta)$ will have the form

$$q(\theta) \propto \exp \langle \log F(\mathbf{Y}, \mathbf{C}, \mathbf{M}, \beta, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \rangle_{q(\Theta \setminus \theta)}, \quad (17)$$

where $\Theta \setminus \theta$ represents all the variables in Θ except θ and $\langle \cdot \rangle_{q(\Theta \setminus \theta)}$ denotes the expected value calculated using the distribution $q(\Theta \setminus \theta)$. For variables with a degenerate posterior approximation, that is, for $\theta \in \{\beta, \alpha_1, \dots, \alpha_{n_s}, \gamma_1, \dots, \gamma_{n_s}\}$, the value where the posterior degenerates is

$$\hat{\theta} = \arg \max_{\theta} \langle \log F(\mathbf{Y}, \mathbf{C}, \mathbf{M}, \beta, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \rangle_{q(\Theta \setminus \theta)}. \quad (18)$$

For the rest of the variables, that is, for $\theta \in \{\mathbf{m}_1, \dots, \mathbf{m}_{n_s}, \mathbf{c}_1, \dots, \mathbf{c}_{n_s}\}$, when point estimates are required, the expected value, that is, $\hat{\theta} = \langle \theta \rangle_{q(\theta)}$ is used.

Let us now explicitly obtain analytical expressions for these estimates.

3.1. Concentration Update

According to (17), the estimation of the distributions on the concentrations $q(\mathbf{c}_s)$ is obtained as

$$q(\mathbf{c}_s) \propto \exp \langle \log F(\mathbf{Y}, \mathbf{C}, \mathbf{M}, \beta, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \rangle_{q(\Theta \setminus \mathbf{c}_s)}, \quad (19)$$

where

$$\langle \log F(\mathbf{Y}, \mathbf{C}, \mathbf{M}, \beta, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \rangle_{q(\Theta \setminus \mathbf{c}_s)} = \langle \log p(\mathbf{Y} | \mathbf{C}, \mathbf{M}, \beta) \rangle_{q(\Theta \setminus \mathbf{c}_s)} + \langle \log \mathcal{M}(\mathbf{C}, \mathbf{U}, \boldsymbol{\alpha}) \rangle_{q(\Theta \setminus \mathbf{c}_s)}. \quad (20)$$

To calculate the first term of the sum, we rewrite the distribution probability in (8) as

$$\begin{aligned}
p(\mathbf{Y}|\mathbf{M}, \mathbf{C}, \beta) &\propto \beta^{\frac{1}{2}} \prod_{i=1}^{MN} \exp \left(-\frac{1}{2} \beta \|\mathbf{y}_{i,:} - \sum_{s=1}^{n_s} c_{is} \mathbf{m}_s\|^2 \right) \\
&= \beta^{\frac{1}{2}} \prod_{i=1}^{MN} \exp \left(-\frac{1}{2} \beta \|\mathbf{y}_{i,:} - c_{is} \mathbf{m}_s - \sum_{k \neq s} c_{ik} \mathbf{m}_k\|^2 \right) \\
&= \beta^{\frac{1}{2}} \prod_{i=1}^{MN} \exp \left(-\frac{1}{2} \beta \sum_{s=1}^{n_s} \left[-2c_{is} \mathbf{m}_s^T \left(\mathbf{y}_{i,:} - \sum_{k \neq s} c_{ik} \mathbf{m}_k \right) + c_{is}^2 \|\mathbf{m}_s\|^2 \right] \right. \\
&\quad \left. + \text{const} \right), \tag{21}
\end{aligned}$$

where we have separated the contribution of the s -th stain to each observed image pixel from the rest of stains and const indicates the term which does not depend on \mathbf{c}_s .

Then, we calculate $\langle \log p(\mathbf{Y}|\mathbf{C}, \mathbf{M}|\beta) \rangle_{q(\Theta \setminus \mathbf{c}_s)}$ as

$$\begin{aligned}
\langle \log p(\mathbf{Y}|\mathbf{C}, \mathbf{M}|\beta) \rangle_{q(\Theta \setminus \mathbf{c}_s)} &= \left\langle -\frac{\beta}{2} \sum_{i=1}^{MN} \sum_{s=1}^{n_s} \left[-2c_{is} \mathbf{m}_s^T \left(\mathbf{y}_{i,:} - \sum_{k \neq s} c_{ik} \mathbf{m}_k \right) + c_{is}^2 \|\mathbf{m}_s\|^2 \right] \right\rangle \\
&= -\frac{\beta}{2} \left(-2\mathbf{c}_s^T \mathbf{z}^{-s} + \|\mathbf{c}_s\|^2 \langle \|\mathbf{m}_s\|^2 \rangle \right), \tag{22}
\end{aligned}$$

where \mathbf{z}^{-s} is a column vector with components

$$z_i^{-s} = \langle \mathbf{m}_s \rangle^T \mathbf{e}_{i,:}^{-s} \quad \text{with} \quad \mathbf{e}_{i,:}^{-s} = \mathbf{y}_{i,:} - \sum_{k \neq s} \langle c_{ik} \rangle \langle \mathbf{m}_k \rangle, \quad i = 1, \dots, MN. \tag{23}$$

From (13), we can calculate

$$\begin{aligned}
\langle \log \mathcal{M}(\mathbf{C}, \mathbf{U}, \boldsymbol{\alpha}) \rangle_{q(\Theta \setminus \mathbf{c}_s)} &= \left\langle -\frac{\boldsymbol{\alpha}_s}{2} \sum_{i=1}^{MN} \frac{(\Delta_i^h(\mathbf{c}_s))^2 + (\Delta_i^v(\mathbf{c}_s))^2 + \mathbf{u}_{is}}{\mathbf{u}_{is}} \right\rangle \\
&= -\frac{\boldsymbol{\alpha}_s}{2} (\mathbf{c}_s)^T \left[(\Delta^h)^T \mathbf{W}(\mathbf{u}) \Delta^h + (\Delta^v)^T \mathbf{W}(\mathbf{u}) \Delta^v \right] \mathbf{c}_s + \text{const}, \tag{24}
\end{aligned}$$

where $\mathbf{W}(\mathbf{u}_s)$ is a diagonal matrix of the form $\mathbf{W}(\mathbf{u}_s) = \text{diag}(u_{is}^{-1/2})$, for $i = 1, \dots, MN$.

Hence,

$$\begin{aligned}
\langle \log F(\mathbf{Y}, \mathbf{C}, \mathbf{M}|\beta, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \rangle_{q(\Theta \setminus \mathbf{c}_s)} &= -\frac{\beta}{2} \left(-2\mathbf{c}_s^T \mathbf{z}^{-s} + \|\mathbf{c}_s\|^2 \langle \|\mathbf{m}_s\|^2 \rangle \right) \\
&\quad -\frac{\boldsymbol{\alpha}_s}{2} (\mathbf{c}_s)^T \left[(\Delta^h)^T \mathbf{W}(\mathbf{u}) \Delta^h + (\Delta^v)^T \mathbf{W}(\mathbf{u}) \Delta^v \right] \mathbf{c}_s + \text{const}, \tag{25}
\end{aligned}$$

which, from (17), produces $q(\mathbf{c}_s) = \mathcal{N}(\mathbf{c}_s | \langle \mathbf{c}_s \rangle, \Sigma_{\mathbf{c}_s})$, where

$$\Sigma_{\mathbf{c}_s}^{-1} = \beta \langle \|\mathbf{m}_s\|^2 \rangle \mathbf{I}_{MN \times MN} + (\Delta^h)^T \mathbf{W}(\mathbf{u}_s) \Delta^h + (\Delta^v)^T \mathbf{W}(\mathbf{u}_s) \Delta^v \quad (26)$$

$$\langle \mathbf{c}_s \rangle = \beta \Sigma_{\mathbf{c}_s} \mathbf{z}^{-s}, \quad (27)$$

where Δ^h and Δ^v represent the convolution matrices associated with the first order horizontal and vertical differences, respectively. Note that the matrix $\mathbf{W}(\mathbf{u}_s)$ can be interpreted as a spatial adaptivity matrix since it controls the amount of smoothing at each pixel location depending on the strength of the intensity variation at that pixel, as expressed by the horizontal and vertical intensity gradient.

3.2. Color-Vector Update

In a similar way, using (23), we calculate the distribution of \mathbf{m}_s ,

$$\begin{aligned} \langle \log F(\mathbf{Y}, \mathbf{C}, \mathbf{M} | \beta, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \rangle_{q(\Theta \setminus \mathbf{m}_s)} &= \langle \log p(\mathbf{Y} | \mathbf{C}, \mathbf{M}, \beta) \rangle_{q(\Theta \setminus \mathbf{m}_s)} + \langle \log p(\mathbf{M}, \boldsymbol{\gamma}) \rangle_{q(\Theta \setminus \mathbf{m}_s)} \\ &= -\frac{\beta}{2} \left(\|\mathbf{m}_s\|^2 \sum_{i=1}^{MN} \langle c_{is}^2 \rangle - 2\mathbf{m}_s^T \sum_{i=1}^{MN} \langle c_{is} \rangle \mathbf{e}_{i,:}^{-s} \right) - \frac{1}{2} \gamma_s \|\mathbf{m}_s - \underline{\mathbf{m}}_s\|^2 + \text{const}, \end{aligned} \quad (28)$$

which, from (17), produces $q(\mathbf{m}_s) = \mathcal{N}(\mathbf{m}_s | \langle \mathbf{m}_s \rangle, \Sigma_{\mathbf{m}_s})$, where

$$\Sigma_{\mathbf{m}_s}^{-1} = \left(\beta \sum_{i=1}^{MN} \langle c_{is}^2 \rangle + \gamma_s \right) \mathbf{I}_{3 \times 3}, \quad (29)$$

$$\langle \mathbf{m}_s \rangle = \Sigma_{\mathbf{m}_s} \left(\beta \sum_{i=1}^{MN} \langle c_{is} \rangle \mathbf{e}_{i,:}^{-s} + \gamma_s \underline{\mathbf{m}}_s \right). \quad (30)$$

Notice that $\langle \mathbf{m}_s \rangle$ may not be a unitary vector even if $\underline{\mathbf{m}}_s$ is. To obtain unitary vectors, we can always replace $\langle \mathbf{m}_s \rangle$ by $\langle \mathbf{m}_s \rangle / \|\langle \mathbf{m}_s \rangle\|$ and $\Sigma_{\mathbf{m}_s}$ by $\Sigma_{\mathbf{m}_s} / \|\langle \mathbf{m}_s \rangle\|^2$.

3.3. U Update

The use of the majorization with the functional in (13) introduces a new set of parameters, \mathbf{U} , that need to be estimated along with the concentrations and the color-vectors matrix. To estimate the \mathbf{U} matrix, we need to solve, for each $s \in \{1, \dots, n_s\}$,

$$\hat{\mathbf{u}}_s = \arg \min_{\mathbf{u}_s} - \langle \log \mathcal{M}_s(\alpha_s, \mathbf{c}_s, \mathbf{u}_s) \rangle_{q(\mathbf{c}_s)}, \quad (31)$$

whose solution is given by

$$\hat{u}_{is} = \arg \min_{u_{is}} \frac{\langle (\Delta_i^h(\mathbf{c}_s))^2 + (\Delta_i^v(\mathbf{c}_s))^2 \rangle + u_{is}}{\sqrt{u_{is}}} = \langle \Delta_i^h(\mathbf{c}_s)^2 \rangle + \langle \Delta_i^v(\mathbf{c}_s)^2 \rangle. \quad (32)$$

3.4. Parameter Update

Finally, the estimates of the noise, concentration, and color-vectors parameters are obtained according (18) as

$$\hat{\beta}^{-1} = \frac{\text{tr}(\langle (\mathbf{Y}^T - \mathbf{M}\mathbf{C}^T)(\mathbf{Y}^T - \mathbf{M}\mathbf{C}^T)^T \rangle_{\mathbf{q}(\Theta)})}{3MN}, \quad (33)$$

$$\hat{\alpha}_s^{-1} = \frac{\text{tr} \left(\left((\Delta^h)^T (\Delta^h) + (\Delta^v)^T (\Delta^v) \right) \langle \mathbf{c}_s \mathbf{c}_s^T \rangle \right)}{MN}, \quad (34)$$

$$\hat{\gamma}_s^{-1} = \frac{\text{tr}(\langle (\mathbf{m}_s - \underline{\mathbf{m}}_s)(\mathbf{m}_s - \underline{\mathbf{m}}_s)^T \rangle)}{3}. \quad (35)$$

3.5. Calculating the expectations and concentration covariance matrices

To estimate the concentrations and color-vectors, the expectations $\langle c_{is}^2 \rangle$ in (30) and $\langle \|\mathbf{m}_s\|^2 \rangle$ in (26) need to be calculated. Also, the computation of the matrix $\Sigma_{\mathbf{c}_s}$, defined in (26), is an issue due to the size of WSI images. In this section we explicitly calculate the mentioned expected values and address the concentrations covariance matrix calculation issue.

Notice that $\langle c_{is}^2 \rangle$ can be calculated using (27) and $\langle \|\mathbf{m}_s\|^2 \rangle$ can be easily calculated from (29) resulting in

$$\sum_{i=1}^{MN} \langle c_{is}^2 \rangle = \sum_{i=1}^{MN} \langle c_{is} \rangle^2 + \text{tr}(\Sigma_{\mathbf{c}_s}), \quad \langle \|\mathbf{m}_s\|^2 \rangle = \|\langle \mathbf{m}_s \rangle\|^2 + \text{tr}(\Sigma_{\mathbf{m}_s}). \quad (36)$$

The matrix $\Sigma_{\mathbf{c}_s}$ must be explicitly calculated to find its trace and also to calculate \hat{u}_{is} . However, since its calculation is very intense, following [27], we approximate the covariance matrix as follows. We first approximate $\mathbf{W}(\mathbf{u}_s)$ using $\mathbf{W}(\mathbf{u}_s) \approx z(\mathbf{u}_s)\mathbf{I}$, where $z(\mathbf{u}_s)$ is calculated as the mean value of the diagonal values in $\mathbf{W}(\mathbf{u}_s)$, that is, $z(\mathbf{u}_s) = \frac{1}{MN} \sum_i \frac{1}{\sqrt{u_{is}}}$. We then use the approximation

$$\Sigma_{\mathbf{c}_s}^{-1} \approx \beta \langle \|\mathbf{m}_s\|^2 \rangle \mathbf{I}_{MN \times MN} + \alpha_s z(\mathbf{u}_s) (\Delta^h)^T (\Delta^h) + \alpha_s z(\mathbf{u}_s) (\Delta^v)^T (\Delta^v) = \mathbf{B}. \quad (37)$$

Note that the matrix \mathbf{B} is a block circulant matrix with circulant blocks (BCCB), thus, computing its inverse can be very efficiently performed in the discrete Fourier domain. Finally, we have

$$\begin{aligned} \langle \Delta_i^h(\mathbf{c}_s)^2 \rangle + \langle \Delta_i^v(\mathbf{c}_s)^2 \rangle &\approx (\Delta_i^h(\langle \mathbf{c}_s \rangle))^2 + (\Delta_i^v(\langle \mathbf{c}_s \rangle))^2 \\ &+ \frac{1}{MN} \text{tr} \left[\mathbf{B}^{-1} \times \left((\Delta^h)^T (\Delta^h) + (\Delta^v)^T (\Delta^v) \right) \right]. \end{aligned} \quad (38)$$

Algorithm 1 Variational Bayesian TV Blind Color Deconvolution

Require: Observed image \mathbf{I} and reference (prior) color-vector matrix $\underline{\mathbf{M}}$.Obtain the observed OD image \mathbf{Y} from \mathbf{I} and set $\langle \mathbf{m}_s \rangle^{(0)} = \underline{\mathbf{m}}_s$, $\Sigma_{\mathbf{m}_s}^{(0)} = \mathbf{0}$, $\Sigma_{\mathbf{c}_s}^{(0)} = \mathbf{0}$, $\langle \mathbf{c}_s \rangle^{(0)}$, $\forall s = 1, \dots, n_s$, from the matrix \mathbf{C} obtained as $\mathbf{C}^T = \underline{\mathbf{M}}^+ \mathbf{Y}^T$, with $\underline{\mathbf{M}}^+$ the Moore-Penrose pseudo-inverse of $\underline{\mathbf{M}}$, and $n = 0$.**while** convergence criterion is not met **do**

1. Set $n = n + 1$.
2. Obtain $\beta^{(n)}$, $\boldsymbol{\alpha}^{(n)}$ and $\boldsymbol{\gamma}^{(n)}$ from (33), (34) and (35), respectively.
3. Using $\langle \mathbf{c}_s \rangle^{(n-1)}$ and $\Sigma_{\mathbf{c}_s}^{(n-1)}$, $\forall s \in \{1, \dots, n_s\}$, update the new variational parameters $\hat{\mathbf{u}}_s^{(n)}$ from (32).
4. Using $\langle \mathbf{c}_s \rangle^{(n-1)}$, $\Sigma_{\mathbf{c}_s}^{(n-1)}$ and $\langle \mathbf{m}_s \rangle^{(n-1)}$, update the color-vectors $\langle \mathbf{m}_s \rangle^{(n)}$ and $\Sigma_{\mathbf{m}_s}^{(n)}$ from (30) and (29), $\forall s$.
5. Using $\langle \mathbf{m}_s \rangle^{(n)}$, $\Sigma_{\mathbf{m}_s}^{(n)}$ and $\hat{\mathbf{u}}_s^{(n)}$, update the concentrations $\Sigma_{\mathbf{c}_s}^{(n)}$ and $\langle \mathbf{c}_s \rangle^{(n)}$ from (26) and (27), $\forall s$.

end whileOutput the color-vector $\hat{\mathbf{m}}_s = \langle \mathbf{m}_s \rangle^{(n)}$ and the concentrations $\hat{\mathbf{c}}_s = \langle \mathbf{c}_s \rangle^{(n)}$.

3.6. Proposed Algorithm

Based on the previous derivations, we propose the Variational Bayesian TV Blind Color Deconvolution in Algorithm 1. The algorithm starts from the observed RGB image and a reference (prior) color vector matrix. Using this reference color-vector matrix as an starting point, the algorithm estimates in an iterative way, the model and variational parameters value, the distribution on the concentrations and distribution on the color-vectors.

The linear equations problem in (27), used in step 5 of Alg. 1, has been solved using the conjugate gradient approach while the color-vectors update in step 4 of the algorithm has been directly calculated from the equations due to the small size of the problem. On convergence, the algorithm returns point estimates of the color-vectors and concentrations as the mean value of the estimated distributions. Finally, from Alg. 1, an RGB image of each separated stain, $\hat{\mathbf{I}}_s^{\text{sep}}$, can be obtained as

$$(\hat{\mathbf{I}}_s^{\text{sep}})^T = \exp_{10}(-\hat{\mathbf{m}}_s \hat{\mathbf{c}}_s^T). \quad (39)$$

4. Experimental results

As previously indicated, blind color deconvolution algorithms are used for visual inspection and automatic classification of images. These may be conflicting goals since the most accurate color deconvolved images, in the sense of closeness to each single dye, are not usually the ones that lead to the highest performance in classification.

In this section, we will show that, depending on the number of components used in the deconvolution process, the proposed methodology can obtain either the most accurate color images or produce stains that lead to the highest classification performance. To do so, we have designed two set of experiments. In the first one, the proposed method is

applied on the *Warwick Stain Separation Benchmark* (WSSB) dataset [16] (a dataset where the ground-truth color-vectors are known) and its results are compared to classical and state-of-the-art deconvolution methods both visually and numerically. We will show that the proposed method outperforms the competing methods when two components are used. We also presents results on prostate cancer detection using the histopathological SICAPv1 database [28]. On this carefully annotated dataset, color deconvolution is used to separate H&E stains from which a set of features are extracted. Following [28], those features are then used to train a group of state-of-the-art supervised classification methods to distinguish between benign and pathological images. In this classification scenario we will show that the proposed method outperforms its competitors when three components are used.

The experiments carried out will then indicate that the introduced framework can be used for accurate reconstruction of original stains and to obtain better classification results depending on the number of stains used to decompose the image.

4.1. Color Deconvolution Experiments

In this first experiment, we assess the quality of the color deconvolution methods for accurate H&E separation. For this purpose, we used the *Warwick Stain Separation Benchmark* (WSSB) [16] dataset as a test-bed. WSSB contains 24 H&E stained images of different tissues (breast, colon and lung) from different laboratories which have been captured with different microscopes. For each image, its ground truth stain color-vector matrix, \mathbf{M}_{GT} , was manually obtained by medical experts. The median value of a set of image pixels with a single stain was used. The pixels were selected based on biological structures: nuclei for hematoxylin and cytoplasm for eosin. The ground truth concentrations were obtained in [16] from the ground-truth color-vector matrix as $\mathbf{C}_{GT}^T = \mathbf{M}_{GT}^+ \mathbf{Y}^T$. From those ground-truth concentrations and color-vectors, a RGB image for each stain separately is obtained by applying (39). A sample breast image from the WSSB dataset is shown in Fig. 1a and its ground truth RGB separation is depicted in Fig. 1b.

The proposed framework was compared with the classical non-blind method by Ruifrok *et al.*[4], the classical blind color deconvolution by Macenko *et al.*[10], and the state-of-the-art methods by Vahadane *et al.*[8], Alsubaie *et al.*[16], and Hidalgo-Gavira *et al.*[21]. The proposed Algorithm 1 was run on this dataset until the criterion $\| \langle \mathbf{c}_s \rangle^{(n)} - \langle \mathbf{c}_s \rangle^{(n-1)} \|^2 / \| \langle \mathbf{c}_s \rangle^{(n)} \|^2 < 10^{-5}$ was met by all stains, that is, $s = 1, 2, \dots, n_s$. Since different tissues may have different color characteristics, the reference (prior) color-vector matrix \mathbf{M} was obtained by selecting, by non-medical experts, a single pixel from each type of tissue, breast, colon and lung, containing mainly hematoxylin and another pixel containing mainly eosin. When a third component is utilized, following the most commonly used implementation of Ruifrok’s method [29], the color representing the third component of each reference color-vector was calculated as the complementary of the first two colors. For all the competing algorithms, parameters were selected following the recommendations in the original paper or the reference software freely available.

The resulting H-only and E-only images were compared both visually and numerically by means of the *Peak Signal to Noise Ratio* (PSNR) and *Structural Similarity* (SSIM) metrics. Numerical results, presented in Table 1, show that using two stains, i. e., $n_s = 2$, the

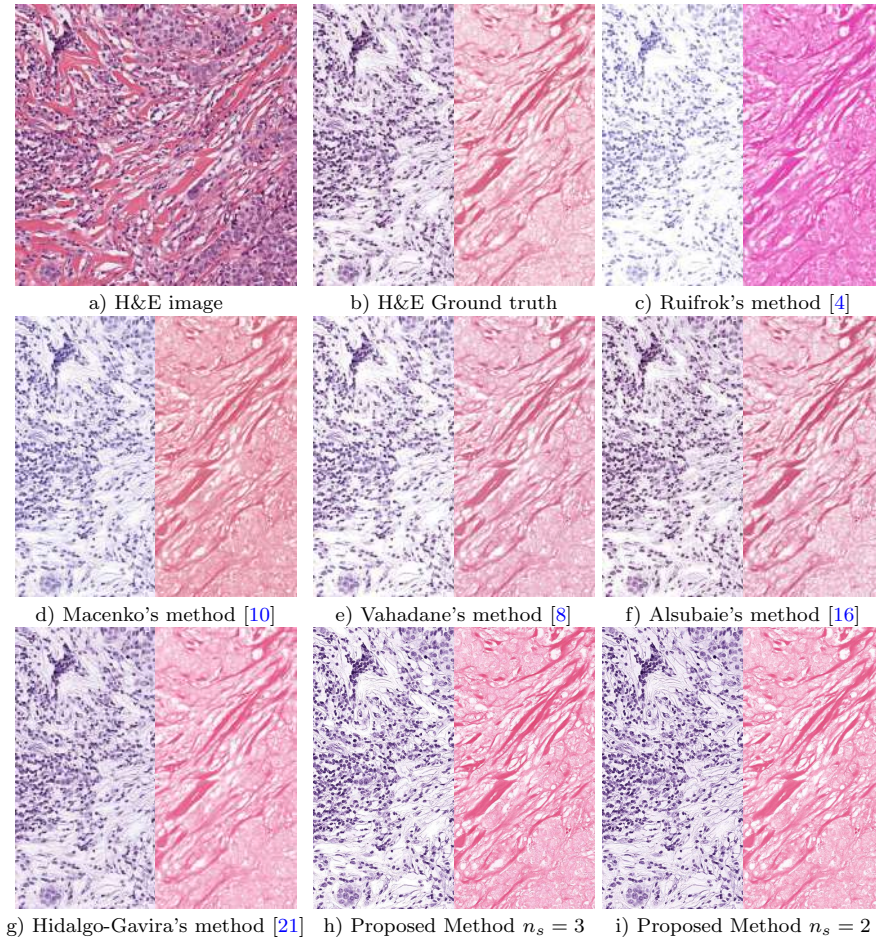


Figure 1: A breast H&E stained image from the WSSB dataset in [16], its ground truth separated H-only and E-only images, and its separation results by the competing and proposed methods. Hematoxylin and eosin separations are presented on the left and right hand sides of each image, respectively.

proposed method produces higher PSNR and SSIM values than the competing models except for SSIM in lung images where a slightly higher value is obtained by the Hidalgo-Gavira's method.

The separated H- and E-only images from the observed image in Fig. 1a are shown in Fig. 1(c-i). The proposed method and the methods by Vahadane and Hidalgo-Gavira produce H&E images very similar to the ground truth separation in Fig. 1b. Note also that the images obtained by Hidalgo-Gavira's method and the proposed one with two and three components are very similar. Notice, however, that the H-only images produced by the proposed method (Fig. 1h-i) are sharper and nuclei are clearer which will be useful, as we will later see, for classification. Both methods use the same prior model on the color-vectors, but they differ on the prior on the concentrations. While Hidalgo-Gavira's method uses a SAR model, ours uses a TV-based one. This model produces sharper images than those

Table 1: PSNR and SSIM for the different methods on the WSSB dataset [16].

Image	Stain	Ruifrok's method [4]		Macenko's method [10]		Vahadane's method [8]		Alsubaie's method [16]		Hidalgo-Gavira's method [21]		Proposed method $n_s = 3$		Proposed method $n_s = 2$	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Colon	H	22.27	0.8141	23.91	0.8095	25.83	0.8851	21.11	0.7241	28.57	0.9542	24.83	0.9005	28.62	0.9544
	E	20.70	0.7456	21.55	0.6365	26.29	0.8904	21.94	0.8540	27.58	0.9139	25.97	0.8695	27.60	0.9161
Breast	H	15.27	0.6215	26.24	0.9552	25.46	0.9239	24.60	0.8068	28.81	0.9528	27.71	0.9538	29.14	0.9560
	E	17.66	0.7644	23.62	0.9336	27.68	0.9550	25.92	0.9380	26.60	0.9464	26.84	0.9510	26.76	0.9492
Lung	H	22.47	0.7987	19.52	0.7389	25.87	0.8912	20.62	0.5551	32.91	0.9763	25.00	0.8374	33.10	0.9757
	E	22.05	0.7734	18.09	0.5088	25.53	0.8195	23.95	0.8939	30.77	0.9306	25.81	0.8426	31.02	0.9353
Mean	H	20.00	0.7448	23.22	0.8345	25.72	0.9100	22.11	0.6953	30.10	0.9611	25.85	0.8972	30.29	0.9621
	E	20.14	0.7611	21.08	0.6930	26.50	0.8883	23.94	0.8953	28.32	0.9303	26.21	0.8877	28.46	0.9336

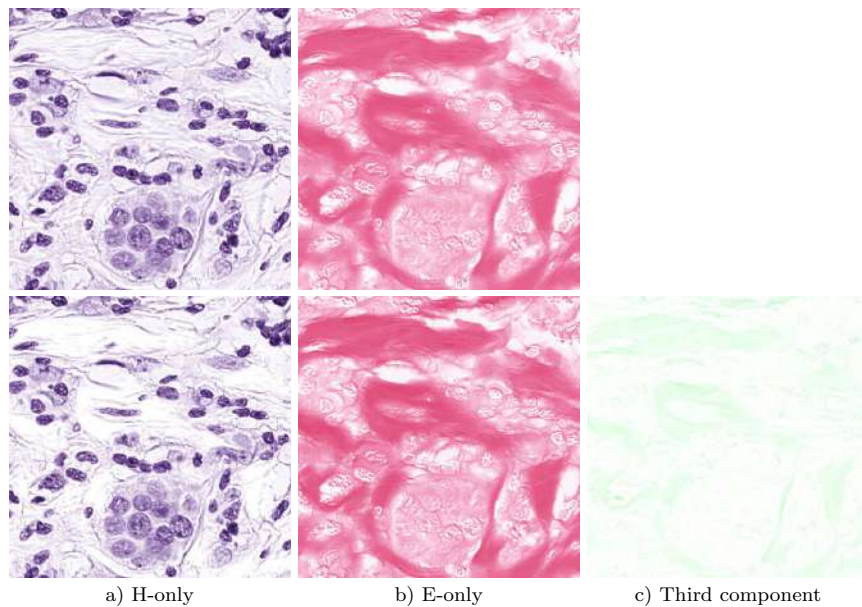


Figure 2: Detail of the H-only, E-only and third component separations of the bottom left corner of Fig. 1a obtained with the proposed method using two components (top) and three components (bottom).

obtained by Hidalgo-Gavira's method and is richer in details than Vahadane's method all the above is reflected in higher PSNR and SSIM values, see Table 1.

When a third component is used, the separation obtained by the proposed method, see Fig. 1h, is not so close to the ground-truth. Zoomed in areas of the bottom left corner of Fig. 1(h-i) are shown in Figure 2 for a better visual inspection. Colors are visually similar to the ones obtained when using two components, but some pixel information, specially from the background in the hematoxylin band, has been displaced to the third component. It can be observed that the third component has bright values, that is, only a small fraction of the information originally in the other bands is captured by this one, and nuclei in the H-only image appear brighter and are more clearly separated when three components are used, which will be extremely useful for classification. However, this implies a separation from the ground-truth images and, hence, lower values of PSNR and SSIM. In spite of the lower objective quality measure values, the separation in three components leads, as we will

Table 2: Computational time in seconds for the different methods on the WSSB dataset [16].

Method	Ruifrok [4]	Macenko [10]	Vahadane [8]	Alsubaie [16]	Hidalgo-Gavira [21]	Proposed $n_s = 3$	Proposed $n_s = 2$
Whole Dataset	147.68	141.47	375.10	210.13	357.03	877.67	507.28
Mean per image	6.15	5.89	15.62	8.75	14.87	36.56	21.13

see in the next section, to a better classification.

To conclude this section, Table 2 contains a computational time comparison between the competing methods. The method by Ruifrok is the fastest one. As complexity increases, the methods require higher computational time. Method by Vahadane requires as much time as the method by Hidalgo-Gavira but achieves lower PSNR and SSIM values. The proposed method takes longer than the competing ones but the higher computational burden is accompanied by higher figures-of-merit as already shown in Table 1. Note, also, that the proposed method estimates the model parameters together with the color-vector matrix and the concentrations, increasing the running time but making the method parameter free.

4.2. Prostate Cancer Classification Experiments

In this section we study how the use of different stain deconvolution methods affects the performance of classifiers. We use the SICAPv1 database, a prostate cancer histopathological database recently presented in [28]. The database contains 79 H&E WSIs from 48 patients scanned at 40x magnification, 19 correspond to benign prostate tissue biopsies (negative class) and 60 to pathological prostate tissue biopsies (positive class). In each pathological WSI, malignant regions were annotated by expert pathologists. The whole dataset was divided into a training set of 60 WSI (17 benign and 43 pathological), and a test set of 19 WSI (2 benign and 17 pathological). The images were downsampled to 10x scale and those in the training set were divided into patches of size 512×512 pixels and 1024×1024 pixels with a 50% overlap. Using this scale and patch size it is possible to capture complete glands in the 512×512 patches. Patches containing more than a 75% of background were discarded. Benign patches were extracted from benign WSI. Malignant patches were considered only if they contain at least a 25% of malignant tissue. Following [28] we will use cross-validation on this subset of the training set to assess the performance of the classifiers. Figure 3 shows an example of malignant patches with the areas annotated by the pathologist. In this experiment we only consider the dataset with patches of size 1024×1024 since it produced the best results in the patch classification experiments carried out in [28]. This training dataset contains 1909 patches from benign WSIs and 344 from pathological ones.

The dataset was color deconvolved using the proposed and competing methods. The H&E concentration image in the OD space was used to extract features to be utilized as input to the classifiers. Following [28], we used the concatenation of Local Binary Patterns Variance (LBPV) [30] and Geodesic granulometries (GeoGran) features [28]. LBPV features capture the texture and contrast information from the hematoxylin. GeoGran is an H&E granulometry based descriptor in the OD space recently proposed in [28] for prostate cancer classification. It encodes the structure of the glands by recovering, from the hematoxylin, the structure of the nuclei which formed the gland frontiers (those that enclosed their lumen

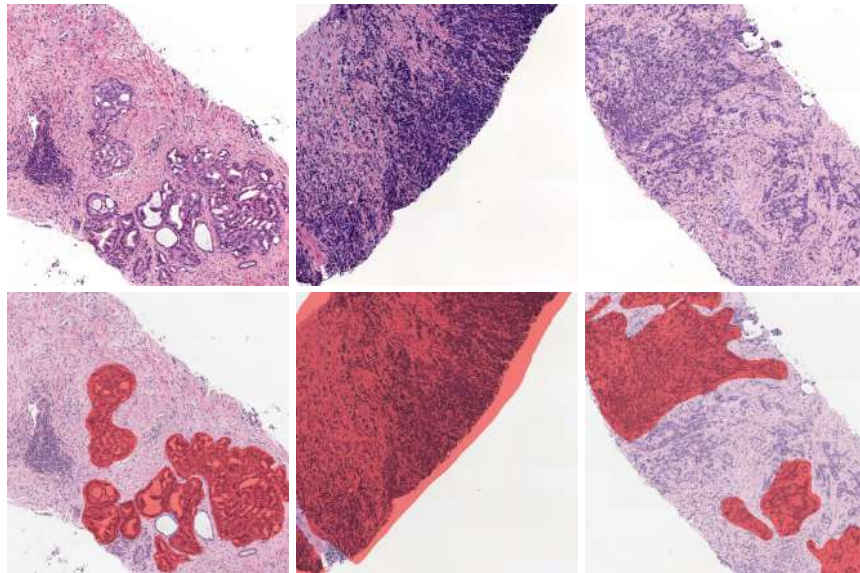


Figure 3: Top row: Patches extracted from SICAPv1 database. Bottom row: The same patches with the pathological areas annotated by the pathologist colored in red.

Table 3: AUC of the proposed and competing deconvolution methods with different classifiers.

Method	RF	GP	XgBoost	DGP
Ruifrok [4]	0.9789±0.0187	0.9855±0.0089	0.9764±0.0218	0.9737±0.0239
Macenko [10]	0.9315±0.0273	0.9535±0.0276	0.9425±0.0209	0.8802±0.0792
Vahadane [8]	0.9222±0.0318	0.9479±0.0321	0.9295±0.0325	0.9420±0.0436
Alsubaie [16]	0.9262±0.0586	0.9442±0.0294	0.9246±0.0612	0.9344±0.0581
Hidalgo-Gavira [21]	0.9157±0.0528	0.9542±0.0332	0.9228±0.0540	0.8997±0.0810
Proposed $n_s = 2$	0.9242±0.0579	0.9498±0.0332	0.9294±0.0824	0.9249±0.0638
Proposed $n_s = 3$	0.9798±0.0174	0.9856±0.0082	0.9797±0.0160	0.9718±0.0208

and cytoplasm). It also utilizes how distinguishable is, in the eosin, the lumen and nuclei structure from the rest of the stroma. This information is relevant to discriminate between pathological and benign tissues. The combination of LBPV and GeoGran features, which obtained the best classification results in the mentioned paper, allows to collect texture and structural information in the image, creating a descriptor able to accurately classify histopathological images.

A set of shallow and deep classifiers were trained with those descriptors and their results were compared. We used Random Forest (RF) [31], Extreme Gradient Boosting (XgBoost) [32], Gaussian Processes (GP)[33] and Deep Gaussian Processes (DGP)[34]. The tree-based ensemble models and the shallow and Deep GP can capture complex patterns in data and they are state-of-art classifiers. RF and XgBoost are configured with 1000 estimators and maximum depth of 20 and 30, respectively. A learning rate of 0.01 is chosen for XgBoost. Following the same approach as in [28] we use variational inference on a single-layer GP classifier with a RBF kernel [35]. For DGP, doubly stochastic variational inference [36] in a

Table 4: Accuracy of the proposed and competing deconvolution methods with different classifiers.

Method	RF	GP	XgBoost	DGP
Ruifrok [4]	0.9408±0.0301	0.9512±0.0272	0.9324±0.0505	0.9349±0.0337
Macenko [10]	0.8656±0.0277	0.8883±0.0561	0.8904±0.0205	0.8043±0.0399
Vahadane [8]	0.8870±0.0284	0.8826±0.0531	0.8830±0.0299	0.8996±0.0317
Alsubaie [16]	0.8825±0.0557	0.8793±0.0438	0.8730±0.0769	0.8885±0.05985
Hidalgo-Gavira [21]	0.8799±0.0105	0.8706±0.0445	0.8881±0.0673	0.8693±0.0810
Proposed $n_s = 2$	0.8914±0.0579	0.9029±0.0426	0.8910±0.0824	0.8797±0.0649
Proposed $n_s = 3$	0.9422±0.0375	0.9519±0.0319	0.9420±0.0339	0.9349±0.0257

Table 5: Accuracy of the proposed methods with different classifiers in train and test.

Method	RF	GP	XgBoost	DGP
$n_s = 2$ train	0.9789±0.0036	0.9794±0.0030	0.9697±0.0033	0.9401±0.0166
$n_s = 2$ test	0.8914±0.0579	0.9029±0.0426	0.8910±0.0824	0.8797±0.0649
$n_s = 3$ train	0.9774±0.0026	0.9878±0.0041	0.9796±0.0001	0.9605±0.0059
$n_s = 3$ test	0.9422±0.0375	0.9519±0.0319	0.9420±0.0339	0.9349±0.0257

three-layer classifier with RBF kernel and 100 inducing points per layer was used.

For each classifier, a five-fold cross-validation was applied to compare its performance with each deconvolution method. To avoid correlation between training and test sets, patches from the same image and patient were assigned to the same fold. Since the training set has more benign than pathological patches, an usual scenario on medical applications, balanced classifiers were built with all the pathological instances and a subset of the benign ones. The final prediction will be the average of the predictions of each classifier. The area under the ROC curve (AUC) obtained by the different deconvolution methods and classifiers is presented in Figure 4 and Table 3. Accuracy is shown in Table 4.

From Table 3, the best results are obtained using the proposed method with $n_s = 3$ and GP, with an AUC of 0.9856. The proposed method with $n_s = 3$ also obtains the best results among the shallow classifiers being the Ruifrok’s method the one obtaining the best result with the DGP classifier. When the proposed method is run with only two components results are also competitive but not as good as the ones obtained with three components. The curves in Figure 4 clearly show the advantage of the proposed $n_s = 3$ method and Ruifrok’s over the others. Average results of the method with $n_s = 2$ are also visible. From Table 4, the proposed method with $n_s = 3$ reaches the highest accuracy for all the classification methods. Notice that Ruifrok’s method was used for color deconvolution in [28] and so the figures of merits reported in the first line of Table 3 coincide with those reported in Table 5 in [28]. Finally, we would like to mention that in [28] a comparison with the deep learning methods VGG19, Inception v3, and Xception was carried out. The deep learning methods use as input the original RGB images, so the values reported for them in [28] are valid here. GPs and DGPs perform similarly and are competitive to VGG19, the best performing deep learning method in [28].

To assess the generalization capability of our model, we show in Table 5 the accuracy of the proposed method obtained for the train and test sets when performing cross validation.

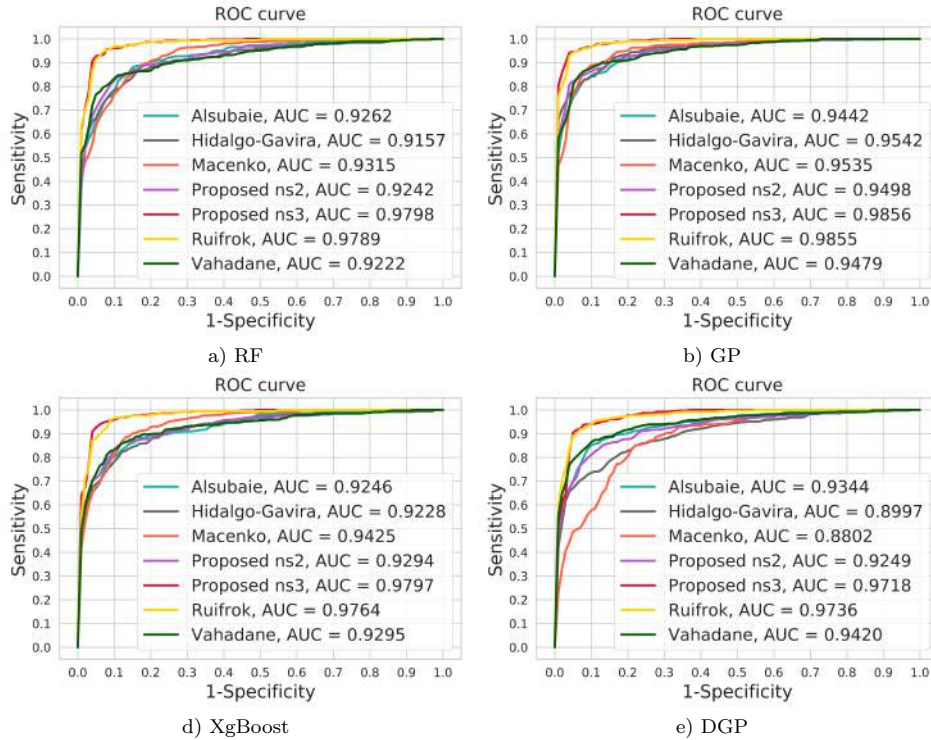


Figure 4: ROC curves for the competing methods and classifiers. Each sub-image contains all deconvolution methods AUC for a single classifier.

The use of $n_s = 2$ induced a higher overfitting to the train data in all the classification methods, reducing their generalization capability. For the GP and DGP models, Figure 5 includes the evolution of accuracy in train and test during the training procedure. Both GP and DGP models obtain a high accuracy from the beginning of the training and quickly converge. The overfitting when using $n_s = 2$ is visible in both models. The values obtained in training data using $n_s = 2$ and $n_s = 3$ are similar while the results obtained in testing data with $n_s = 2$ are much lower than the ones obtained with $n_s = 3$.

For classification, the use of a third component capturing residual information is clearly an advantage although the obtained images are not as close to the ground-truth separations as those obtained using $n_s = 2$. As seen in section 4.1, the third component is mainly capturing background information from the hematoxylin channel. An example of component concentration values in the OD space, which are used to extract the features, is shown in Fig. 6. The hematoxylin is used to extract LBPV and GeoGran features, that is, textures and nuclei structure. Due to the prostate tissue characteristics, the cytoplasm captures eosin and, partly, hematoxylin, so it appears also on the background of the hematoxylin band (see Fig. 6a). When three components are used, this background information is displaced to the third component. This also leads to a clearer hematoxylin (Fig. 6c) where nucleus information, belonging to the gland frontiers, is enhanced while the nucleus information

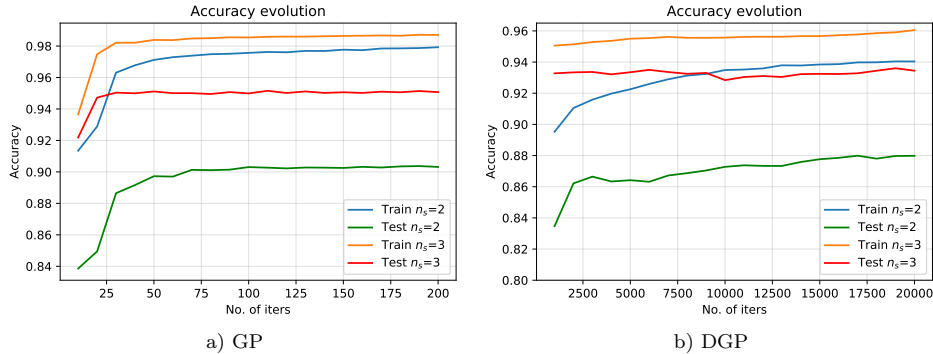


Figure 5: Train and test accuracy during the training procedure. a) GP model. b) DGP model.

belonging to the stroma (non-discriminative) appears in the third component. This allows to obtain less noisy features.

The eosin band is used to obtain GeoGran features to capture stroma information and identifies whether is invaded by nuclei or not. The use of three components makes the eosin band slightly more contrasted, which allows to obtain better descriptors. The joint use of descriptors extracted from hematoxylin and eosin bands by the proposed method using three component leads to an increased classification performance. The use of the TV prior, which produces sharper edges, also helps the feature extractors and, hence, the classifier.

4.2.1. Whole slide image evaluation

Our ultimate goal is to analyse full WSI images. To extend patch-wise classification to WSI classification, each WSI was split into overlapping patches. For each pixel, the probability of being cancerous was estimated by bilinearly interpolating the predicted probabilities of its four closest patches (using Euclidean distance to the center of the patches). A pixel-wise probability map was then obtained for each WSI. To assess the proposed method performance on this task, we deconvolved the train and test sets using $n_s = 3$. The GP classifier was then trained with the 60 images of the training set and used to predict the 19 WSIs in the test set. To obtain a better map resolution, 512×512 patches were used with 75% overlap. Figure 7 illustrates the result on a WSI of the test set. Probability maps are represented as heat maps. Red and blue colors are assigned to highest and lowest probabilities of being cancerous, respectively. The obtained probabilities correctly identify the annotated areas. Figure 8 shows zoomed in regions of interest.

5. Conclusions

In this work we have presented a framework for blind color deconvolution and classification of histological images. In this framework, we have developed a novel variational Bayesian blind color deconvolution algorithm which automatically estimates the concentration of stains, the color-vector matrix, and all the model parameters. It takes into account the spatial relations between pixels by means of a TV prior model, as well as the similarity

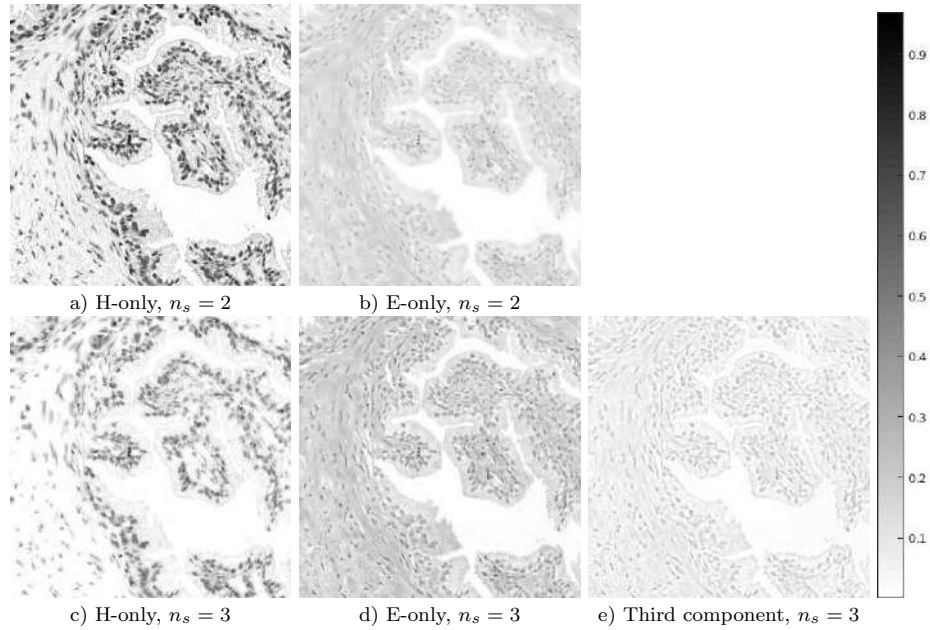


Figure 6: Detail of the H-only, E-only and third component concentration values in the OD space for a patch of SICAPv1 database obtained by the proposed method using two (top) and three components (bottom). The color-map of the images is inverted for a better visualization.

to a reference color-vector matrix. The use of the non-quadratic TV energy helps to reduce the noise in the images while preserving sharp edges.

For H&E stained images, color deconvolution with two components can be used in order to capture all stain details when visual inspection is needed. Classification algorithms, however, benefit from a clearer separation between classes. The use of a third, residual, component helps that separation by capturing information that is not completely explained by only one of the two stains. We found that, when using a third component, we obtain a clearer hematoxylin background, while nucleus information is enhanced and nuclei appear more clearly. The eosin is not severely modified, but the contrast of the image is increased which meliorates the discrimination power of this band. The use of a third component reduces the SSIM and PSNR values, but it helps the geodesic and LBPV descriptors to extract the relevant information and leads to better classification results.

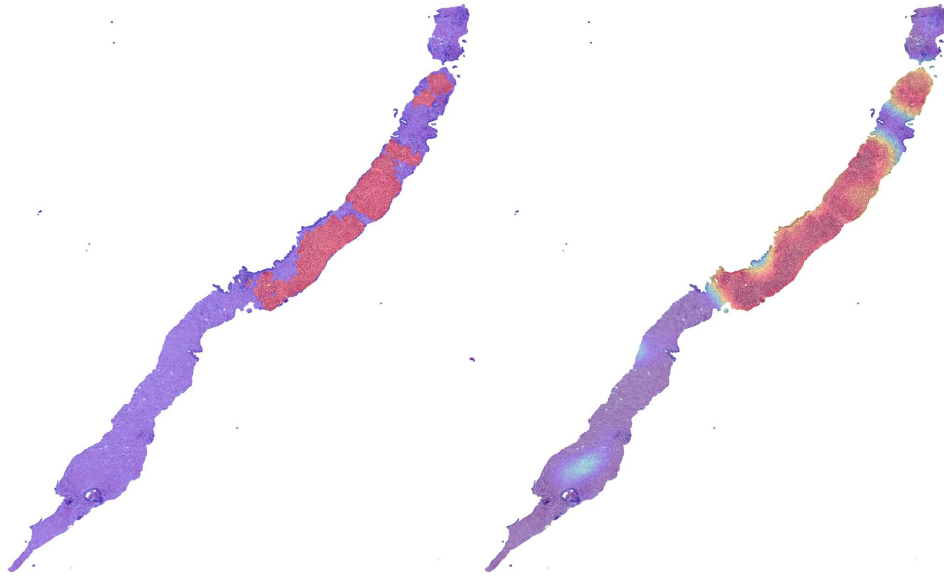


Figure 7: Full WSI comparison. Left: Areas annotated by the pathologists. Right: Probability maps (heat maps) obtained by the proposed method with $n_s = 3$ and GP classifier with 512×512 patches.

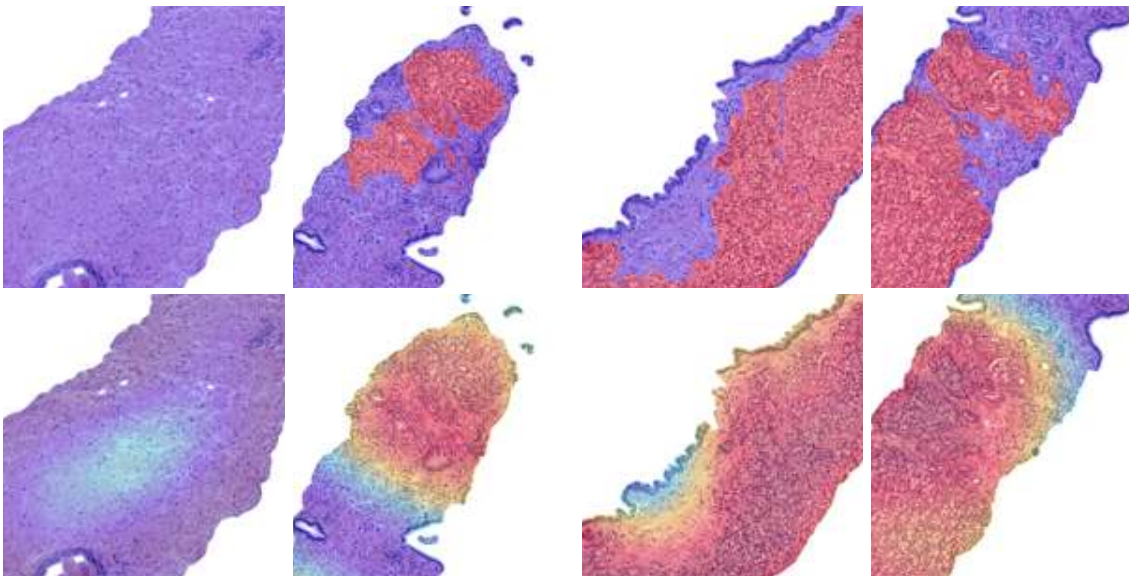


Figure 8: Regions of interest from Figure 7. Top row: annotations by a pathologist. Bottom row: Probability maps (heat maps) obtained.

6. References

References

- [1] A. H. Fischer, K. A. Jacobson, J. Rose, R. Zeller, *Hematoxylin and Eosin Staining of Tissue and Cell Sections*, Cold Spring Harbor Protocols (2008).
- [2] A. Rabinovich, S. Agarwal, C. Laris, J. H. Price, S. J. Belongie, Unsupervised color decomposition of histologically stained tissue samples, in: *Advances in Neural Information Processing Systems*, 2004, pp. 667–674.
- [3] T. A. A. Tosta, P. R. de Faria, L. A. Neves, M. Z. do Nascimento, Computational normalization of H&E-stained histological images: Progress, challenges and future potential, *Artificial Intelligence in Medicine* 95 (2019) 118 – 132.
- [4] A. C. Ruifrok, D. A. Johnston, Quantification of histochemical staining by color deconvolution, *Analytical and quantitative cytology and histology* 23 (2001) 291–299.
- [5] P. A. Bautista, Y. Yagi, Staining correction in digital pathology by utilizing a dye amount table, *Journal of digital imaging* 28 (3) (2015) 283–294.
- [6] T. Abe, H. Haneishi, P. A. Bautista, Y. Murakami, M. Yamaguchi, N. Ohyama, Y. Yagi, Color correction of red blood cell area in H&E stained images by using multispectral imaging, in: *4th European Conference on Colour in Graphics, Imaging, and Vision and 10th International Symposium on Multi-spectral Colour Science, CGIV*, 2008, pp. 432–436.
- [7] E. Reinhard, M. Adhikhmin, B. Gooch, P. Shirley, Color transfer between images, *IEEE Computer Graphics and Applications* 21 (5) (2001) 34–41.
- [8] A. Vahadane, T. Peng, A. Sethi, S. Albarqouni, L. Wang, M. Baust, K. Steiger, A. M. Schlitter, I. Esposito, N. Navab, Structure-preserving color normalization and sparse stain separation for histological images, *IEEE Transactions on Medical Imaging* 35 (2016) 1962–1971.
- [9] J. Xu, L. Xiang, G. Wang, S. Ganesan, M. Feldman, N. N. Shih, H. Gilmore, A. Madabhushi, Sparse non-negative matrix factorization (SNMF) based color unmixing for breast histopathological image analysis, *Computerized Medical Imaging and Graphics* 46 (2015) 20–29.
- [10] M. Macenko, M. Niethammer, et al., A method for normalizing histology slides for quantitative analysis, in: *International Symposium on Biomedical Imaging (ISBI)*, 2009, pp. 1107–1110.
- [11] M. T. McCann, J. Majumdar, et al., Algorithm and benchmark dataset for stain separation in histology images, in: *International Conference on Image Processing (ICIP)*, 2014, pp. 3953–3957.
- [12] D. Carey, V. Wijayathunga, A. Bulpitt, D. Treanor, A novel approach for the colour deconvolution of multiple histological stains, in: *Proceedings of the 19th Conference of Medical Image Understanding and Analysis*, 2015, pp. 156–162.
- [13] M. Gavrilovic, J. C. Azar, et al., Blind color decomposition of histological images, *IEEE Transactions on Medical Imaging* 32 (2013) 983–994.
- [14] A. M. Khan, N. Rajpoot, D. Treanor, D. Magee, A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution, *IEEE Transactions on Biomedical Eng.* 61 (6) (2014) 1729–1738.
- [15] N. Alsubaie, S. E. A. Raza, N. Rajpoot, Stain deconvolution of histology images via independent component analysis in the wavelet domain, in: *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, 2016, pp. 803–806.
- [16] N. Alsubaie, N. Trahearn, S. E. A. Raza, D. Snead, N. Rajpoot, Stain deconvolution using statistical analysis of multi-resolution stain colour representation, *PLOS ONE* 12 (2017) e0169875.
- [17] N. Trahearn, D. Snead, I. Cree, N. Rajpoot, Multi-class stain separation using independent component analysis, in: *Medical Imaging 2015: Digital Pathology*, 2015, p. 94200J.
- [18] L. Astola, Stain separation in digital bright field histopathology, in: *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, 2016, pp. 1–6.
- [19] Y. Zheng, Z. Jiang, H. Zhang, F. Xie, J. Shi, C. Xue, Adaptive color deconvolution for histological WSI normalization, *Computer Methods and Programs in Biomedicine* 170 (2019) 107–120.

- [20] S. Roy, A. K. Jain, S. Lal, J. Kini, A study about color normalization methods for histopathology images, *Micron* 114 (2018) 42–61.
- [21] N. Hidalgo-Gavira, J. Mateos, M. Vega, R. Molina, A. K. Katsaggelos, Variational Bayesian blind color deconvolution of histopathological images, *IEEE Transactions on Image Processing* accepted for publication.
- [22] M. Vega, J. Mateos, R. Molina, A. K. Katsaggelos, Variational Bayes color deconvolution with a total variation prior, in: *27th European Signal Processing Conference, EUSIPCO 2019*, 2019, p. TuEP3.7.
- [23] S. Villena, M. Vega, R. Molina, A. Katsaggelos, A non-stationary image prior combination in super-resolution, *Digital Signal Processing* 32 (2014) 1–10.
- [24] P. Ruiz, X. Zhou, J. Mateos, R. Molina, A. Katsaggelos, Variational Bayesian blind image deconvolution: A review, *Digital Signal Processing* 47 (2015) 116–127.
- [25] C. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006, pp. 454–455.
- [26] S. Kullback, *Information Theory and Statistics*, Dover Pub., 1959.
- [27] S. D. Babacan, R. Molina, A. K. Katsaggelos, Parameter estimation in TV image restoration using variational distribution approximation, *IEEE Transactions Image Processing* (2008) 326–339.
- [28] A. E. Esteban, M. Lopez-Perez, A. Colomer, M. A. Sales, R. Molina, V. Naranjo, A new optical density granulometry-based descriptor for the classification of prostate histological images using shallow and deep Gaussian processes, *Computer Methods and Programs in Biomedicine* 178 (2019) 303–317.
- [29] G. Landini, Colour deconvolution, <https://blog.bham.ac.uk/intellimic/g-landini-software/colour-deconvolution/>, accessed: 2019-10-30.
- [30] Z. Guo, L. Zhang, D. Zhang, Rotation invariant texture classification using LBP variance (LBPV) with global matching, *Pattern Recognition* 43 (3) (2010) 706–719.
- [31] M. Valkonen, K. Kartasalo, K. Liimatainen, M. Nykter, L. Latonen, P. Ruusuvoori, Metastasis detection from whole slide images using local features and random forests, *Cytometry Part A* 91 (6) (2017) 555–565.
- [32] A. Pimkin, G. Makarchuk, V. Kondratenko, M. Pisov, E. Krivov, M. Belyaev, Ensembling neural networks for digital pathology images classification and segmentation, *Lecture Notes in Computer Science* 10882 LNCS (2018) 877–886.
- [33] C. Rasmussen, C. Williams, *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*, The MIT Press, 2006.
- [34] A. Damianou, N. Lawrence, Deep Gaussian processes, *Journal of Machine Learning Research* 31 (2013) 207–215.
- [35] M. Opper, C. Archambeau, The variational Gaussian approximation revisited, *Neural Comput.* 21 (3) (2009) 786–792.
- [36] H. Salimbeni, M. Deisenroth, Doubly stochastic variational inference for deep Gaussian processes, in: *Advances in Neural Information Processing Systems*, 2017, pp. 4591–4602.

CHAPTER 3

Reference-based Blind Color Deconvolution Using General Super Gaussian Priors

3.1 JCR Publication Details

Authors: Fernando Pérez-Bueno, Miguel Vega, María A. Sales, José Aneiros-Fernández, Valery Naranjo, Rafael Molina, Aggelos K. Katsaggelos

Title: Blind Color Deconvolution, Normalization, and Classification of Histological Images Using General Super Gaussian Priors and Bayesian Inference

Reference: Computer Methods and Programs in Biomedicine, Volume 211, 2021, 106453

Status: Published

DOI: <https://doi.org/10.1016/j.cmpb.2021.106453>

Quality indices:

- Impact Factor (JCR 2021): 7.027
 - Rank: 12/109 (Q1) in Computer Science, Theory and Methods
 - Rank: 20/98 (Q1) in Engineering, Biomedical
- Journal Citation Indicator (JCR 2021): 1.63
 - Rank: 13/142 (D1) in Computer Science, Theory and Methods
 - Rank: 10/115 (D1) in Engineering, Biomedical

3.2 Main Contributions

- We propose the use of Super Gaussian (SG) sparse priors to represent the sharp image features in the stain concentrations of histological images.

- The model is evaluated with two representative members of the SG distributions, those corresponding to l_p and log energy functions.
- The proposed approach was successfully evaluated on five real histopathological image datasets and three different histological tasks: stain separation, color normalization and cancer classification. Time requirements and dependency to the reference prior were also evaluated. Our method improved stain separation, color normalization and cancer classification in comparison with state-of-the-art methods for BCD.

3.3 Related Conference Papers

3.3.1 Super Gaussian Priors for Blind Color Deconvolution of Histological Images

JCR Publication Details

Authors: Fernando Pérez-Bueno, Miguel Vega, Valery Naranjo, Rafael Molina, Aggelos K. Katsaggelos

Title: Super Gaussian Priors for Blind Color Deconvolution of Histological Images.

Reference: 2020 IEEE International Conference on Image Processing (ICIP), Abu Dhabi (United Arab Emirates), October 2020.

Status: Published

DOI: 10.1109/ICIP40778.2020.9191200

Quality indices:

- GGS Rating: A- (2021)
- GGS Class: 2 (2021)
- CORE: B (2020)

Abstract

Color deconvolution aims at separating multi-stained images into single stained ones. In digital histopathological images, true stain color vectors vary between images and need to be estimated to obtain stain concentrations and separate stain bands. These band images can be used for image analysis purposes and, once normalized, utilized with other multi-stained images (from different laboratories and obtained using different scanners) for classification purposes. In this paper we propose the use of Super Gaussian (SG) priors for each stain concentration together with the similarity to a given reference matrix for the color vectors. Variational inference and an evidence lower bound are utilized to automatically estimate all the latent variables. The proposed methodology is tested on real images and compared to classical and state-of-the-art methods for histopathological blind image color deconvolution.

3.3.2 Fully Automatic Blind Color Deconvolution of Histological Images Using Super Gaussians

JCR Publication Details

Authors: Fernando Pérez-Bueno, Miguel Vega, Valery Naranjo, Rafael Molina, Aggelos K. Katsaggelos

Title: Fully Automatic Blind Color Deconvolution of Histological Images Using Super Gaussians.

Reference: 28th European Signal Processing Conference, EUSIPCO 2020, Amsterdam (Netherland), January 2021.

Status: Published

DOI: 10.23919/Eusipco47968.2020.9287497

Quality indices:

- GGS Rating: B (2021)
- GGS Class: 3 (2021)
- CORE: B (2018)

3.3.3 Abstract

In digital pathology blind color deconvolution techniques separate multi-stained images into single stained bands. These band images are then used for image analysis and classification purposes. This paper proposes the use of Super Gaussian priors for each stain band together with the similarity to a given reference matrix for the color vectors. Variational inference and an evidence lower bound are then utilized to automatically estimate the latent variables and model parameters. The proposed methodology is tested on real images and compared to classical and state-of-the-art methods for histopathological blind image color deconvolution. Its use as a preprocessing step in prostate cancer classification is also analysed.

Blind Color Deconvolution, Normalization, and Classification of Histological Images using General Super Gaussian Priors and Bayesian Inference

Fernando Pérez-Bueno^{a,1,*}, Miguel Vega^b, María A. Sales^c, José Aneiros-Fernández^d, Valery Naranjo^e, Rafael Molina^a, Aggelos K. Katsaggelos^f

^a*Dpto. Ciencias de la Computación e Inteligencia Artificial, Universidad de Granada, Spain*

^b*Dpto. de Lenguajes y Sistemas Informáticos, Universidad de Granada, Spain*

^c*Anatomical Pathology Service, University Clinical Hospital of Valencia, Valencia, Spain*

^d*Intercenter Unit of Pathological Anatomy, San Cecilio University Hospital, Granada, Spain*

^e*Dpto. de Comunicaciones, Universidad Politécnica de Valencia, Spain*

^f*Dept. of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL, USA*

Abstract

Background and Objective:

Color variations in digital histopathology severely impact the performance of computer-aided diagnosis systems. They are due to differences in the staining process and acquisition system, among other reasons. Blind color deconvolution techniques separate multi-stained images into single stained bands which, once normalized, can be used to eliminate these negative color variations and improve the performance of machine learning tasks.

Methods:

In this work, we decompose the observed RGB image in its hematoxylin and eosin components. We apply Bayesian modeling and inference based on the use of Super Gaussian sparse priors for each stain together with prior closeness to a given reference color-vector matrix. The hematoxylin and eosin components are then used for image normalization and classification of histological images. The proposed framework is tested on stain separation, image normalization, and cancer classification problems. The results are measured using the peak signal to noise ratio, normalized median intensity and the area under ROC curve on five different databases.

Results:

The obtained results show the superiority of our approach to current state-of-the-art blind color deconvolution techniques. In particular, the fidelity to the tissue improves 1,27 dB in mean PSNR. The normalized median intensity shows a good normalization quality of the

*Corresponding author

Email addresses: fpb@ugr.es (Fernando Pérez-Bueno), mvega@ugr.es (Miguel Vega), sales_man@gva.es (María A. Sales), janeirosf@hotmail.com (José Aneiros-Fernández), vnaranjo@dcom.upv.es (Valery Naranjo), rms@decsai.ugr.es (Rafael Molina), aggk@eecs.northwestern.edu (Aggelos K. Katsaggelos)

proposed approach on the tested datasets. Finally, in cancer classification experiments the area under the ROC curve improves from 0.9491 to 0.9656 and from 0.9279 to 0.9541 on Camelyon-16 and Camelyon-17, respectively, when the original and processed images are used. Furthermore, these figures of merits are better than those obtained by the methods compared with.

Conclusions:

The proposed framework for blind color deconvolution, normalization and classification of images guarantees fidelity to the tissue structure and can be used both for normalization and classification. In addition, color deconvolution enables the use of the optical density space for classification, which improves the classification performance.

Keywords: Blind Color Deconvolution, Image normalization, histopathological images, Variational Bayes, Super Gaussian

1. Introduction

Histopathological tissues utilized for cancer diagnosis are stained using different dyes, commonly Hematoxylin-Eosin (H&E)[1]. This process facilitates the analysis made by pathologists. The Whole-Slide Images (WSIs) obtained by high-resolution scanners have many advantages: images do not deteriorate over time, they can be easily accessed and shared and, very importantly, enable pathologists to study slides on a screen and the development of Computer-Aided-Diagnosis (CAD) systems. The performance of CAD systems can be significantly affected by color variations of histological images[2]. These variations, which can be inter- and intra- laboratory are introduced in the acquisition procedure. Caused by variables like fixatives, staining manufactures, lab condition and temperatures, and the use of different scanners, among others, see [3] for details. Two main approaches have been proposed to minimize the influence of color variations on the obtained images and their posterior analysis. Blind Color Deconvolution (BCD) and Color Normalization (CN).

BCD techniques separate the stains in an image by estimating its stain color-vectors and the corresponding stain concentrations. The process should lead to structure, nuclei (hematoxylin), cytoplasm and collagen of the stroma (eosin), etc, preservation. BCD can be used for image normalization (by normalizing each stain separately), but this is only one of the possible solutions it offers to deal with color variation. Stain separation also allows CAD systems to use the information provided by each stain separately [4]. Furthermore, concentrations can be directly used for classification [4, 5].

CN focuses on transforming histological images to a common color range, usually obtained from a reference WSI. Tosta *et al.* [3] classifies direct CN methods into histogram matching and color transfer. Histogram matching techniques adjust image colors using histogram information. This is a common solution for general images but it is not appropriate for histological images as it assumes that stains are equally distributed and disregards local information. Stain concentration is closely related to the tissue and cell structures which need to be preserved. Color transfer often includes a segmentation step to identify histological regions or dyes. Then a stain-specific based color correction is applied. However, the selective transformation occasionally causes artifacts on the images. Most Deep Learning (DL) methods are included in this category as they usually perform CN without Color Deconvolution

(CD) [6, 7].

1.1. Related work

A wide range of solutions have been proposed to find the stain color-vector in the images. They can be experimentally obtained as Ruifrok et al. [8] did in one of the pioneer works in the CD field. The empirically obtained color-vectors proposed in [8] do not tackle stain color variation. To take variability into account, the selection of pixels corresponding to each stain was proposed in [9]. The amount of slides available quickly made this solution obsolete. Formulating the problem as blind source separation, Non-negative Matrix Factorization (NMF) was used in [10]. Using the same principles [11] and [12] further developed this research approach including regularization and sparsity terms which encapsulate the assumption that each stain fixes only to specific tissue structures, forcing most of the pixels to respond to one type of stain only. Singular Value Decomposition (SVD), was applied in [13] for H&E stain separation and then further developed in [14] by considering the interaction between stains. It was recently revisited in [15] where the steps were reorganized to obtain a time-optimized pipeline. The NMF memory and time requirements were reduced in [16] with the use of Non-Negative Least Squares (NNLS). In [17], stain vectors were estimated through clustering in the Maxwellian chromacity plane. In [18], supervised relevant vector machines are used to segment background, hematoxylin and eosin pixels. The color-vector for each stain is then defined as the mean of the pixels in each class. Recently, Salvi *et al.*[19] have presented a three steps method using Gabor kernels, structure segmentation and a final deconvolution step. Independent Component Analysis (ICA) was utilized in [20] and extended in [21, 22], using the wavelet transform that reduces the independence condition between sources. The method in [13] was revised in [23], where the author state that they obtained better result applying it in the linearly inverted RGB-space instead of the (logarithmically inverted) absorbency space. The work by Zheng *et al.*[24] includes the deconvolution by Ruifrok as a starting point and optimizes the color-vector and concentration values using a prior knowledge-based objective function.

In this work, we develop a Bayesian framework for BCD, CN, and classification of histological images using both normalized and stain separated images. Like the approaches presented in [25] and [5], this work uses Bayesian modeling and inference. In [25], a similarity prior to reference stain color-vectors, together with a smoothness Simultaneous Autoregressive (SAR) prior model on the stain concentrations were used. Since the SAR prior oversmooths edges, in [5], we presented the use of a Total Variation (TV) prior on the stain concentrations. The TV prior preserves sharp edges while reducing noise in the images[26], but unfortunately, in some cases, it tends to flatten areas which, together with the edges, are essential for image classification. For blind natural image deconvolution, we proposed in [26, 27] a general framework to model and restore the the image from its blurred and noisy version. We introduced a large class of sparse image priors, the so called Super Gaussians (SGs) which represent well sharp image characteristics. Most sparse image models used in the literature are included in the formulation as special cases. In this work we provide a complete mathematical derivation of how to combine SG prior models with the likelihood associated to blind color deconvolution of histological images. The proposed approach is tested on stain separation, image normalization, and classification problems using five different databases. Preliminary results were presented in [28, 29] where a limited theoretical derivation was provided and a reduced

set of SG priors and datasets were utilized in the experimental validation. In this work we extend [28, 29] by providing a complete and clearer mathematical derivation of the model. We also provide an extensive experimental validation using three additional databases including images from different laboratories. The validation now includes: Application of the SG prior models to stain normalization, a complete evaluation of the stain normalization results, additional classification experiments using normalized images and stain concentrations separately, time comparison of the competing methods, and analysis of the similarity prior on the color-vectors. Furthermore we also evaluate the use of normalized images or stain concentrations for classification tasks, and discuss the use of a third residual stain.

The paper is organized as follows: Section 2 introduces the BCD problem and its mathematical formulation. Section 2.2 presents the modeling and Bayesian inference proposed for the estimation of the color-vector matrix, the stain concentrations, and all the model parameters. In section 4, we use H&E stained images to evaluate the proposed framework and provide a comparison with classical and state-of-the-art CD methods using four different histopathology related tasks: BCD stain separation, image normalization, deconvolution based prostate cancer classification, and breast cancer classification using normalized images and stain concentrations. Section 5 includes the discussion and finally, section 6 concludes the paper.

2. Methods

2.1. Problem Formulation

For each WSI, the tissue observed by a brightfield microscope is represented as an $MN \times 3$ matrix \mathbf{I} . Each color plane is stacked into a $MN \times 1$ column vector $\mathbf{i}_c = (i_{1c}, \dots, i_{MNc})^T$, $c \in \{R, G, B\}$. The transmitted light on the color band $c \in \{R, G, B\}$ for the i -th pixel in the slide is stored in i_{ic} . Stain deconvolution methods usually apply the Beer-Lambert law to transform slide images to the *Optical Density* (OD) space, where the n_s stained slide can be expressed as

$$\mathbf{Y}^T = \mathbf{M}\mathbf{C}^T + \mathbf{N}^T, \quad (1)$$

The observed OD image $\mathbf{Y} \in \mathbb{R}^{MN \times 3}$ contains three channels, i.e., $\mathbf{Y} = [\mathbf{y}_R \ \mathbf{y}_G \ \mathbf{y}_B]$ and each channel $\mathbf{y}_c \in \mathbb{R}^{MN \times 1}$ is defined as $\mathbf{y}_c = -\log_{10}(\mathbf{i}_c/\mathbf{i}_c^0)$, with \mathbf{i}_c^0 the incident light (Typically 255 for RGB images). The values for \mathbf{y}_c are computed element-wise. The matrix $\mathbf{C} \in \mathbb{R}^{MN \times n_s}$ contains the stain concentration, $\mathbf{M} \in \mathbb{R}^{3 \times n_s}$ is the color-vector matrix and $\mathbf{N} \in \mathbb{R}^{MN \times 3}$ is a random noise matrix with i.i.d. zero mean Gaussian components with variance β^{-1} .

The BCD approach aims to estimate both \mathbf{C} and \mathbf{M} . In \mathbf{C} , the concentration of each stain in the i -th \mathbf{Y} pixel value, $\mathbf{y}_{i,:}$, is expressed as the i -th row $\mathbf{c}_{i,:}^T = (c_{i1}, \dots, c_{in_s})$ and the whole contribution of the s -th stain to the image is the s -th column $\mathbf{c}_s = (c_{1s}, \dots, c_{MN_s})^T$. In the color-vector matrix \mathbf{M} , each \mathbf{m}_s column contains the color composition of the s -th stain.

2.2. SG Bayesian Model

Using the Beer-Lambert model in (1), the observation model is

$$p(\mathbf{Y}|\mathbf{C}, \mathbf{M}, \beta) = \prod_{i=1}^{MN} \mathcal{N}(\mathbf{y}_{i,:} | \mathbf{M}\mathbf{c}_{i,:}, \beta^{-1}\mathbf{I}_{3 \times 3}). \quad (2)$$

Table 1: Some penalty functions

Label	$\rho(\mathbf{s})$	$\rho'(\mathbf{s})/ \mathbf{s} $
$\ell_p, 0 < p \leq 1$	$\frac{1}{p} \mathbf{s} ^p$	$ \mathbf{s} ^{p-2}$
log	$\log(\epsilon + \mathbf{s})$	$(\epsilon + \mathbf{s})^{-1} \mathbf{s} ^{-1}$

The Bayesian approach requires to select a prior distribution on the unknowns. Here we adopt SG distributions as priors for the stain concentrations in the filtered space. SG priors are known to preserve sharp images [26]. They induce sparsity and allow us to find the key values for each stain. We use a set of J high-pass filters noted as $\{\mathbf{D}_\nu\}_{\nu=1}^J$ to obtain the filtered concentrations $\mathbf{c}_{\nu_s} = \mathbf{D}_\nu \mathbf{c}_s$. The filtered space remarks the edges in the image that we want to preserve.

$$\begin{aligned} p(\mathbf{C}|\boldsymbol{\alpha}) &= \prod_{\nu=1}^J \prod_{s=1}^{n_s} p(\mathbf{c}_{\nu_s}|\alpha_{\nu_s}) \\ &= \prod_{\nu=1}^J \prod_{s=1}^{n_s} \prod_{i=1}^{MN} Z(\alpha_{\nu_s}) \exp[-\alpha_{\nu_s} \rho(c_{\nu_s}(i))], \end{aligned} \quad (3)$$

with $\alpha_{\nu_s} > 0$ and $Z(\alpha_{\nu_s})$ a partition function. For $p(\mathbf{c}_{\nu_s}|\alpha_{\nu_s})$ in (3) to be SG, the penalty function $\rho(\cdot)$ has to be symmetric around zero. In addition, $\rho(\sqrt{s})$ has to be increasing and concave for $s \in (0, \infty)$, which is equivalent to $\rho'(s)/s$ being decreasing on $(0, \infty)$. The latter condition allows ρ to be written as follows

$$\rho(c_{\nu_s}(i)) = \inf_{\eta_{\nu_s}(i) > 0} L(c_{\nu_s}(i), \eta_{\nu_s}(i)) \quad (4)$$

where $L(c_{\nu_s}(i), \eta_{\nu_s}(i)) = \frac{1}{2} \eta_{\nu_s}(i) c_{\nu_s}^2(i) - \rho^*\left(\frac{1}{2} \eta_{\nu_s}(i)\right)$, \inf denotes infimum and $\rho^*(\cdot)$ is the concave conjugate of $\rho(\cdot)$ and $\boldsymbol{\eta}_{\nu_s} = \{\eta_{\nu_s}(i)\}_{i=1}^{MN}$ are positive parameters. The relationship dual to (4) is given by [30]

$$\rho^*\left(\frac{1}{2} \eta_{\nu_s}(i)\right) = \inf_{c_{\nu_s}(i)} \frac{1}{2} \eta_{\nu_s}(i) c_{\nu_s}^2(i) - \rho(c_{\nu_s}(i)). \quad (5)$$

Table 1 and figure 1 show possible choices for the penalty function and their corresponding SG distributions (for additional SG distributions, see [26]).

The color-vector matrix $\mathbf{M} = [\mathbf{m}_1, \dots, \mathbf{m}_{n_s}]$ is also unknown, but it is expected to be similar to a reference color-vector matrix $\underline{\mathbf{M}} = [\underline{\mathbf{m}}_1, \dots, \underline{\mathbf{m}}_{n_s}]$. Therefore we use a similar prior as

$$\begin{aligned} p(\mathbf{M}|\boldsymbol{\gamma}) &= \prod_{s=1}^{n_s} p(\mathbf{m}_s|\gamma_s) \\ &\propto \prod_{s=1}^{n_s} \gamma_s^{\frac{3}{2}} \exp\left(-\frac{1}{2} \gamma_s \|\mathbf{m}_s - \underline{\mathbf{m}}_s\|^2\right), \end{aligned} \quad (6)$$

where the parameter γ_s , $s = 1, \dots, n_s$, measures the confidence on the accuracy of the reference $\underline{\mathbf{m}}_s$.

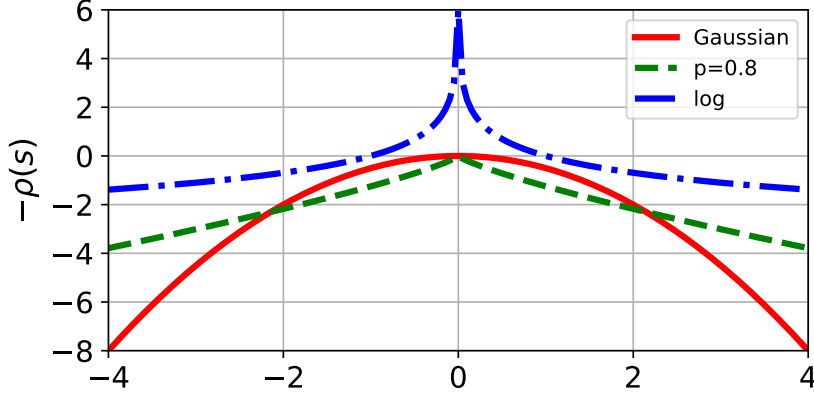


Figure 1: Penalties corresponding to functions in Table 1. $\log |s|$ is bounded for better visualization.

The joint probability distribution is then defined as

$$\begin{aligned}
 p(\mathbf{Y}, \mathbf{C}, \mathbf{M}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) &= p(\mathbf{M}|\boldsymbol{\gamma})p(\boldsymbol{\gamma})p(\mathbf{Y}|\mathbf{C}, \mathbf{M}, \boldsymbol{\beta})p(\boldsymbol{\beta}) \\
 &\quad \times \prod_{\nu=1}^J \prod_{s=1}^{n_s} p(\mathbf{c}_{\nu s}|\alpha_{\nu s})p(\alpha_{\nu s}), \tag{7}
 \end{aligned}$$

where we include the hyperpriors $p(\boldsymbol{\gamma})$, $p(\boldsymbol{\beta})$ and $p(\alpha_{\nu s})$ on the model hyperparameters for automatic estimation.

Following the Bayesian paradigm, the estimation of \mathbf{M} and \mathbf{C} is based on our estimation of the posterior distribution $p(\Theta|\mathbf{Y})$ with $\Theta = \{\mathbf{C}, \mathbf{M}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}\}$ including all the unknowns. Our approach approximates $p(\Theta|\mathbf{Y})$ using the mean-field variational Bayesian model [31], by the distribution $q(\Theta)$ of the form $q(\Theta) = \prod_{s=1}^{n_s} q(\mathbf{m}_s) \prod_{\nu=1}^J q(\mathbf{c}_{\nu s})$ that minimizes the Kullback-Leibler (KL) divergence [32] defined as

$$\begin{aligned}
 \mathbf{KL}(q(\Theta) || p(\Theta|\mathbf{Y})) &= \int q(\Theta) \log \frac{q(\Theta)}{p(\Theta, \mathbf{Y})} d\Theta \\
 &\quad + \log p(\mathbf{Y}). \tag{8}
 \end{aligned}$$

However, the SG prior for \mathbf{C}_ν makes the evaluation of this divergence intractable. To tackle this problem we will make use of the quadratic bound for ρ to bound the prior in (3) with a Gaussian form

$$p(c_{\nu s}(i)|\alpha_{\nu s}) \geq Z(\alpha_{\nu s}) \exp[-\alpha_{\nu s}L(c_{\nu s}(i), \eta_{\nu s}(i))], \tag{9}$$

$\forall \eta_{\nu s}(i) > 0$. Then we define

$$\mathcal{M}_\nu(\mathbf{C}, \boldsymbol{\eta}_\nu | \boldsymbol{\alpha}_\nu) = \prod_{s=1}^{n_s} \prod_{i=1}^{MN} Z(\alpha_{\nu s}) \exp[-\alpha_{\nu s}L(c_{\nu s}(i), \eta_{\nu s}(i))] \tag{10}$$

and

$$\begin{aligned}
F(\Theta, \mathbf{Y}) &= p(\mathbf{M}, \boldsymbol{\gamma})p(\boldsymbol{\gamma})p(\mathbf{Y}|\mathbf{C}, \mathbf{M}, \beta)p(\beta) \\
&\quad \times \prod_{\nu} \mathcal{M}_{\nu}(\mathbf{C}, \boldsymbol{\eta}_{\nu}|\boldsymbol{\alpha}_{\nu})p(\boldsymbol{\alpha}_{\nu}),
\end{aligned} \tag{11}$$

obtaining the bound $\log p(\Theta, \mathbf{Y}) \geq \log F(\Theta, \mathbf{Y})$.

Using $F(\Theta, \mathbf{Y})$ for the posterior distribution in (8) we can now minimize $\mathbf{KL}(q(\Theta) \| F(\Theta, \mathbf{Y}))$ instead of $\mathbf{KL}(q(\Theta) \| p(\Theta|\mathbf{Y}))$.

As described in [31], $q(\theta)$, for each unknown $\theta \in \Theta$, can be written as

$$q(\theta) \propto \exp \langle \log F(\Theta, \mathbf{Y}) \rangle_{q(\Theta \setminus \theta)}, \tag{12}$$

where $\langle \cdot \rangle$ is the expectation and $q(\Theta \setminus \theta)$ indicates that it is taken with respect to all parameters in Θ except θ . The mean is used when a point estimation is required.

2.3. Updating the Concentrations

We define

$$\begin{aligned}
\mathbf{e}_{i,:}^{-s} &= \mathbf{y}_{i,:} - \sum_{k \neq s} \langle c_{ik} \rangle \langle \mathbf{m}_k \rangle \\
z_i^{-s} &= \langle \mathbf{m}_s \rangle^T \mathbf{e}_{i,:}^{-s}, \quad i = 1, \dots, MN,
\end{aligned} \tag{13}$$

from eq. (12) we can obtain that $q(\mathbf{c}_s) = \mathcal{N}(\mathbf{c}_s | \langle \mathbf{c}_s \rangle, \boldsymbol{\Sigma}_{\mathbf{c}_s})$, where the inverse of the covariance matrix is given by

$$\boldsymbol{\Sigma}_{\mathbf{c}_s}^{-1} = \beta \langle \|\mathbf{m}_s\|^2 \rangle \mathbf{I}_{MN \times MN} + \sum_{\nu} \alpha_{\nu s} \mathbf{D}_{\nu}^T \text{diag}(\boldsymbol{\eta}_{\nu s}) \mathbf{D}_{\nu} \tag{14}$$

and the mean is obtained as

$$\boldsymbol{\Sigma}_{\mathbf{c}_s}^{-1} \langle \mathbf{c}_s \rangle = \beta \mathbf{z}^{-s}. \tag{15}$$

2.4. Updating the Color-Vector matrix

Similarly, from (13), we obtain that $q(\mathbf{m}_s) = \mathcal{N}(\mathbf{m}_s | \langle \mathbf{m}_s \rangle, \boldsymbol{\Sigma}_{\mathbf{m}_s})$, where

$$\begin{aligned}
\boldsymbol{\Sigma}_{\mathbf{m}_s}^{-1} &= \left(\sum_{\nu=1}^J \beta_{\nu} \sum_{i=1}^{MN} \langle c_{\nu is}^2 \rangle + \gamma_s \right) \mathbf{I}_{3 \times 3}, \\
\boldsymbol{\Sigma}_{\mathbf{m}_s}^{-1} \langle \mathbf{m}_s \rangle &= \left(\sum_{\nu=1}^J \beta_{\nu} \sum_{i=1}^{MN} \langle c_{\nu is} \rangle \mathbf{e}_{\nu i,:}^{-s} + \gamma_s \mathbf{m}_s \right).
\end{aligned} \tag{16}$$

To ensure $\langle \mathbf{m}_s \rangle$ to be a unitary vector, we replace $\langle \mathbf{m}_s \rangle$ by $\langle \mathbf{m}_s \rangle / \|\langle \mathbf{m}_s \rangle\|$ and $\boldsymbol{\Sigma}_{\mathbf{m}_s}$ by $\boldsymbol{\Sigma}_{\mathbf{m}_s} / \|\langle \mathbf{m}_s \rangle\|^2$.

2.5. Updating the Variational Parameter

The estimation of the $\boldsymbol{\eta}$ matrix, requires to solve, for each $s \in \{1, \dots, n_s\}$, $\nu \in \{1, \dots, J\}$ and $i \in \{1, \dots, MN\}$

$$\begin{aligned}\hat{\eta}_{\nu s}(i) &= \arg \min_{\eta_{\nu s}(i)} \langle L(c_{\nu s}(i), \eta_{\nu s}(i)) \rangle_{\mathbf{q}(\mathbf{c}_s)} \\ &= \arg \min_{\eta_{\nu s}(i)} \frac{1}{2} \eta_{\nu s}(i) u_{\nu s}^2(i) - \rho^* \left(\frac{1}{2} \eta_{\nu s}(i) \right)\end{aligned}\quad (17)$$

where $u_{\nu s}(i) = \sqrt{\langle c_{\nu s}^2(i) \rangle}$. Since

$$\rho^* \left(\frac{\hat{\eta}_{\nu s}(i)}{2} \right) = \min_x \frac{1}{2} \hat{\eta}_{\nu s}(i) x^2 - \rho(x) \quad (18)$$

whose minimum is achieved at $x = u_{\nu s}(i)$. Then, differentiating the right hand side of (18) with respect to x , equating it to zero and substituting the value for x at its minimum, we have,

$$\hat{\eta}_{\nu s}(i) = \rho'(u_{\nu s}(i)) / |u_{\nu s}(i)|. \quad (19)$$

2.6. Updating the Hyperparameters

The estimates of the parameters controlling the noise and color-vectors confidence are calculated from

$$\hat{\beta}^{-1} = \frac{\text{tr} \langle (\mathbf{Y}^T - \mathbf{M}\mathbf{C}^T)(\mathbf{Y}^T - \mathbf{M}\mathbf{C}^T)^T \rangle_{\mathbf{q}(\boldsymbol{\Theta})}}{3MN}, \quad (20)$$

$$\hat{\gamma}_s^{-1} = \frac{\text{tr} \langle (\mathbf{m}_s - \underline{\mathbf{m}}_s)(\mathbf{m}_s - \underline{\mathbf{m}}_s)^T \rangle}{3}. \quad (21)$$

Using (12) the distribution for $\alpha_{\nu s}$ is written as follows

$$\mathbf{q}(\alpha_{\nu s}) = \text{const} + \sum_{i=1}^{MN} \log Z(\alpha_{\nu s}) \exp[-\alpha_{\nu s} \rho(u_{\nu s}(i))], \quad (22)$$

where $u_{\nu s}(i)$ was defined in section 2.5. Estimating $\alpha_{\nu s}$ with the mode of (22), we obtain $\hat{\alpha}_{\nu s}$ from

$$\frac{\partial \log Z(\hat{\alpha}_{\nu s})}{\partial \hat{\alpha}_{\nu s}} = \frac{1}{MN} \sum_{i=1}^{MN} \rho(u_{\nu s}(i)). \quad (23)$$

From the penalty functions shown in Table 1, ℓ_p produces proper priors, where we can evaluate the partition function. However, the log penalty function produces an improper prior. To tackle this problem we examine the behaviour of

$$Z(\alpha_{\nu s}, K)^{-1} = \int_{-K}^K \exp[-\alpha_{\nu s} \rho(t)] dt \quad (24)$$

when $\alpha_{\nu s} \neq 1$, and keeping in $\partial Z(\alpha_{\nu s}) / \partial \alpha_{\nu s}$ the term that depends on $\alpha_{\nu s}$. This produces for the log prior

$$\frac{\partial Z(\hat{\alpha}_{\nu s})}{\partial \hat{\alpha}_{\nu s}} = (\hat{\alpha}_{\nu s} - 1)^{-1}. \quad (25)$$

Values for $\hat{\alpha}_{\nu s}$ can be obtained substituting this last expression into (23). Flat hyperpriors have been used for all the hyperparameters.

Algorithm 1 Fully Variational Bayesian SG BCD

Require: Observed RGB image \mathbf{I} and reference (prior) color-vector matrix $\underline{\mathbf{M}}$.

Obtain the OD image \mathbf{Y} from \mathbf{I} and set $\langle \mathbf{m}_s \rangle^{(0)} = \mathbf{m}_s$, $\Sigma_{\mathbf{m}_s}^{(0)} = \mathbf{0}$, $\Sigma_{\mathbf{c}_s}^{(0)} = \mathbf{0}$, $\langle \mathbf{c}_s \rangle^{(0)}$, $\forall s = 1, \dots, n_s$, from the matrix \mathbf{C} obtained as $\mathbf{C}^T = \underline{\mathbf{M}}^+ \mathbf{Y}^T$, with $\underline{\mathbf{M}}^+$ the Moore-Penrose pseudo-inverse of $\underline{\mathbf{M}}$, and $n = 0$. **while** convergence criterion is not met **do**

1. Set $n = n + 1$.
2. Obtain $\beta^{(n)}$, $\gamma_s^{(n)}$ and $\alpha_{\nu_s}^{(n)}$ from (20), (21) and (23).
3. Using $\langle \mathbf{c}_s \rangle^{(n-1)}$ and $\Sigma_{\mathbf{c}_s}^{(n-1)}$ $\forall s$, update variational parameters $\hat{\boldsymbol{\eta}}_{\nu_s}^{(n)}$ from (19) $\forall \nu$.
4. Using $\langle \mathbf{c}_s \rangle^{(n-1)}$, $\Sigma_{\mathbf{c}_s}^{(n-1)}$ and $\langle \mathbf{m}_s \rangle^{(n-1)}$ update $\Sigma_{\mathbf{m}_s}^{(n-1)}$ and solve (16) for the color-vectors $\langle \mathbf{m}_s \rangle^{(n)}$, $\forall s$.
5. Using $\langle \mathbf{m}_s \rangle^{(n)}$, $\Sigma_{\mathbf{m}_s}^{(n)}$ and $\hat{\boldsymbol{\eta}}_{\nu_s}^{(n)}$ $\forall \nu$ update $\Sigma_{\mathbf{c}_s}^{(n-1)}$ from (14) and solve (15) for the concentrations $\langle \mathbf{c}_s \rangle^{(n)}$, $\forall s$.

end while

Output color-vector $\hat{\mathbf{m}}_s = \langle \mathbf{m}_s \rangle^{(n)}$ and $\hat{\mathbf{c}}_s = \langle \mathbf{c}_s \rangle^{(n)}$.

2.7. Covariance matrices for the concentration:

We have to find the covariance matrix $\Sigma_{\mathbf{c}_s}$ in order to calculate its trace as well as $\hat{\boldsymbol{\eta}}_{\nu_s}(i)$. Unfortunately, this is computationally intensive. To reduce the impact of the calculation, we propose to approximate $\Sigma_{\mathbf{c}_s}$ as follows. First, we approximate $\text{diag}(\boldsymbol{\eta}_{\nu_s})$ by

$$\text{diag}(\boldsymbol{\eta}_{\nu_s}) \approx z(\boldsymbol{\eta}_{\nu_s})\mathbf{I}, \quad (26)$$

where we use the mean of the diagonal values to calculate $z(\boldsymbol{\eta}_{\nu_s})$. Then we approximate

$$\Sigma_{\mathbf{c}_s}^{-1} \approx \beta \langle \|\mathbf{m}_s\|^2 \rangle \mathbf{I}_{MN \times MN} + \sum_{\nu} \alpha_{\nu_s} z(\boldsymbol{\eta}_{\nu_s}) \mathbf{D}_{\nu}^T \mathbf{D}_{\nu} = \mathbf{B}.$$

Finally we have $\langle c_{\nu_s}(i) \rangle \approx (\langle c_{\nu_s}(i) \rangle)^2 + \frac{1}{MN} \text{tr} [\mathbf{B}^{-1} \mathbf{D}_{\nu}^T \mathbf{D}_{\nu}]$.

2.8. Proposed Algorithm

Considering the previous inference, we propose the Fully Variational Bayesian SG BCD in Algorithm 1. Figure 2 depicts the pipeline followed by the proposed framework. We use the Conjugate Gradient approach to solve the linear equation problem in step 4 of Alg. 1. The inference procedure iterates between concentration update, color-vector update, variational parameter update, and parameter update. When necessary, a single-stain RGB image $\hat{\mathbf{I}}_s^{\text{sep}}$, can be obtained from the outputs in Alg. 1 as follows

$$(\hat{\mathbf{I}}_s^{\text{sep}})^T = \exp_{10}(-\hat{\mathbf{m}}_s \hat{\mathbf{c}}_s^T) \quad (27)$$

2.9. Use of the algorithm on WSIs

The size of WSI images is usually on the order of Gigapixels, making their processing challenging. The proposed method could, in principle, be used directly on WSIs but Bayesian methods are computationally expensive and the computational burden would be considerable, notice that M and N would be huge. However, WSIs are not usually processed at once and most classification or analysis tasks require patching [4, 33] or focusing only on Regions-of-Interest (RoI) [22].

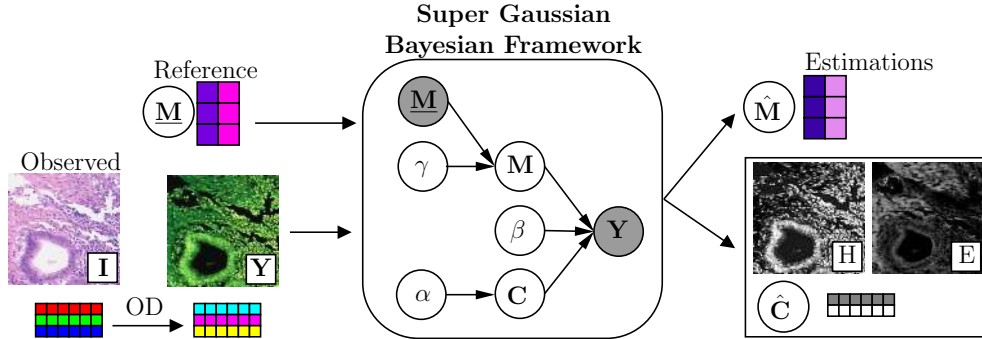


Figure 2: Pipeline of the proposed framework. First, the H&E image is converted to the OD space. The OD image \mathbf{Y} and the reference matrix \mathbf{M} are given to the SG Bayesian framework. The values of all parameters are automatically estimated during the inference procedure using the KL divergence. Finally, the estimated color-vector matrix $\hat{\mathbf{M}}$ and concentrations $\hat{\mathbf{C}}$ are obtained.

For classification purposes it is possible to deconvolve patches separately. This approach can tackle local variations but will create variations in the estimated color-vector matrix for each patch. Another possible solution is to select a RoI in order to obtain the color matrix. This is the approach we follow in this paper. First, we select the biggest connected RoI within the patches of interest and estimate the color-vector matrix $\hat{\mathbf{M}}$ for the complete WSI. Then, the concentrations of the remaining patches are obtained using $\mathbf{C}^T = \hat{\mathbf{M}}^+ \mathbf{Y}^T$, with $\hat{\mathbf{M}}^+$ the Moore-Penrose pseudo-inverse of $\hat{\mathbf{M}}$. Notice that a single color-vector matrix is obtained for all patches belonging to the same WSI and that they can be stitched together without artifacts if necessary.

An alternative approach is to use the prior on the concentrations in eq. (3) and the observation model in eq. (2) for all the patches in the WSI we want to use. In other words, eq. (2) becomes a product over patches of interest. This is the approach we follow in the paper. Notice that the new variational distributions are similar to those derived in the paper but now have to consider all the utilized patches.

3. Data Material

Five databases, were used in the experiments *Warwick Stain Separation Benchmark* (WSSB) [22], SICAPv1 [4], SICAP-GR, Camelyon-16 [34] and Camelyon-17 [35]. Details for each database are provided below:

3.0.1. WSSB

WSSB is a multi-tissue dataset (breast, colon, and lung) that contains 24 H&E stained images from different laboratories and captured with different microscopes. Colon images were captured at 20x magnification and Breast and Lung at 40x. Hematoxylin- and Eosin-only pixels manually selected by expert pathologists were used to obtain the ground truth stain color-vector matrix for each image. Then, the ground truth concentration is calculated in [22] as

$$\mathbf{C}_{GT}^T = \mathbf{M}_{GT}^+ \mathbf{Y}^T. \quad (28)$$

Table 2: Camelyon 17 dataset labeling structure

Subset	WSI total	negative	stage label		
			itc	micro	macro
Whole training set	500	318	36	59	87
annotated	50	0	16	17	17
no annotated	450	318	20	42	70

Then using (27), a single-stain RGB image was calculated for both hematoxylin and eosin. This database will be used for BCD evaluation.

3.0.2. SICAPv1

This database comes from *Hospital Clínico Universitario de Valencia*, Spain, it contains 79 H&E WSI from 48 patients, 19 benign prostate tissue biopsies (negative class) and 60 pathological prostate tissue biopsies (positive class). The images were digitized using a Ventana iScan Coreo scanner at 40x magnification. Malignant regions of each pathological WSI were annotated by expert pathologists. 60 WSI (17 benign and 43 pathological) were used as training set and the remaining 19 WSI (2 benign and 17 pathological) were utilized for testing. This database will be used for classification purposes and some of its slides will also be used for CN as we describe next.

3.0.3. SICAP-HUVNGR

This dataset contains 26 prostate H&E WSI: 13 slides at 40x magnification from *Hospital Universitario Virgen de las Nieves de Granada (HUVNGR)* and 13 slides from *Hospital Clínico Universitario de Valencia* (randomly extracted from SICAPv1 dataset). These WSIs will be used for CN evaluation.

3.0.4. Camelyon-16 and 17

These two databases are part of the Camelyon challenge¹ for cancer metastasis detection in the lymph node. We will use them in CN and classification experiments. Both Camelyon databases were scanned at 40x. They are described below.

- Camelyon-16 contains 400 H&E-stained lymph node multiresolution WSIs from 2 different laboratories. 270 are used from training (159 referred as normal and 111 as tumor) and 130 for testing. Cancer regions were annotated by expert pathologists in tumor and test images. All the annotations are available.
- Camelyon-17 contains 1000 WSIs from 5 medical centers. Only the training set, which contains 500 WSIs, was used since the annotations for the testing WSIs are not yet available. The dataset comprises 20 patients per center and 5 slides per patient. Cancer regions were annotated by pathologists only on 50 WSIs, but the stage label: negative, isolated tumor cells (itc), micrometastasis (micro), macrometastasis (macro) is available for all the slides in the training set. See Table 2 for details.

For a clearer perspective, we include Table 3 that shows the experiments performed for each database.

¹<https://camelyon17.grand-challenge.org/>.

Table 3: Experiments performed for each database

database	Stain separation	Color normalization	Classification
WSSB	✓		
SICAPv1			✓
HUVNGR		✓	
Camelyon-16		✓	✓
Camelyon-17		✓	✓

4. Experiments and Results

As mentioned previously, BCD techniques are used to facilitate the visual analysis and to improve the automatic classification of WSIs. These are frequently conflicting goals due to the differences between the human eye and computer vision. Usually, the highest classification performance is not obtained with the most accurate color deconvolved images, where each stain is accurately separated.

We have designed a set of experiments to test the performance of the algorithms on the most common histological color deconvolution related tasks: stain separation, image normalization, and classification. Our first experiment is devoted to assess the quality of the stain separated images, that is, of the concentration matrices and color vector. Then, in the second one, we analyze the quality of these matrices in CN. In the CN step a reference WSI is selected and the color-vectors of the image to normalize are substituted by those of the reference image, keeping the concentrations. Finally, the obtained deconvolved and normalized images are evaluated on histological classification problems.

The proposed SG framework was compared with the following (B)CD methods frequently used in the literature: the classical non-blind CD method by Ruifrok *et al.* [8] and the BCD methods by Macenko *et al.* [13], Vahadane *et al.* [11], Alsubaie *et al.* [22], Hidalgo-Gavira *et al.* [25], Pérez-Bueno *et al.* [5] and Zheng *et al.* [24]. They will be denoted by RUI, MAC, VAH, ALS, HID, PER, and ZHE, respectively. Since SG represents a family of prior distributions, we have selected two of its representative members, the corresponding to ℓ_p and log energy functions. They will be denoted by L1 and LOG, respectively, in the experiments.² For the ℓ_p function we experimentally compared values for p in the interval $[0.6, 1]$ and found no significant differences. For simplicity, we choose ℓ_1 . The proposed L1 and LOG methods were run until the criterion $\| \langle \mathbf{c}_s \rangle^{(n)} - \langle \mathbf{c}_s \rangle^{(n-1)} \|^2 / \| \langle \mathbf{c}_s \rangle^{(n)} \|^2 < 10^{-3}$ was met by all the stains. Vertical, horizontal and diagonal differences were used as high-pass filters in the concentration prior (eq. (3)). All the model parameters are automatically estimated.

4.1. BCD Stain Separation Experiments

We begin the experimental assessment by comparing the fidelity to the H&E separation obtained by the different BCD methods on the WSSB database, see Section 3. From this dataset, we show an observed RGB image (Fig. 3(a)) and the corresponding ground truth H&E-only RGB image (Fig. 3(b)).

To set an adequate prior for our method, we consider that the stain color properties may change for the different tissues types in WSSB (Colon, breast, lung). For each tissue, an

²The code used in the experiments will be made available at <https://github.com/vipgugr> upon acceptance of the paper.

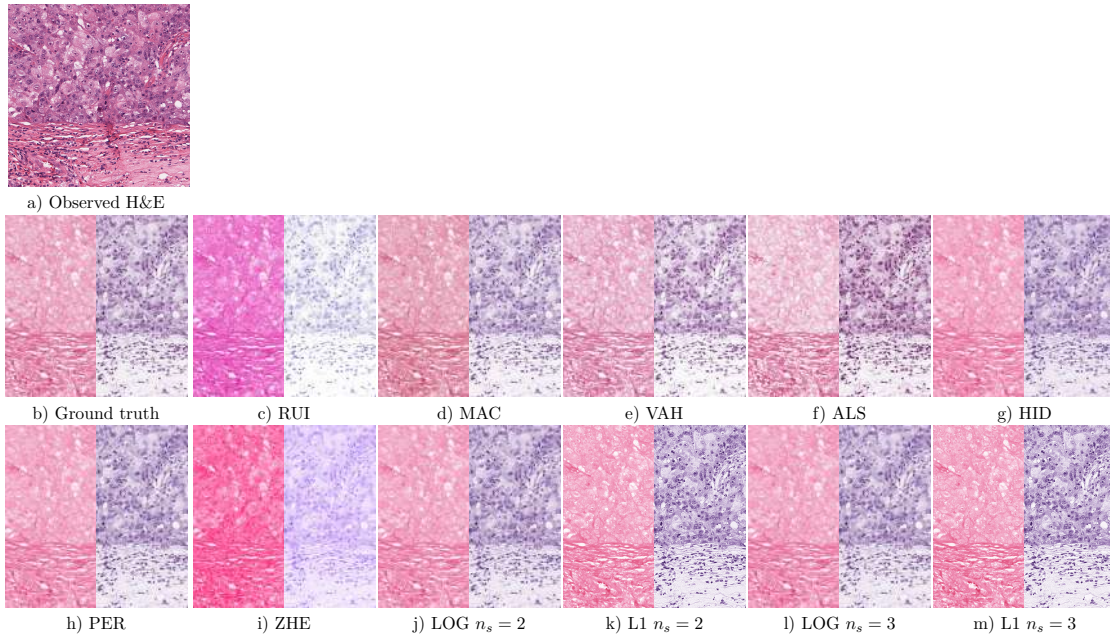


Figure 3: Single breast observed H&E RGB image from WSSB [22], corresponding ground truth single stain E-only and H-only images and separation obtained by the BCD methods. Eosin and hematoxylin separations are presented on the left and right hand sides of each image, respectively.

H&E reference color-vector matrix \mathbf{M} was selected by a non-medical expert using a single pixel for each stain. Following the widely used implementation [36] of Ruifrok’s method, when a third residual component is used, the reference color-vector is calculated using the vector product of the H&E components in the color matrix.

The single stain images obtained from the observed image in Fig. 3(a) are shown in Fig. 3(c-k). The standard color vector used by RUI obtains a separation that do not represent the ground truth. The proposed methods, L1 and LOG, and MAC, HID, and PER are able to find colors that are close to the ground truth separation in Fig. 3(b). The Bayesian methods HID, PER and the proposed ones share the same prior for the color-vectors, but their differences lay on the concentration prior. HID uses a SAR model, that tends to oversmooth images. The TV based method PER keeps edges sharp, but flattens the inner area of the tissues. The proposed SG methods does not suffer from the Gaussian oversmoothing, obtaining sharper edges depending on the prior chosen and richer details than MAC and the just described methods.

The quantitative comparison on the stain separated images was performed using the Quaternion Structural Similarity (QSSIM)[37] and the Peak Signal to Noise Ratio (PSNR) metrics. The mean value for each tissue in the dataset is presented in Table 4. The results show that the proposed L1 with $n_s = 2$ achieves outperform the competitors. The proposed LOG slightly improves the results of the TV based method PER. This table also includes the performance of our proposed methods when three color vectors are used. As we will later show, the use of a residual component facilitates the classification task, see also [5]. Although the use of three components deteriorates the quality of the stained separated images, our

Table 4: Mean PSNR and QSSIM values for all the methods on the WSSB dataset [22].

		RUI	MAC	VAH	ALS	HID	PER	ZHE	LOG $n_s = 2$	LOG $n_s = 3$	L1 $n_s = 2$	L1 $n_s = 3$
Image	Stain											
Colon	H	22.27	23.91	25.83	21.11	28.57	28.62	17.89	<u>28.66</u>	24.12	29.01	24.12
	E	20.70	21.55	26.29	21.94	27.58	27.60	14.76	<u>27.74</u>	25.31	28.38	25.31
Breast	H	15.27	26.24	25.46	24.60	28.81	29.14	15.31	<u>29.23</u>	27.56	30.50	27.56
	E	17.66	23.62	27.68	25.92	26.60	26.76	14.99	<u>26.74</u>	27.19	27.71	27.19
Lung	H	22.47	19.52	25.87	20.62	32.91	<u>33.10</u>	19.51	31.21	24.69	35.21	24.69
	E	22.05	18.09	25.53	23.95	30.77	<u>31.02</u>	16.23	29.99	25.50	33.07	25.50
Mean	H	20.00	23.22	25.72	22.11	30.10	<u>30.29</u>	17.57	29.70	25.46	31.57	25.46
	E	20.14	21.08	26.50	23.94	28.32	<u>28.46</u>	15.33	28.16	26.00	29.72	26.00
QSSIM												
Image	Stain											
Colon	H	0.8841	0.8581	0.9536	0.5369	<u>0.9635</u>	0.9163	0.7490	0.9556	0.9168	0.9696	0.9168
	E	0.5670	0.6133	0.8656	0.7642	<u>0.8713</u>	0.6111	0.4407	0.8455	0.8404	0.9011	0.8404
Breast	H	0.7721	0.9859	0.9881	0.7347	0.9919	0.6813	0.5231	0.9903	0.9852	<u>0.9918</u>	0.9852
	E	0.7721	0.8907	0.9695	0.8068	0.9598	0.5527	0.3108	0.9567	0.9594	<u>0.9605</u>	0.9594
Lung	H	0.9206	0.6973	0.9489	0.4603	0.9959	0.9519	0.7747	0.9894	0.9442	<u>0.9957</u>	0.9442
	E	0.5368	0.3500	0.8064	0.7983	<u>0.9401</u>	0.6226	0.3359	0.8807	0.8405	0.9433	0.8405
Mean	H	0.8589	0.8471	0.9635	0.5773	<u>0.9838</u>	0.8499	0.6823	0.9784	0.9488	0.9857	0.9488
	E	0.6253	0.6180	0.8805	0.7898	<u>0.9237</u>	0.5955	0.3624	0.8943	0.8801	0.9349	0.8801

methods perform similarly to some other methods (not the worst ones) in terms of PSNR and QSSIM values.

As it can be observed in Figs 3(j-m) the differences when a third component is used are difficult to distinguish. For a better visual comparison, Figure 4 shows zoomed in details from Fig. 3(k&m). Notice that we report L1 results since this method obtains the best PSNR and QSSIM values with $n_s = 2$ and the difference with the $n_s = 3$ results is wider. The difference between hematoxylin (Figs 4(a&d) and eosin (Figs 4(b&c) colors is small. The third component captures only residual information extracted from the H&E bands. The third band is discarded, which implies less fidelity to the original image. Then, the experimental design in [22] implies that removing information will lead to lower PSNR and QSSIM values. In spite of the lower figures of merit, we will see in following sections that the use of a third component leads to better classification performances.

4.1.1. Dependency on the reference color-vector $\underline{\mathbf{M}}$

The similarity prior in (6) requires the use of a reference color-vector matrix $\underline{\mathbf{M}}$. On one hand, the prior on $\underline{\mathbf{M}}$ ensures that the obtained result agrees with our previous knowledge on the H&E channels. On the other hand, it reduces the search space of feasible solutions. The prior for our model should be as accurate as possible. However, the color variability in the WSIs might hamper the accuracy of our prior. To assess the impact of the reference matrix $\underline{\mathbf{M}}$ we have evaluated a breast image on the WSSB dataset using different values of $\underline{\mathbf{M}}$. Variations of $\underline{\mathbf{M}}$ were obtained by adding random values sampled from an uniform distribution $U(-\sigma, \sigma)$, with $\sigma \in [0.05, 0.3]$. Then, each row is normalized to achieve $\|\underline{\mathbf{m}}_s\| = 1$. Twenty different color-vectors were used as prior for the L1 method. Figure 5 depicts some values for $\underline{\mathbf{M}}$, PSNR, and QSSIM as σ increases. Values of $\sigma \geq 0.2$ produce low quality values for the prior, as they do not represent the stains in the image and even reach unreal values for the H&E channels. The variations on the prior have a considerable impact on the obtained separation. The proposed method is able to deal with variations up to $\sigma = 0.1$ while obtaining values comparable to the competing methods.

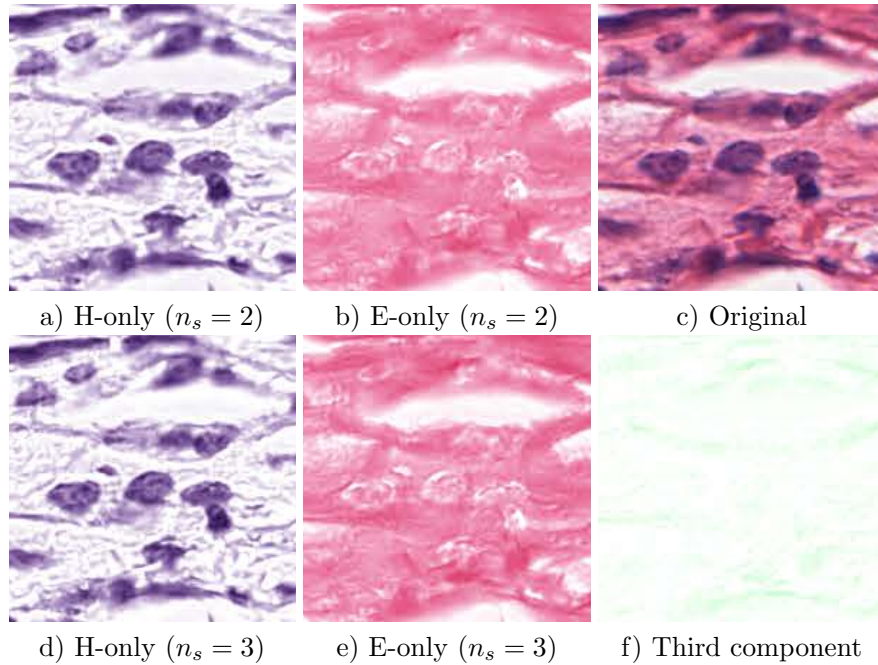


Figure 4: Detail of the of the bottom left corner of Fig. 3(a) and its H-only, E-only and third component separations. Separations in the top and bottom rows were obtained with the proposed method L1 with two components Fig. 3(k) and three components Fig. 3(m), respectively.

4.1.2. Time comparison

Using a single WSSB image, we measured the time needed for the competing methods to deconvolve the image. The comparison is shown in figure 6 as a joint plot with the PSNR values obtained. RUI, which does not require color-vector estimation, obtains the lowest time. More complex blind methods require more computational time to estimate the color-vector matrix. ZHE implements a similar deconvolution step to RUI using a similar time. ALS requires as much time as HID but its PSNR and QSSIM values are lower. The proposed approach is severely impacted by the chosen prior. Using LOG the proposed method is expensive in time cost. However L1 reduces by half the time spent by the TV-based method, PER. L1 requires a longer time than some of the competing methods but also obtains the best figures-of-merit as already reported in Table 4. Considering a third stain component increases the time required by L1 but reduces it for LOG. This is due to a higher number of parameters to estimate but less iterations required to converge, specially for LOG. L1 required 6 iterations to deconvolve the image in both cases while LOG used 10 and 6 with $n_s = 2$ and $n_s = 3$, respectively. Notice, also, that the proposed fully Bayesian approach includes estimation of all model parameters together with the stain concentrations and color-vector matrix. All these estimations increase the running time but make our methods parameter free.

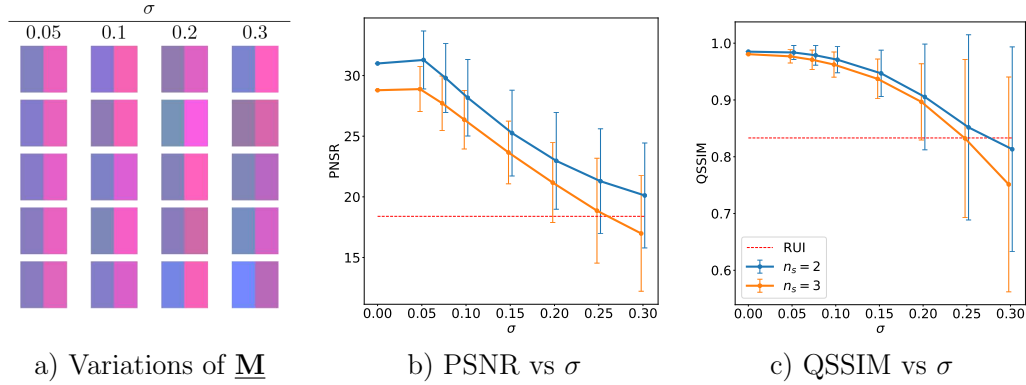


Figure 5: Evolution of the results for different values of $\underline{\mathbf{M}}$. a) Different combinations of H&E color-vectors used as $\underline{\mathbf{M}}$. Each column shows different values obtained with a fixed variance. b&c) Evolution of PSNR and QSSIM as the variance in $\underline{\mathbf{M}}$ increases, respectively. The red dashed line indicates the performance of the separation by RUI.

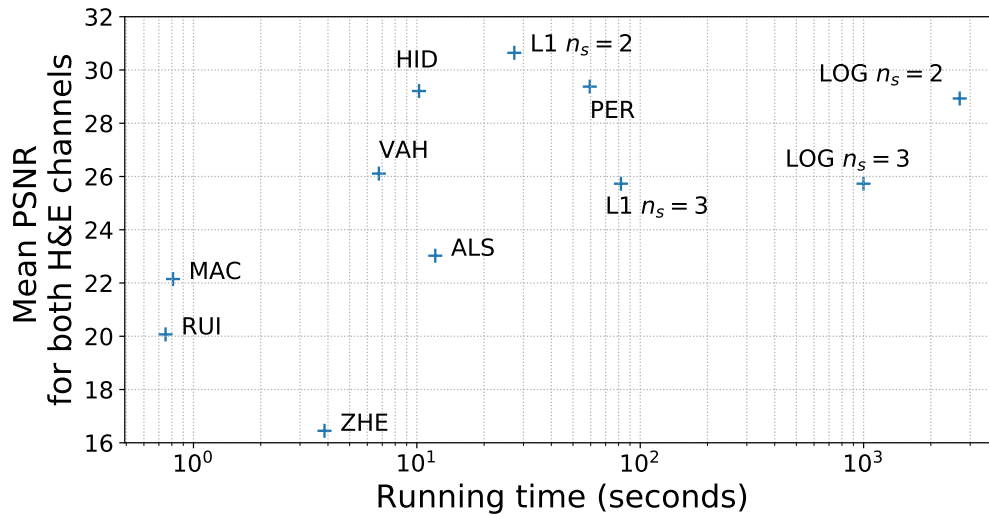


Figure 6: Joint plot of mean PSNR and running time for deconvolving a 2000x2000 image. The time is counted in seconds and the x axis is presented in logarithmic scale. The time was measured in a shared server running CentOS 7 with 32 CPU Intel(R) Xeon(R) (2.4 GHz).

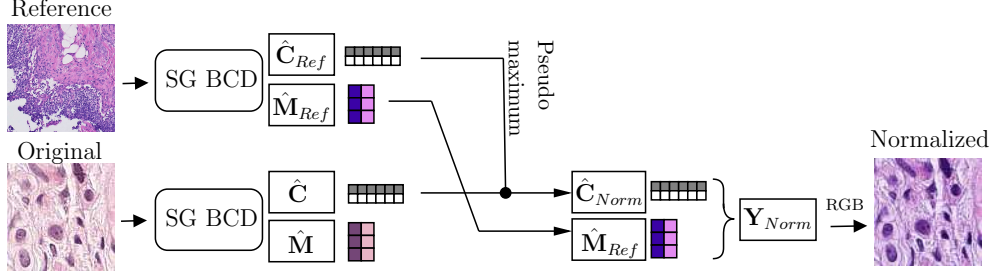


Figure 7: Pipeline of the normalization procedure. Both, reference and original image are color deconvolved. To obtain the normalized image, the dynamic range of the concentration $\hat{\mathbf{C}}$ is adjusted to be the same as that of $\hat{\mathbf{C}}_{reference}$ and the color matrix $\hat{\mathbf{M}}$ is substituted by $\hat{\mathbf{M}}_{reference}$. Then, the normalized image $\hat{\mathbf{Y}}_{Normalized}$ is transformed back to RGB space.

4.2. CN Experiments

Deep learning based CAD systems usually make use of the observed H&E images instead of the separated bands [38]. Therefore, they are highly affected by stain color variations. CN aims to provide an improved input to CAD system. The images are preprocessed to reduce the staining variations without modifying their structure. CN can easily be achieved as an additional step after BCD, as stain color information is separated from the structure of stain concentration. This section performs a comparison on the color variations between the original data and the CN obtained by all the competing methods.

To normalize the images a reference image, $\mathbf{I}_{reference}$ is used. Let $\hat{\mathbf{M}}_{reference}$ and $\hat{\mathbf{C}}_{reference}$ be the estimated color and concentration matrices in the OD space obtained using one of our proposed methods on the image $\mathbf{Y}_{reference}$ (obtained from $\mathbf{I}_{reference}$). Following [11], given a new image \mathbf{I} , the dynamic range of its corresponding $\hat{\mathbf{C}}$ is adjusted to be the same as that of $\hat{\mathbf{C}}_{reference}$ and the color matrix $\hat{\mathbf{M}}$ is substituted by $\hat{\mathbf{M}}_{reference}$ to obtain the normalized image as follows:

$$(\hat{\mathbf{Y}}_{normalized})^T = \sum_{s=1}^{n_s} -(\hat{\mathbf{m}}_s)_{reference} \hat{\mathbf{c}}_s^T \frac{P_{99}((\hat{\mathbf{c}}_s)_{reference})}{P_{99}(\hat{\mathbf{c}}_s)} \quad (29)$$

where $P_{99}(\mathbf{v})$ represents the pseudo maximum (99%) of vector \mathbf{v} . The normalized RGB image $\hat{\mathbf{I}}_{normalized}$ is then

$$\hat{\mathbf{I}}_{normalized} = \exp_{10} \hat{\mathbf{Y}}_{normalized} \quad (30)$$

Figure 7 depicts the pipeline followed to obtain the normalized image.

To measure the quality of a CN procedure, we use the normalized median intensity (NMI) measure [39] defined as

$$NMI(\mathbf{I}) = Median(\mathbf{u})/P_{95}(\mathbf{u}) \quad (31)$$

where \mathbf{I} denotes a WSI and \mathbf{u} is a vector where each u_i component is the mean value of the R, G, and B channels at the i -th pixel, [40].

The NMI value is calculated for each WSI in a given dataset. However, we require information about the distribution of the NMI values in the dataset. Then, the standard deviation of the NMI values in the dataset (NMI SD) and the coefficient of the variation

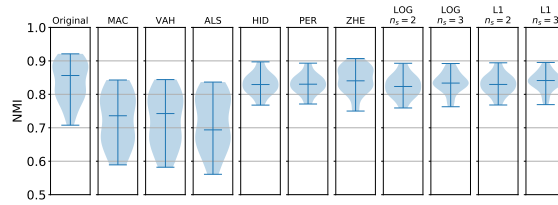


Figure 8: Violin plots of NMI values for the original and normalized images by the compared methods on the SICAP-HUVNGR dataset. The bars mark the maximum, median and minimum values for each plot.

(NMI CV), i.e., NMI SD divided by mean of the dataset, were used as metrics. Lower values of NMI SD and NMI CV indicate a more consistent normalization.

Three datasets containing images from different centers were used in this section. SICAP-HUVNGR, Camelyon-16 and Camelyon-17, see Section 3.

In the SICAP-HUVNGR dataset, to avoid the influence of large background regions, 512x512 pixel patches at 40x magnification, with at least 70% tissue, were sampled from each WSI. This patch size is motivated by the prostate slide appearance. They are narrow tissue segments surrounded by background which is also visible inside glands. The use of a larger patch size, while maintaining the above tissue percentage, discard most patches containing glands and keep only stroma patches mainly stained with eosin, because they have low nuclear density. The NMI for each WSI is calculated over all the pixels in the patches. The number of patches used from each WSI was evaluated from 20 to 120, observing that beyond 60 the NMI value did not change.

For Camelyon-16 and Camelyon-17 datasets, 224x224 pixel patches, with at least 70% tissue, were sampled from each WSI. This will also be the patch size used for classification, see section 4.4. Following [24], 500 patches were sampled from each WSI in the datasets for CN and classification purposes.

Let us now describe the obtained results. First we notice that RUI does not estimate the color-vectors in the images, therefore it is not possible to use it for CN. Furthermore, the prior color-vector matrix $\underline{\mathbf{M}}$ used by our method is fixed to the standard proposed by Ruifrok *et al.*[8].

Table 5: NMI SD and NMI CV comparison for different normalization methods on SICAP-HUVNGR dataset.

Method	NMI SD	NMI CV
Original Data	0.0591	0.0705
MAC	0.0782	0.1079
VAH	0.0796	0.1099
ALS	0.0799	0.1114
HID	0.0313	0.0378
PER	0.0296	<u>0.0356</u>
ZHE	0.0398	0.0472
LOG $n_s = 2$	0.0330	0.0400
LOG $n_s = 3$	0.0307	0.0368
L1 $n_s = 2$	<u>0.0306</u>	0.0368
L1 $n_s = 3$	0.0287	0.0342

NMI values for the SICAP-HUVNGR dataset are shown in Table 5. The proposed methods, LOG and L1, reduce by half the NMI SD and NMI CV values of the original data. L1 obtains the best value with $n_s = 3$. ZHE significantly reduces both values, but the results

are not as clustered as the obtained by HID and PER. MAC, VAH, and ALS do not improve the initial NMI values. Figure 8 depicts the distribution of NMI values using violin plots. In the first column of the figure, two different NMI distributions can be appreciated on the original data. They correspond to the two centers the images come from. The two centers are still visible when MAC, VAH, ALS and ZHE are used, but disappear when HID, PER, and the proposed L1 and LOG are utilized. The proposed L1 and LOG correctly identify the H&E distribution on the WSIs. When CN is applied, the color distribution is equalized for all the WSIs and the color properties of each stain are fixed to those in the reference image, reducing the NMI SD and CV values.

The CN analysis on Camelyon databases is provided below. Due to the computational cost of CD and parameter estimation (See figure 6) on the large volume of WSIs in those databases and also to the superior performance in previous experiments (See tables 4 & 5) only the proposed L1, and not LOG, was used in the comparison.

In addition to undesired color variance due to the staining procedure and also to the acquisition system used, pathology related color variations also appear in the WSIs (e.g: tumor images usually have a higher percentage of hematoxylin pixels). The fully labeled Camelyon-16 allows us to study the pathological color variance. For that matter, NMI SD and NMI CV were calculated for the whole dataset and for the tumor, normal and test WSIs as separated subsets. NMI SD and NMI CV values obtained for the Camelyon-16 dataset are shown in Table 6 and Figure 9. The best result for the complete dataset is obtained by ZHE, closely followed by our proposed L1. However, in the separated normal and tumor subsets, the proposed method obtained the best values. Images normalized by our method are more similar to those in the same subset, but the difference between classes is preserved. The proposed L1 method with $n_s = 3$ obtains higher NMI values than the original dataset when all images are considered, however it is reduced in the normal and tumor subsets. This is caused by a wide separation on the colors for the hematoxylin and eosin channels, that will be useful for classification as we will see in the following sections.

Table 6: NMI SD and NMI CV comparison for diferent normalization methods on Camelyon-16.

database subset Method	Camelyon-16							
	All images		Tumor		Normal		Test	
	SD	CV	SD	CV	SD	CV	SD	CV
Original Data	0.0629	0.0860	0.0497	0.0693	0.0528	0.0684	0.0538	0.0778
Macenko	0.0799	0.1359	0.0553	0.0826	0.0629	0.1122	0.0678	0.1221
Vahadane	0.1127	0.2112	0.0877	0.1404	0.0741	0.1471	0.1274	0.2573
Alsubaie	0.0698	0.1262	0.1186	0.2015	0.1048	0.1540	0.1923	0.3271
Hidalgo-Gavira	0.0645	0.0915	0.0373	0.0480	0.0552	0.0795	<u>0.0378</u>	<u>0.0572</u>
Pérez-Bueno	0.0624	0.0900	<u>0.0375</u>	0.0492	0.0506	0.0740	0.0351	0.0539
Zheng	0.0477	0.0616	0.0394	0.0519	<u>0.0396</u>	0.0516	0.0551	0.0693
ℓ_1 prior $n_s = 2$	<u>0.0532</u>	<u>0.0775</u>	0.0376	<u>0.0491</u>	0.0357	<u>0.0549</u>	0.0532	0.0785
ℓ_1 prior $n_s = 3$	0.0793	0.1136	0.0493	0.0622	0.0617	0.0910	0.0457	0.0708

In Figure 9.(a-c) we observe that the NMI variation on Camelyon-16 dataset comes not only from different centers but also from different pathologies. Images in tumor, and normal image subsets show different distributions on the original data. The normalized images by HID, PER, and the proposed L1 preserve those differences, keeping a separation on the median NMI value of both subsets. ZHE, designed to optimize NMI values, tends to over-normalize the images, eliminating most of the NMI difference between tumor and normal subsets.

Table 7: NMI SD and NMI CV comparison for different normalization methods on Camelyon-17 dataset.

Method	All images		Non-Negative		Negative	
	SD	CV	SD	CV	SD	CV
Original Data	0.0773	0.1035	0.0812	0.1087	0.0750	0.1004
Macenko	0.1031	0.1689	0.0993	0.1581	0.1040	0.1731
Vahadane	0.1058	0.1823	0.1010	0.1685	0.1069	0.1878
Alsubaie	0.0992	0.1806	0.0989	0.1753	0.0984	0.1819
Hidalgo-Gavira	0.0635	0.0948	0.0671	0.0987	0.0606	0.0913
Pérez-Bueno	0.0629	0.0941	<u>0.0668</u>	<u>0.0984</u>	0.0598	0.0902
Zheng	0.0489	0.0631	0.0488	0.0628	0.0489	0.0632
ℓ_1 prior $n_s = 2$	<u>0.0624</u>	<u>0.0935</u>	0.0720	0.1051	<u>0.0534</u>	<u>0.0813</u>
ℓ_1 prior $n_s = 3$	0.0793	0.1136	0.0684	0.1037	0.0638	0.0994

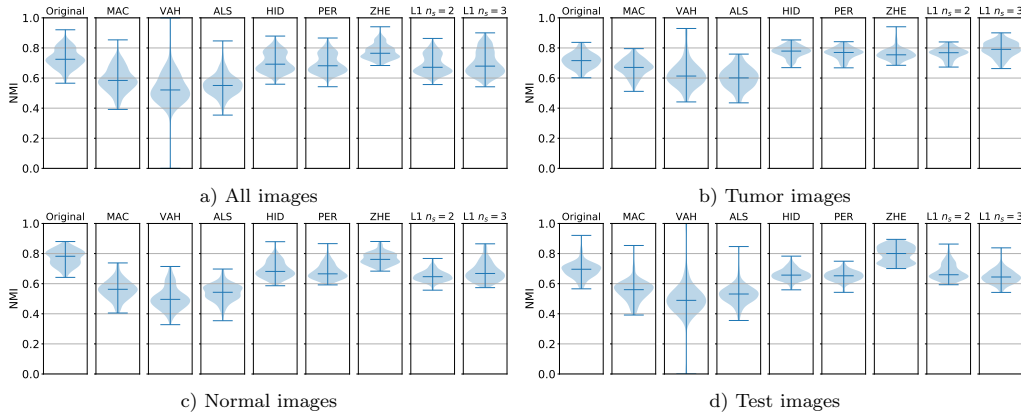


Figure 9: Violin plots of NMI values for the normalized patches from Camelyon-16 dataset by the compared methods. The bars mark the maximum, median and minimum values for each plot.

The NMI values obtained for Camelyon-17 are shown in Table 7 and Figure 10. The labeling is more complex on this dataset (See Table 2), so the NMI is calculated for the negative (normal WSIs) and non negative (itc, micro and macro) subsets, along with the full dataset. The original WSIs from Camelyon-17 have larger color variations than previous datasets, although they are not as balanced in terms of normal and tumoral WSIs. From Figure 10(a-c) it can be appreciated that the subset distributions are similar to the whole dataset distribution, meaning that the NMI differences caused by pathologies are overwhelmed by the differences between centers. In the non-negative subset there is also variation due to the significant differences between itc, micro and macro. The lower NMI SD and CV values are obtained by ZHE. The proposed L1 with $n_s = 2$ obtains the second lowest value in most cases. L1 obtained its lowest values on the negative subset while maintaining a wide distribution on the non negative, probably due to the inter-subset differences mentioned. The Bayesian methods HID and PER show similar results to the proposed one.

To conclude this section we include Figure 11 to qualitatively compare the CNs obtained by the competing methods. The reference image and some of their 224x224 extracted patches are shown in the first row. The remaining rows contain patches from different WSIs in the Camelyon-16 dataset normalized using the competing methods. MAC and VAH tend to saturate the color in the images. ALS introduces artifacts in some of the patches. ZHE overbrightens the images. HID and PER effectively transformed the color to that of the reference image. The proposed L1 keeps the structure and tissue differences but set the stain properties

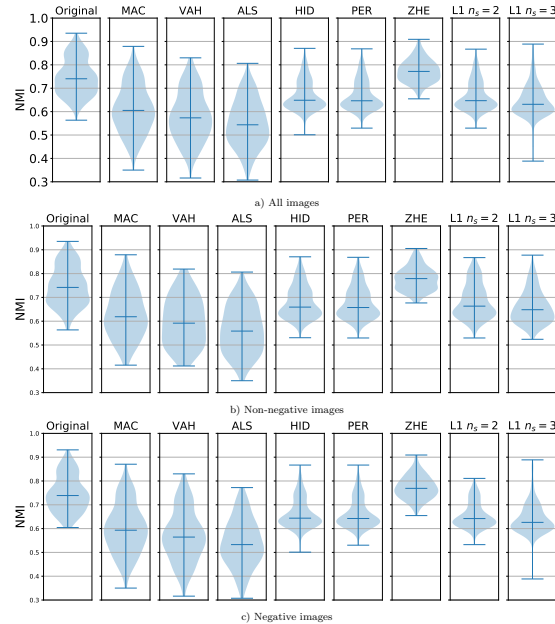


Figure 10: Violin plots of NMIs for the normalized patches from Camelyon-17 dataset by the compared methods. The bars mark the maximum, median and minimum values for each plot.

to those observed in the reference image. The normalization with the proposed L1 and $n_s = 2$ is the most similar to the reference image. When using a third component, the eosin is clearer, and more distinguishable from the hematoxylin. The difference between patches is higher, but the stains keep the common color properties. The effect of the residual component is clearly appreciated in the first and third rows of the last column. Although the removal of the residual produces artifacts, small hematoxylin structures are eliminated and nuclei appear more clearly separated. As discussed in previous sections, discarding the residual reduces the fidelity to the original image. In the following sections, we will demonstrate the beneficial effect of the third component on classification tasks.

4.3. Deconvolution based classification

BCD allows CAD systems to use the single stained bands separately, which can improve the classifier performance [4]. The separated H&E concentrations are used to extract features and train four different classifiers. The prostate cancer histopathological SICAPv1 database [4] was used for this purpose. In the 10x scale, we use patches of size 1024x1024 pixels with the purpose of capturing complete glands within the patches. Training patches have 50% overlap and we discarded those containing mostly background (75%). From the WSI annotated as benign we obtained 1909 negative patches. A minimum of 25% of malignant tissue was required for malignant patches, obtaining 344 pathological ones.

The proposed and competing methods were used to color deconvolve the dataset. Following [4], the hematoxylin and eosin OD concentration images were used to extract the concatenation of Geodesic granulometries (GeoGran)[4] and Local Binary Patterns Variance (LBPV) [41] features. The H&E GeoGran descriptor was proposed in [4] for prostate cancer

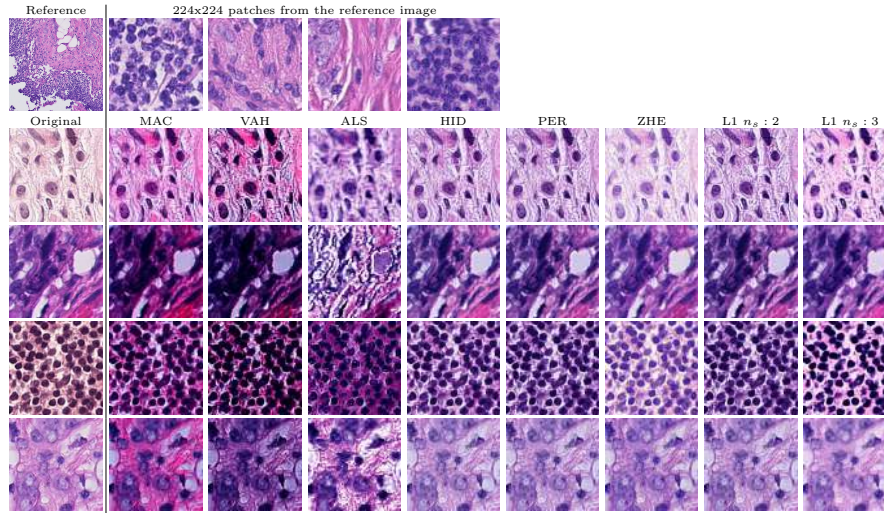


Figure 11: Example patches from different WSIs in Camelyon-16. The first row shows the reference image used as a sample of 224x224 patches extracted from the reference. The original patch is shown in the first column.

classification, obtaining stain-specific information. From the hematoxylin, it recovers the gland frontiers formed by the nuclei structure (those that enclosed their lumen and cytoplasm). From the eosin, it encodes how the stroma is affected by the lumen and nuclei structure. LBPV features are extracted from the hematoxylin band to capture texture and contrast information. The use of both Geogran and LBPV features, recovers texture and structural information in the stain separated bands, and has been proven to be an accurate descriptor for histopathological image classification [4].

With the described descriptors, the following set of state-of-art classifiers were trained: Random Forest (RF) [42, 43], Extreme Gradient Boosting (XgBoost) [44], Gaussian Processes (GP) [45] and Deep Gaussian Processes (DGP)[46]. The classifiers were configured following [4] to achieve an unbiased classification benchmark. For RF and XgBoost we use 1000 estimators and a maximum depth of 20 and 30, respectively. The learning rate for XgBoost is fixed to 0.01. GP and DGP classifiers were configured following the same approach as in [4]. A GP classifier with Radial Basis Function (RBF) kernel [47] using variational inference and a three-layer DGP classifier with RBF kernel and 100 inducing points per layer, following the doubly stochastic inference proposed in [48]. DGP uses a mini-batch size of 1000 and the inducing points were initialized using kmeans. Both models GP and DGP were optimized using Adam with a learning rate of 0.01.

To tackle the unbalance of positive and negative patches (common in cancer classification), we use a five-fold cross-validation. Each patient is assigned to a single fold to avoid correlation between training and testing sets. With this configuration, each classifier was built using all positive patches and a subset of the negative ones. The classifiers were trained from scratch using each deconvolution method

AUCs obtained by all the compared methods are shown in Table 8 and Figure 12. Since HID oversmooths the images, it performs worse as it happens to methods like MAC which

Table 8: AUC obtained by different classifiers when trained with different deconvolution methods on the SICAPv1 cross validation.

Method	RF	GP	XgBoost	DGP
RUI	0.9789	<u>0.9855</u>	0.9764	0.9737
MAC	0.9315	0.9535	0.9425	0.8802
VAH	0.9222	0.9479	0.9295	0.9420
ALS	0.9262	0.9442	0.9246	0.9344
HID	0.9157	0.9542	0.9228	0.8997
PER	0.9798	0.9856	<u>0.9797</u>	0.9718
ZHE	0.9194	0.9420	0.9263	0.9251
LOG $n_s = 2$	0.9256	0.9497	0.9281	0.9303
LOG $n_s = 3$	<u>0.9796</u>	0.9842	0.9798	0.9723
L1 $n_s = 2$	0.9256	0.9497	0.9281	0.9303
L1 $n_s = 3$	<u>0.9796</u>	0.9842	0.9796	<u>0.9729</u>

obtain less detailed images. ZHE scores poorly even when its deconvolution step is based on RUI. Although L1 and LOG using $n_s = 2$ do not obtain the best results, the use of $n_s = 3$ leads to a performance comparable to RUI and PER. With XgBoost, LOG obtains the best result. Notice that the best AUC (0.9856) is obtained using PER and GP, closely followed by RUI, L1, and LOG. Notice also that L1 and LOG perform very similarly. This is due to the very close estimated color vector matrix which leads to very similar extracted features.

The results obtained by L1 and LOG are in agreement with those obtained in our previous work [5]. Including a third residual component ($n_s = 3$) in the deconvolution step leads to better classification performance although the obtained stain separation is not as close to the ground-truth separation as that obtained using $n_s = 2$. Despite of a lower fidelity, the information captured by the residual channel makes the nuclei in the hematoxylin channel to appear more clearly separated and with less noise. The distribution of nuclei is usually considered be the most determinant feature for classification [4]. We believe this is the most plausible reason for the discriminative power of the residual band.

4.4. Normalization based classification

As we have already indicated, CN can be considered as a preprocessing step whose goal is to increase the performance of CAD systems[49], specially those using as input the original RGB images. To conclude the experimental section, in our last experiment we compare the performance of VGG19 [50], a common CNN used in cancer classification [38, 4], when it is fed with the original and color normalized patches. We also analyze the VGG19 performance when trained and tested using the OD concentrations obtained by the different methods, as they can be seen as a two channel image. Figure 13 shows an example of Camelyon-16 patches and their OD concentration channels.

From the patches extracted in section 4.2, 55,000 tumor annotated (positive class) patches and 55,000 normal (negative class) patches from negative WSIs were randomly sampled from each Camelyon dataset training set, see Section 3. Since Camelyon-17 contains only 50 tumor annotated WSIs, to complete its 55,000 tumor annotated patches, additional tumor patches were extracted following the procedure described in section 4.2. Using the above protocol, Camelyon-16 testing set contains approximately 19000 tumor patches, and from this testing dataset 19000 normal patches were sampled. VGG19 was trained and tested in two scenarios. In the first case, we explore how normalization affects performance within a single database (using Camelyon-16 training and testing set). In the second scenario we use Camelyon-17 for

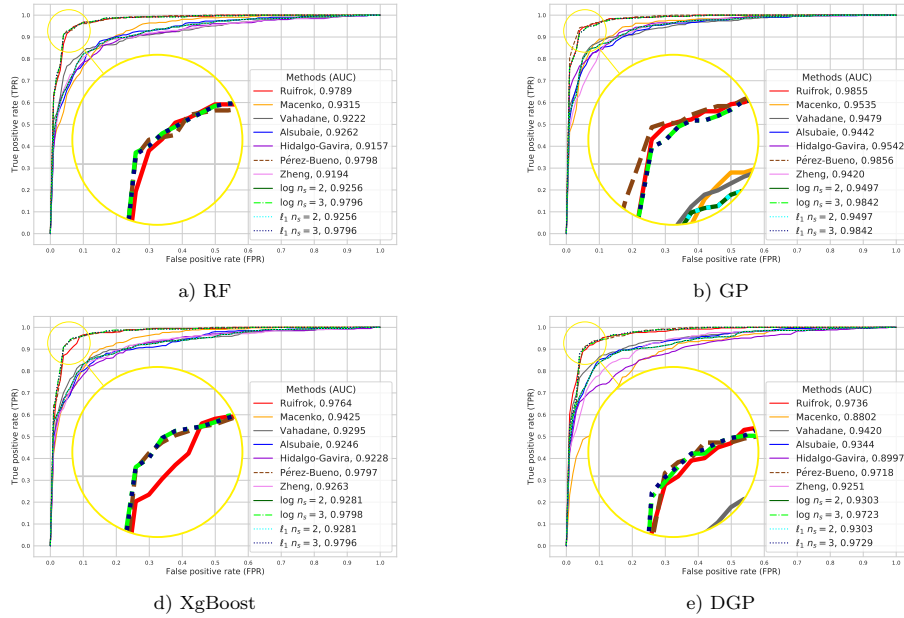


Figure 12: ROC curves and AUC for the competing methods and classifiers on the SICAPv1 dataset. Each sub-image contains a single classifier trained with all deconvolution methods.

training and since Camelyon-17 test set labels are not available, Camelyon-16 is used for test. This experiment provides information on the inter-generalization capabilities of the model.

VGG19 with batch normalization was trained during 100 epochs in each case, which was enough for the network to converge. A batch size of 64 samples was used, constrained by the available memory of the Nvidia Titan X GPU utilized in this work. The learning rate was initially set to 0.01 and was reduced by factor 0.5 each 30 epochs. AUCs were calculated on the test set using the training best performing epoch for each method.

Table 9: AUC Performance of the VGG19 over Camelyon-16 testing set using CN images and OD concentrations obtained by the proposed and competing deconvolution methods.

Method	Training set			
	Camelyon-16 CN	Camelyon-16 OD	Camelyon-17 CN	Camelyon-17 OD
Original images	0.9491	NA	0.9279	NA
RUI	NA	0.9458	NA	0.9003
MAC	0.9564	0.9608	0.8652	0.8503
ALS	0.9557	0.9556	0.9144	0.8874
HID	0.9479	0.9558	0.9042	0.7994
PER	<u>0.9627</u>	0.9552	0.9106	0.8941
ZHE	0.9466	<u>0.9621</u>	<u>0.9370</u>	<u>0.9380</u>
L1 $n_s = 2$	0.9656	0.9429	0.9289	0.9009
L1 $n_s = 3$	0.9505	0.9634	0.9378	0.9541

The obtained AUCs are shown in Table 9 and the ROC curves in Figure 14. Notice that for Camelyon-16, VGG19 performs well on the original images (better than some of the methods). The proposed L1 with $n_s = 2$ is however the best feed to VGG19 since its AUC increases from 0.9479 (original data) to 0.9656. The oversmoothing of the edges by HID

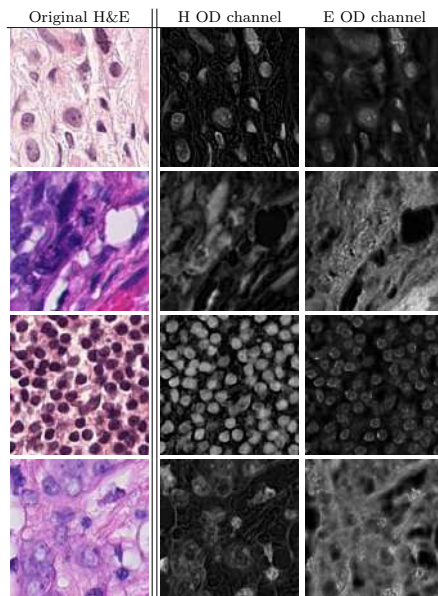


Figure 13: Examples of the OD concentrations channels obtained by the proposed method L1 $n_s = 2$ for different patches and used to train VGG19 using a 2-band image as input.

and the overnormalization of ZHE obtained a slightly lower value than the non-normalized original data. The use as input to VGG19 of Camelyon-16 OD concentrations, was a boost for the methods ZHE and the proposed L1 using $n_s = 3$, and had a slightly beneficial effect for most methods.

Camelyon-17 training set contains more WSIs than Camelyon-16, furthermore its color variance is considerable as images come from 5 different centers. An adequate preprocessing has a higher impact on the generalization capability of the CNN. In this case, the original data reached an AUC=0.9279. Using CN, the proposed L1 with $n_s = 3$ obtained the best result with 0.9377. ZHE performs better in this experiment than in the previous one. In this inter database case, only ZHE and L1 with $n_s = 3$ were boosted by the use of OD concentrations to train the network. The L1 AUC raised to 0.9541 when using $n_s = 3$.

The effect of using a third component was limited using normalized images in the Camelyon-16 dataset. However, this configuration obtained the best performance in OD and in both cases when using Camelyon-17. As discussed in previous sections, the most plausible reason is that the third residual component makes nuclei to appear more different from other structures. This effect can be also appreciated on normalized images in Figure 11. Where the patches in the last column show a bigger difference between hematoxylin and eosin colors.

5. Discussion

BCD is a critical step towards normalization and classification of histological images. The stain separation allows to measure the fidelity to the tissue and facilitate feature extraction. The obtained results clearly show that SG priors are a good choice for color deconvolution of histopathological images. As previously indicated, each stain should fix only and com-

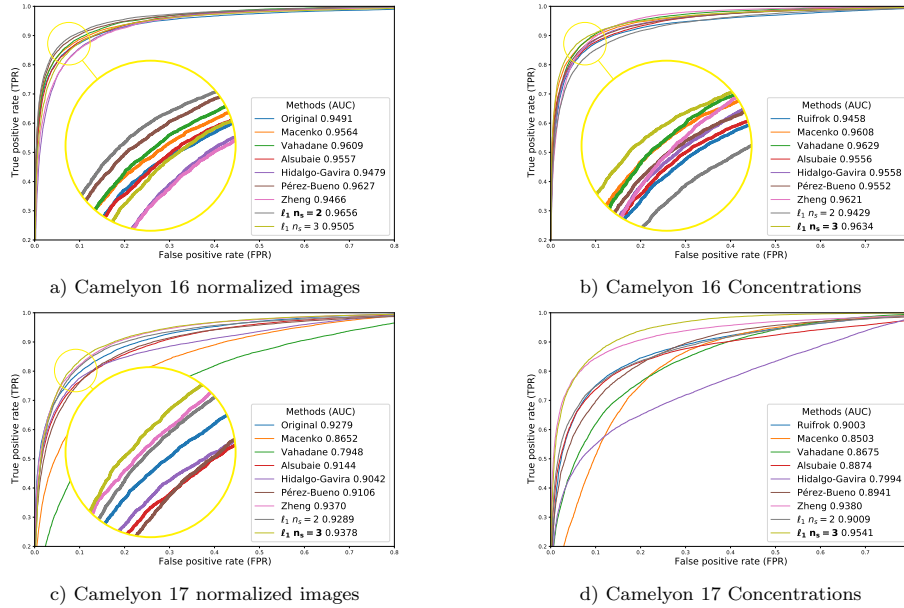


Figure 14: ROC curve and AUC obtained by the VGG19 when trained with normalized images and OD concentrations by competing methods. Testing set is always the one from Camelyon 16. a) Training set from Camelyon 16 normalized. b) Training set from Camelyon 16 concentrations. c) Training set from Camelyon 17 normalized. d) Training set from Camelyon 17 concentrations.

pletely to specific proteins on the tissue, leading to sparse stain concentration differences at neighbouring pixels [11]. However, the experimental results show that the sparsity on the differences is moderated. The l_1 prior, with a lower kurtosis than the log prior, allows to keep more non-zero values. This makes l_1 a good prior for this problem, as its induced sparsity is softer than that of the log prior.

We have analyzed the effect of using two or three stain components in our proposed approach to deconvolution. The carried out experiments indicate that using two components produces stains closer to the original ones and also provides good CN. The use of a third component to capture residual information from the H&E images, makes it possible to obtain a clearer stain separation. In the hematoxylin band, the nuclei appear more clearly enhancing nucleus information and the noisy background is reduced. The effect of the third component in the eosin band is reduced but the contrast is increased. Then, we should choose whether to use the third component for BCD depending on our goal. Its use may reduce the fidelity to the tissue in terms of PSNR and SSIM values, but it improves the performance of feature based and CNN classification methods, improving class separation and helping the descriptors or CNN layers to capture the relevant information.

Finally, the use of BCD allows to extract stain-specific information from H&E channels. Our comparison between classification using the normalized images and OD concentrations have shown that CN of histopathological images improves the performance of CNN methods, however the use of CD to obtain the separated H&E concentrations leads to better performance. The H&E separation is directly provided to the CNN by the OD concentration and directly related to a better class separation.

6. Conclusions

In this work we have proposed the use of SG priors for blind color deconvolution of histological images. The framework presented includes a novel variational Bayesian blind color deconvolution algorithm which automatically estimates the color-vector matrix, the concentration of stains, and all the model parameters. SG priors are used to model neighbouring pixel differences. The use of the SG family is a powerful tool to fine tuning the sparsity of concentration differences, reducing the noise in the images while preserving the tissue structure without oversmoothing the edges. Two penalty functions, named L1 and LOG, corresponding to SG distribution have been used. The information obtained through the proposed deconvolution guarantees fidelity to the tissue structure and can be used both for normalization and classification of histological images.

The proposed LOG and L1 methods have been experimentally compared to classical and state-of-art methods on a set of experiments covering the most common histological color deconvolution related tasks: stain separation, image normalization and cancer classification. They obtained very good results on all the performed experiments.

We have analyzed the effect of using a third residual stain component during deconvolution, showing that an affordable reduction of the fidelity to the tissue improves classification performance using descriptors or CNN classifiers

Finally, our study includes a comparison between classification using the normalized images and OD concentrations showing that although CN improves the performance of classifiers over the raw data, stain separated OD concentrations lead to better classification performance.

Acknowledgments

This work was sponsored in part by the Agencia Estatal de Investigación under project PID2019-105142RB-C22 / AEI / 10.13039/501100011033 and the work by Fernando Pérez-Bueno was sponsored by Ministerio de Economía, Industria y Competitividad under FPI contract BES-2017-081584.

Statements of ethical approval

The data used in this publication has been either publicly available or have been ethically approved for scientific use by the Ethics committee of the San Cecilio University Hospital (Granada, Spain).

Conflict of interest

The authors have no conflicts of interest to declare.

References

- [1] A. H. Fischer, K. A. Jacobson, J. Rose, and R. Zeller, *Hematoxylin and Eosin Staining of Tissue and Cell Sections*, Cold Spring Harbor Protocols, 2008.

- [2] M. Salvi, U. R. Acharya, F. Molinari, and K. M. Meiburger, “The impact of pre- and post-image processing techniques on deep learning frameworks: A comprehensive review for digital pathology image analysis,” *Computers in Biology and Medicine*, vol. 128, p. 104129, 2021.
- [3] T. A. A. Tosta, P. R. de Faria, L. A. Neves, and M. Z. do Nascimento, “Computational normalization of H&E-stained histological images: Progress, challenges and future potential,” *Artificial Intelligence in Medicine*, vol. 95, pp. 118 – 132, 2019.
- [4] A. E. Esteban, M. Lopez-Perez, A. Colomer, M. A. Sales, R. Molina, and V. Naranjo, “A new optical density granulometry-based descriptor for the classification of prostate histological images using shallow and deep Gaussian processes,” *Computer Methods and Programs in Biomedicine*, vol. 178, pp. 303–317, 2019.
- [5] F. Pérez-Bueno, M. López-Pérez, M. Vega, J. Mateos, V. Naranjo, R. Molina, and A. K. Katsaggelos, “A tv-based image processing framework for blind color deconvolution and classification of histological images,” *Digital Signal Processing*, vol. 101, p. 102727, 2020.
- [6] F. G. Zanjani, S. Zinger, B. E. Bejnordi, J. A. W. M. van der Laak, and P. H. N. de With, “Stain normalization of histopathology images using generative adversarial networks,” in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI)*, 4 2018, pp. 573–577.
- [7] M. Shaban, C. Baur, N. Navab, and S. Albarqouni, “Staingan: Stain style transfer for digital histological images,” vol. 2019-April, 2019, pp. 953–956, cited By 41.
- [8] A. C. Ruifrok and D. A. Johnston, “Quantification of histochemical staining by color deconvolution,” *Analytical and quantitative cytology and histology*, vol. 23, pp. 291–299, 2001.
- [9] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley, “Color transfer between images,” *IEEE Computer Graphics and Applications*, vol. 21, no. 5, pp. 34–41, 9 2001.
- [10] A. Rabinovich, S. Agarwal, C. Laris, J. H. Price, and S. J. Belongie, “Unsupervised color decomposition of histologically stained tissue samples,” in *Advances in Neural Information Processing Systems*, 2004, pp. 667–674.
- [11] A. Vahadane, T. Peng, A. Sethi, S. Albarqouni, L. Wang, M. Baust, K. Steiger, A. M. Schlitter, I. Esposito, and N. Navab, “Structure-preserving color normalization and sparse stain separation for histological images,” *IEEE Transactions on Medical Imaging*, vol. 35, pp. 1962–1971, 2016.
- [12] J. Xu, L. Xiang, G. Wang, S. Ganesan, M. Feldman, N. N. Shih, H. Gilmore, and A. Madabhushi, “Sparse non-negative matrix factorization (SNMF) based color unmixing for breast histopathological image analysis,” *Computerized Medical Imaging and Graphics*, vol. 46, pp. 20–29, 2015.

- [13] M. Macenko, M. Niethammer *et al.*, “A method for normalizing histology slides for quantitative analysis,” in *International Symposium on Biomedical Imaging (ISBI)*, 2009, pp. 1107–1110.
- [14] M. T. McCann, J. Majumdar *et al.*, “Algorithm and benchmark dataset for stain separation in histology images,” in *International Conference on Image Processing (ICIP)*, 2014, pp. 3953–3957.
- [15] A. Anghel, M. Stanisavljevic, S. Andani, N. Papandreou, J. H. Rüschoff, P. Wild, M. Gabrani, and H. Pozidis, “A high-performance system for robust stain normalization of whole-slide images in histopathology,” *Frontiers in Medicine*, vol. 6, p. 193, 2019.
- [16] D. Carey, V. Wijayathunga, A. Bulpitt, and D. Treanor, “A novel approach for the colour deconvolution of multiple histological stains,” in *Proceedings of the 19th Conference of Medical Image Understanding and Analysis*, 2015, pp. 156–162.
- [17] M. Gavrilovic, J. C. Azar *et al.*, “Blind color decomposition of histological images,” *IEEE Transactions on Medical Imaging*, vol. 32, pp. 983–994, 2013.
- [18] A. M. Khan, N. Rajpoot, D. Treanor, and D. Magee, “A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution,” *IEEE Transactions on Biomedical Eng.*, vol. 61, no. 6, pp. 1729–1738, 2014.
- [19] M. Salvi, N. Michielli, and F. Molinari, “Stain color adaptive normalization (scan) algorithm: Separation and standardization of histological stains in digital pathology,” *Computer Methods and Programs in Biomedicine*, vol. 193, p. 105506, 2020.
- [20] N. Trahearn, D. Snead, I. Cree, and N. Rajpoot, “Multi-class stain separation using independent component analysis,” in *Medical Imaging 2015: Digital Pathology*, 2015, p. 94200J.
- [21] N. Alsubaie, S. E. A. Raza, and N. Rajpoot, “Stain deconvolution of histology images via independent component analysis in the wavelet domain,” in *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, 2016, pp. 803–806.
- [22] N. Alsubaie, N. Trahearn, S. E. A. Raza, D. Snead, and N. Rajpoot, “Stain deconvolution using statistical analysis of multi-resolution stain colour representation,” *PLOS ONE*, vol. 12, p. e0169875, 2017.
- [23] L. Astola, “Stain separation in digital bright field histopathology,” in *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, 2016, pp. 1–6.
- [24] Y. Zheng, Z. Jiang, H. Zhang, F. Xie, J. Shi, and C. Xue, “Adaptive color deconvolution for histological WSI normalization,” *Computer Methods and Programs in Biomedicine*, vol. 170, pp. 107–120, Mar. 2019.

- [25] N. Hidalgo-Gavira, J. Mateos, M. Vega, R. Molina, and A. K. Katsaggelos, “Variational Bayesian blind color deconvolution of histopathological images,” *IEEE Transactions on Image Processing*, vol. 29, no. 1, pp. 2026–2036, 2020.
- [26] S. D. Babacan, R. Molina, M. N. Do, and A. K. Katsaggelos, “Blind deconvolution with general sparse image priors,” in *ECCV*, 2012.
- [27] X. Zhou, M. Vega, F. Zhou, R. Molina, and A. K. Katsaggelos, “Fast bayesian blind deconvolution with huber super gaussian priors,” *Digital Signal Processing*, vol. 60, pp. 122–133, 2017.
- [28] F. Pérez-Bueno, M. Vega, V. Naranjo, R. Molina, and A. K. Katsaggelos, “Super gaussian priors for blind color deconvolution of histological images,” in *International Conference on Image Processing (ICIP)*, 2020.
- [29] —, “Fully automatic blind color deconvolution of histological images using super gaussians,” in *27th European Signal Processing Conference, EUSIPCO 2020*, 2020.
- [30] R. Rockafellar, *Convex analysis*. Princeton University Press, 1996.
- [31] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006, pp. 454–455.
- [32] S. Kullback, *Information Theory and Statistics*. Dover Pub., 1959.
- [33] C. L. Srinidhi, O. Ciga, and A. L. Martel, “Deep neural network models for computational histopathology: A survey,” 2019.
- [34] G. Litjens, P. Bandi, B. Ehteshami Bejnordi, O. Geessink, M. Balkenhol, P. Bult, A. Halilovic, M. Hermsen, R. van de Loo, R. Vogels, Q. F. Manson, N. Stathonikos, A. Baidoshvili, P. van Diest, C. Wauters, M. van Dijk, and J. van der Laak, “1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset,” *GigaScience*, vol. 7, no. 6, 05 2018, giy065.
- [35] P. Bándi, O. Geessink, Q. Manson, M. Van Dijk, M. Balkenhol, M. Hermsen, B. Ehteshami Bejnordi, B. Lee, K. Paeng, A. Zhong, Q. Li, F. G. Zanjani, S. Zinger, K. Fukuta, D. Komura, V. Ovtcharov, S. Cheng, S. Zeng, J. Thagaard, A. B. Dahl, H. Lin, H. Chen, L. Jacobsson, M. Hedlund, M. Çetin, E. Halıcı, H. Jackson, R. Chen, F. Both, J. Franke, H. Küsters-Vandeveldel, W. Vreuls, P. Bult, B. van Ginneken, J. van der Laak, and G. Litjens, “From detection of individual metastases to classification of lymph node status at the patient level: The camelyon17 challenge,” *IEEE Transactions on Medical Imaging*, vol. 38, no. 2, pp. 550–560, 2019.
- [36] G. Landini, “Colour deconvolution,” <https://blog.bham.ac.uk/intellimic/g-landini-software/colour-deconvolution/>, accessed: 2019-10-30.
- [37] A. Kolaman and O. Yadid-Pecht, “Quaternion structural similarity: A new quality index for color images,” *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, vol. 21, pp. 1526–36, 12 2011.

- [38] C. L. Srinidhi, O. Ciga, and A. L. Martel, “Deep neural network models for computational histopathology: A survey,” *Medical Image Analysis*, vol. 67, p. 101813, 2021.
- [39] A. Basavanthally and A. Madabhushi, “Em-based segmentation-driven color standardization of digitized histopathology,” vol. 8676, 03 2013, p. 86760G.
- [40] B. E. Bejnordi, G. Litjens, N. Timofeeva, I. Otte-Höller, A. Homeyer, N. Karssemeijer, and J. A. v. d. Laak, “Stain Specific Standardization of Whole-Slide Histopathological Images,” *IEEE Transactions on Medical Imaging*, vol. 35, no. 2, pp. 404–415, Feb. 2016.
- [41] Z. Guo, L. Zhang, and D. Zhang, “Rotation invariant texture classification using LBP variance (LBPV) with global matching,” *Pattern Recognition*, vol. 43, no. 3, pp. 706–719, 2010.
- [42] M. Valkonen, K. Kartasalo, K. Liimatainen, M. Nykter, L. Latonen, and P. Ruusuvoori, “Metastasis detection from whole slide images using local features and random forests,” *Cytometry Part A*, vol. 91, no. 6, pp. 555–565, 2017.
- [43] A. Paul, D. P. Mukherjee, P. Das, A. Gangopadhyay, A. R. Chintha, and S. Kundu, “Improved random forest for classification,” *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 4012–4024, 2018.
- [44] A. Pimkin, G. Makarchuk, V. Kondratenko, M. Pisov, E. Krivov, and M. Belyaev, “Ensembling neural networks for digital pathology images classification and segmentation,” *Lecture Notes in Computer Science*, vol. 10882 LNCS, pp. 877–886, 2018.
- [45] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2006.
- [46] A. Damianou and N. Lawrence, “Deep Gaussian processes,” *Journal of Machine Learning Research*, vol. 31, pp. 207–215, 2013.
- [47] M. Opper and C. Archambeau, “The variational Gaussian approximation revisited,” *Neural Comput.*, vol. 21, no. 3, pp. 786–792, Mar. 2009.
- [48] H. Salimbeni and M. Deisenroth, “Doubly stochastic variational inference for deep Gaussian processes,” in *Advances in Neural Information Processing Systems*, 2017, pp. 4591–4602.
- [49] D. Tellez, G. Litjens, P. Bándi, W. Bulten, J.-M. Bokhorst, F. Ciompi, and J. van der Laak, “Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology,” *Medical Image Analysis*, vol. 58, p. 101544, 2019.
- [50] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *CoRR*, vol. abs/1409.1556, 2015.

CHAPTER 4

Dictionary Learning for Blind Color Deconvolution

4.1 JCR Publication Details

Authors: Fernando Pérez-Bueno, Juan G. Serra, Miguel Vega, Javier Mateos, Rafael Molina, Aggelos K. Katsaggelos

Title: Bayesian K-SVD for H&E Blind Color Deconvolution. Applications to Stain Normalization, Data Augmentation and Cancer Classification.

Reference: Computerized Medical Imaging and Graphics, Volume 97, 2022, 102048

Status: Published

DOI: <https://doi.org/10.1016/j.compmedimag.2022.102048>

Quality indices:

- Impact Factor (JCR 2021): 7.422
 - Rank: 14/136 (Q1) in Radiology, Nuclear medicine and Medical Imaging
 - Rank: 15/98 (Q1) in Engineering, Biomedical

- Journal Citation Indicator (JCR 2021): 1.40
 - Rank: 21/200 (Q1) in Radiology, Nuclear medicine and Medical Imaging
 - Rank: 17/115 (Q1) in Engineering, Biomedical

4.2 Main Contributions

- We propose the use of Bayesian K -SVD for the estimation of the color-vector matrix and deconvolution of histological images with fully automatic estimation of the model parameters.
- Two Bayesian inference approaches to K -SVD are presented: variational and empirical Bayes.
- We present a novel methodology for the use of Bayesian approaches in massive WSI images.
- We propose a combination of color augmentation and color normalization using the BCD results. With this combination, the data is normalized before being augmented, with the idea that future test data will be normalized before being fed to the CAD system.
- The proposed approach was successfully evaluated on stain separation on a multi tissue dataset. It was also applied to color normalization and CNN-based cancer classification on a multi-center dataset, where we include the use of color augmentation.

Bayesian K-SVD for H&E blind color deconvolution. Applications to stain normalization, data augmentation and cancer classification.

Fernando Pérez-Bueno^{a,*}, Juan G. Serra^b, Miguel Vega^c, Javier Mateos^a, Rafael Molina^a,
Aggelos K. Katsaggelos^b

^a*Dpto. Ciencias de la Computación e Inteligencia Artificial, Universidad de Granada, Spain*

^b*Dept. of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL, USA*

^c*Dpto. de Lenguajes y Sistemas Informáticos, Universidad de Granada, Spain*

Abstract

Stain variation between images is a main issue in the analysis of histological images. These color variations, produced by different staining protocols and scanners in each laboratory, hamper the performance of computer-aided diagnosis (CAD) systems that are usually unable to generalize to unseen color distributions. Blind color deconvolution techniques separate multi-stained images into single stained bands that can then be used to reduce the generalization error of CAD systems through stain color normalization and/or stain color augmentation. In this work, we present a Bayesian modeling and inference blind color deconvolution framework based on the K-Singular Value Decomposition algorithm. Two possible inference procedures, variational and empirical Bayes are presented. Both provide the automatic estimation of the stain color matrix, stain concentrations and all model parameters. The proposed framework is tested on stain separation, image normalization, stain color augmentation, and classification problems.

Keywords: Bayesian modelling, Histological images, Blind Color Deconvolution, Stain Normalization

1. Introduction

The analysis of Whole-Slide Images (WSI), i.e., digitalized histological slides of tissue sections, is a crucial step towards the development of Computer Aided Diagnosis (CAD) systems. The tissues in a WSI are stained with different dyes to make their structure visible under the microscope. Hematoxylin-Eosin (H&E) is the most common combination, highlighting cell nuclei in blue color and cytoplasm and connective tissue in pink, respectively

*Corresponding author

Email addresses: fpb@ugr.es (Fernando Pérez-Bueno), jgserra@northwestern.edu (Juan G. Serra), mvega@ugr.es (Miguel Vega), jmd@decsai.ugr.es (Javier Mateos), rms@decsai.ugr.es (Rafael Molina), aggk@eecs.northwestern.edu (Aggelos K. Katsaggelos)

(Fischer et al., 2008). However, the color distribution of H&E WSI is affected by the staining and scanning procedures (Tosta et al., 2019b), resulting in inter- and intra- laboratory color variations. These variations hamper the performance of CAD systems, which are usually unable to generalize to unseen color distributions. The different approaches proposed to minimize the influence of color variation on CAD systems can be categorized into three groups: Blind Color Deconvolution (BCD), Color Normalization (CN) and Stain Color Augmentation (SCA). Let us review the most important contributions in each of these groups.

1.1. Blind Color Deconvolution

BCD techniques deal with color variation by estimating the image specific stain color-vectors and stain concentrations. The pioneer approach by Ruifrok and Johnston (2001) experimentally obtained a standard color vector matrix that is still used today. More recent methods tackle inter-slide variations by using different techniques. The use of Non-Negative Matrix Factorization (NMF) was proposed by Rabinovich et al. (2004), (Vahadane et al., 2016; Xu et al., 2015) added regularization and sparsity terms which encapsulate the assumption that a type of stain is only bound to certain structures. In Tosta et al. (2019a) the sparsity parameter was estimated using a fuzzy set method. Independent Component Analysis (ICA) was utilized in Trahearn et al. (2015) and extended in (Alsubaie et al., 2016, 2017) by applying ICA in the wavelet domain where the independence condition among sources is relaxed. The use of Singular Value Decomposition (SVD) was proposed in Macenko et al. (2009) to separate H&E channels. In McCann et al. (2014), the authors take into account the interaction between dyes. The method in Macenko et al. (2009) was revised in Astola (2016), where the author states that better results are obtained by applying it in the linearly inverted RGB-space instead of the (logarithmically inverted) absorbency space. Clustering was utilized in Gavrilovic et al. (2013) using the Maxwellian chromacity plane to obtain the stain vectors. Vicory et al. (2015) used K-means and a prior on the stain vectors to prevent misclustering when the amount of each stain is not balanced. In Khan et al. (2014), images are segmented into background and pixels belonging to each stain using supervised relevant vector machines. The color-vector for each stain is then defined as the mean of the pixels in each class. The work in Zheng et al. (2019) includes the deconvolution by Ruifrok as starting point and optimizes the color-vector and concentration values using a prior knowledge based objective function. Recently, a three-step method using Gabor kernels, structure segmentation and a final deconvolution step has been presented in Salvi et al. (2020).

Several Bayesian approaches have already been presented. In Hidalgo-Gavira et al. (2018), a similarity prior on the color-vectors as well as a smoothness Simultaneous Autorregressive (SAR) prior model on each stain concentration were used. This work was extended with the use of a TV prior in Pérez-Bueno et al. (2020) and with the use of sparse general Super Gaussian priors on the high-pass filtered concentrations in Pérez-Bueno et al. (2021).

Despite the Deep Learning popularity, few works have used it for BCD. Based on Macenko et al. (2009), the work in Duggal et al. (2017) implements a stain deconvolution layer for CNNs to provide a stain separated input to CNN-classifiers. Similarly, Zheng et al. (2021) use a Capsule Network that produces multiple stain separation candidates using 1 by 1 convolution operators and finally assembles the output based on a sparse constraint.

All BCD techniques have in common that they separate color from structural information, offering a strong control on the information preprocessing. They can preserve the tissue

structure and lead to high fidelity to the original images. Often, BCD methods are presented as CN methods since the obtained results can be used for CN (by normalizing each stain separately), but this is only one of the possible solutions BCD offers to deal with color variation.

1.2. Color Normalization

CN aims to reduce the stain variation by matching the color in the images to a selected template or reference. Direct CN methods do not necessarily estimate the concentrations and color vector matrix as BCD methods do. In [Tosta et al. \(2019b\)](#), they are classified into histogram matching, color transfer, and spectral matching. The first one adjusts image colors using histogram information ([Reinhard et al., 2001](#)), a common solution for general images. However, this is not appropriate for histological images as it ignores the local information and the unequally distribution of the stains. This work was further developed with three fuzzy normalization steps and adapted to histopathological images in [Vijh et al. \(2021\)](#). Color transfer methods assume that stain concentration is closely related to tissue structure and usually include region or dye segmentation. The latest color transfer methods, based on deep generative models ([Janowczyk et al., 2017](#); [Zanjani et al., 2018](#); [Bentaieb and Hamarneh, 2018](#)), perform CN without a previous color deconvolution by formulating the problem as a style transfer task where the style is the color distribution of a selected laboratory. Recently, other popular CNN architectures have been adapted to CN problems, such as Pix2pix ([Salehi and Chalechale, 2020](#)), disentangled representations ([Xiang et al., 2020](#)), CycleGAN ([Runz et al., 2021](#)) or Invertible Neural Networks ([Lan et al., 2021](#)). Since they require large datasets to train the networks that transform to an specific stain distribution, usually, they cannot handle intra-laboratory variations. Spectral matching methods typically perform BCD as a first step to CN. In [Tosta et al. \(2019b\)](#), most of the BCD methods mentioned in the previous section are reviewed as CN methods. The normalization is usually performed by replacing the stain color vectors obtained using BCD by the reference color vectors, often obtained from a template image ([Vahadane et al., 2016](#); [Vicory et al., 2015](#); [Zheng et al., 2019](#)). Different approaches are used to adjust the concentration intensity of both source and target images. In [Macenko et al. \(2009\)](#) each concentration intensity is scaled by using the 99th percentile to compute a robust estimation of the maximum. In [Vicory et al. \(2015\)](#) the median of the concentrations is used while in [Zheng et al. \(2019\)](#) the parameters normalizing the intensities are estimated jointly with the stain color vectors. Recently, [Hoque et al. \(2021\)](#) presented a multiscale Retinex model, that estimates and corrects the reflectance and illumination map for pixels of both stains separately.

1.3. Stain Color Augmentation

Data augmentation is a popular solution to reduce generalization error on CNN-based classifiers ([Zheng et al., 2021](#)). In contrast to BCD and CN, which aim to avoid the unseen stain distribution by eliminating the color variation, the augmentation approach aims to simulate unseen data by producing realistic variations of the available data. Although for histological images, morphological, generative ([Wei et al., 2020](#); [Zhu et al., 2017](#)), and color augmentation techniques can be used ([Tellez et al., 2019](#); [Mpinda Ataky et al., 2020](#)), in this study, we will focus on the latter to study the effect of color augmentation on classification in comparison to BCD and CN techniques. Color augmentation techniques do not modify the

image morphological features and only generate color variations. In [Liu et al. \(2017\)](#) common computer vision perturbations of brightness, contrast and hue are used. Furthermore, an specific histological stain color augmentation (SCA) technique was recently proposed in [Tellez et al. \(2018\)](#) where the method in [Ruifrok and Johnston \(2001\)](#) is applied to obtain the H&E concentration and variations of the observed data are created. In [Tellez et al. \(2019\)](#), several SCA and CN methods were evaluated on classification tasks with CNN. Additionally, a new CNN based CN method is proposed which is trained on SCA data.

1.4. Contributions

In the recent years, the field of BCD has received few contributions as CN approaches using Deep Learning usually avoid this step. However, BCD has some advantages for histological image analysis that should not be ignored. Its structure preserving properties, interpretability by doctors, and potential for classification purposes make this a field of interest for new works. The use of Bayesian models for BCD has been hardly explored and previous contributions are dependant on a similarity prior on the color-vectors. The choice of a reference color-vector matrix used for that prior, becomes a problem when working with images from different laboratories. Finally, BCD is required for the recently proposed SCA, which has been only compared to CN in [Tellez et al. \(2019\)](#). SCA and BCD have never been directly compared. For those reasons, in this work we propose a novel Bayesian K-SVD approach to perform BCD of histological images. K-SVD ([Aharon et al., 2006](#)) is a popular greedy algorithm for dictionary learning and sparse representation of signals. In BCD of histological images, the dictionary and the sparse representation will be the stain color vectors and the stain concentrations, respectively ([Vahadane et al., 2016](#)). However, K-SVD has two mayor drawbacks that need to be addressed for its use in BCD, the lack of uncertainty in the estimation procedure and the need to know in advance the number of non-zero components in the signal. The Bayesian K-SVD model ([Serra et al., 2017](#)) we adapt in this paper to BCD tackles these problems allowing its use for BCD of histological images. Using the obtained stain concentrations and color-vectors, our method can be utilized for CN and SCA. Our contributions are summarized as follows:

- Proposal of a new BCD framework that is able to preserve histological structures, with two possible inference approaches: variational and empirical Bayes.
- Unsupervised estimation of the stain concentrations and color properties.
- Automatic estimation of all model parameters.
- Stain specific data augmentation using the stain concentrations and color-vector matrix.
- Performance evaluation on large histological datasets with intra- an inter-laboratory variations.
- Analysis of classification performance when using normalized images or stain concentrations.

The proposed method is tested on a set of experiments designed to cover the main tasks of digital histopathology.

The paper is organized as follows. In section 2 we present the mathematical formulation of the BCD problem. In section 3, this problem is cast into the hierarchical Bayesian paradigm and inference is carried out to estimate the stain concentrations and color-vectors as well as all model parameters. Using Empirical Bayes, in section 4 we modify the inference already presented in section 3 to increase the sparsity of the obtained solution. Section 5 adapts the proposed methods to its application in massive WSI. Section 6 describes the utilized images and methods. The effectiveness of the proposed framework is experimentally assessed in Section 7, where the proposed methods are compared to classical and state-of-the-art alternatives. Finally, section 8 concludes the paper.

2. Problem Formulation

Each WSI is stored as an $M \times N \times 3$ RGB intensity image which is rearranged into the matrix $\mathbf{I} \in \mathbb{R}^{3 \times Q}$, $Q = MN$, where each value $i_{cq} \in \mathbf{I}$ represents the transmitted light across the slide for pixel q and channel c . Diagnosis protocols use the contribution of each stain to this value, that is, its absorbency or *optical density* (OD). The OD corresponding to intensity i_{cq} , $y_{cq} \in \mathbf{Y}$, is defined as $y_{cq} = -\log_{10}(i_{cq}/i_{cq}^0)$, where i_{cq}^0 denotes the incident light. The monochromatic Beer-Lambert law establishes that a slide \mathbf{Y} stained with N_s stains follows the equation

$$\mathbf{Y} = \mathbf{MC} + \mathbf{N}, \quad (1)$$

where $\mathbf{M} = [\mathbf{m}_1, \dots, \mathbf{m}_{N_s}] \in \mathbb{R}^{3 \times N_s}$ is the normalized stains' specific color-vector matrix; $\mathbf{C} \in \mathbb{R}^{N_s \times Q}$ is the stain concentration matrix, its q th column, $\mathbf{c}_q = [c_{1,q}, \dots, c_{N_s,q}]^T$, represents the contribution of each stain to the q th pixel value in \mathbf{Y} ; and, finally, $\mathbf{N} \in \mathbb{R}^{3 \times Q}$ is a random matrix with i.i.d. zero-mean Gaussian components with unknown variance β^{-1} . Each column, \mathbf{m}_s , in matrix \mathbf{M} is assumed to be a unit ℓ^2 -norm stain color-vector containing the relative RGB color composition of the corresponding stain in the OD space.

Notice that each column \mathbf{y}_q can be represented as a linear combination of the color vectors weighted by the corresponding concentrations, that is, $\mathbf{y}_q = \sum_{s=1}^{N_s} c_{sq} \mathbf{m}_s + \mathbf{n}_q$. Hematoxylin is a basic stain that dyes basophilic structures, namely nuclei, while eosin is an acidic stain that fixes to cytoplasm and other structures, usually referred to as eosinophilic. Although the actual color of biological structures will be influenced by both stains, they will present structure-specific color properties (Vahadane et al., 2016) (effective stains) that are the basis of differential staining. Therefore, we can assume that most pixels in the image are stained by a single effective stain (Vahadane et al., 2016), making our stain concentration matrix sparse, in other words, most of the weights, $c_{sq} \in \mathbf{c}_q$, in this linear combination are expected to be zero (or very small). We would like to find not only the sparse coefficients of these linear combinations, but at the same time, also estimate the color vectors \mathbf{m}_s which result in the best, most sparse, solution. This dual estimation can clearly be understood as a dictionary learning problem. Notice that we estimate the effective stains that allow to sparsely separate biological structures.

The original problem of finding an exactly sparse solution minimizing the number of non-zero elements in each \mathbf{c}_q (i.e., minimizing $\|\mathbf{c}_q\|_0, \forall q$)¹, is known to be NP-hard, see (Babacan

¹ $\|\cdot\|_0$ denotes the ℓ^0 -(pseudo)norm, which counts the number of non-zero elements in a vector.

et al., 2010), for example. The true solution can be approximated with greedy methods (e.g. the popular K-SVD Aharon et al. (2006) method). Alternatively, the sparsity constraint on the concentration vectors can be relaxed by using the ℓ^1 -norm instead. Formally, we can formulate this problem as

$$\begin{aligned} \min_{\mathbf{M}, \mathbf{C}} & \|\mathbf{Y} - \mathbf{MC}\|_{\text{F}}^2 \\ \text{s.t.} & \|\mathbf{c}_q\|_1 \leq T, \quad \forall q, \end{aligned} \quad (2)$$

where $\|\cdot\|_{\text{F}}$ and $\|\cdot\|_1$ denote the Frobenius and ℓ^1 -norms respectively, and T is nonnegative real parameter that determines the degree of regularization. The main advantage of this relaxation is that convex optimization techniques can be used to solve this problem (e.g., (Aharon et al., 2006; Mairal et al., 2009; Zhou et al., 2009)).

The novel Bayesian framework we propose solves the histological color deconvolution as an ℓ^1 dictionary learning problem, following the method introduced in Serra et al. (2017), automatically estimating the optimal color-vector matrix \mathbf{M} , the posterior distribution of \mathbf{C} considering the uncertainty of the coefficients, along with all model parameters. The next section gives detailed intuition on the modelling and inference of the proposed method, albeit not a full derivation. We encourage the interested reader to consult Appendix A for further explanation.

3. Bayesian Model and Inference

Our Bayesian model for solving the dictionary learning problem in (2) relies on defining suitable probability distributions on the observations \mathbf{Y} and on the set of unknowns $\{\beta, \mathbf{M}, \mathbf{C}\}$. The observation model in (1) described above corresponds to the isometric Gaussian distribution on \mathbf{Y} given by

$$p(\mathbf{Y}|\beta, \mathbf{M}, \mathbf{C}) \propto \beta^{\frac{3Q}{2}} \exp\left(-\frac{\beta}{2} \|\mathbf{Y} - \mathbf{MC}\|_{\text{F}}^2\right). \quad (3)$$

The obvious choice for the noise precision, β , since it is a positive-valued variable, is a gamma distribution, thus, $p(\beta) = \Gamma(\beta|a^\beta, b^\beta)$ with $a^\beta, b^\beta > 0$. Our modelling for the stain vector matrix \mathbf{M} focuses on imposing unit norm for each column \mathbf{m}_s ; for this purpose we use a flat prior on the columns of \mathbf{M} such that $p(\mathbf{m}_s) = \text{const.}$, if $\|\mathbf{m}_s\| = 1$, 0 otherwise, and assume independent column vectors. Finally, notice that the sparsity constraint on the coefficient vectors \mathbf{c}_q in (2) is equivalent to imposing a zero-mean Laplace distribution with scale parameter $\lambda_q > 0$, $p(\mathbf{c}_q) \propto \exp(-\sqrt{\lambda_q} \|\mathbf{c}_q\|_1)$. The Laplace prior is more peaked than the normal distribution with longer tails, which is also interesting for structure preserving (Babacan et al., 2012). Unfortunately, the non-conjugacy of this distribution with the likelihood in (3) makes inference intractable. We circumvent this problem by using a two-tiered hierarchical prior on \mathbf{c}_q instead. First, we impose a zero-mean normal distribution with diagonal covariance matrix $\mathbf{\Gamma}_q = \text{diag}(\gamma_q)$, i.e., $\mathbf{c}_q \sim \mathcal{N}(\mathbf{c}_q|\mathbf{0}_{N_s}, \mathbf{\Gamma}_q)$. And secondly, we use the Gamma hyperpriors on the positive-valued γ_{sq} given by $\gamma_{sq} \sim \Gamma(1, \lambda_q/2)$ and assume independence yet again, so that $p(\gamma_q) = \prod_s p(\gamma_{sq})$. This two-tier prior can be further expanded with a third prior on the scale parameters λ_q , however, although it gives more flexibility to the model, in practice does not turn into noticeable estimation improvement. The idea behind

this hierarchical prior is to sample the covariance matrix of the normal distribution $p(\mathbf{c}_q|\gamma_q)$ from a Gamma distribution with shape 1 and variable scale (an exponential distribution). The samples produced using this scheme follow a Laplace distribution, which can be shown by marginalization of γ_q , i.e., $\int p(\mathbf{c}_q|\gamma_q)p(\gamma_q|\lambda_q)d\gamma_q \sim \text{Laplace}(\mathbf{c}_q|\lambda_q)$.

In order to estimate the whole set of unknowns, Θ , that includes the noise precision, the color-vector matrix and the coefficient matrix along with the corresponding hyperparameters, $\Theta = \{\beta, \mathbf{M}, \mathbf{C}, \Gamma, \boldsymbol{\lambda}\}$ with $\Gamma = \{\gamma_q\}_{q=1}^Q$ and $\boldsymbol{\lambda} = \{\lambda_q\}_{q=1}^Q$, we make use of Bayesian inference. The exact calculation of the true posterior $p(\Theta|\mathbf{Y}) = p(\mathbf{Y}, \Theta)/p(\mathbf{Y})$, with joint distribution

$$p(\mathbf{Y}, \Theta) = p(\mathbf{Y}|\beta, \mathbf{M}, \mathbf{C})p(\beta)p(\mathbf{M})p(\mathbf{C}|\Gamma)p(\Gamma|\boldsymbol{\lambda}), \quad (4)$$

cannot be done analytically since it requires the marginal $p(\mathbf{Y}) = \int p(\mathbf{Y}, \Theta)d\Theta$ which is intractable. We use variational inference to approximate the true posterior, which requires the assumption of simplifications on the form of the posterior. These simplifications should render the inference tractable, while at the same time ensure that the model is flexible enough to closely approximate the true posterior distribution. Concretely, we will assume that our approximate posterior $q(\Theta)$ factorizes as $q(\Theta) = q(\beta)q(\Gamma)q(\boldsymbol{\lambda})q(\mathbf{C})\prod_s q(\mathbf{m}_s)$, which is referred to as mean-field factorization in the literature, see (Bishop, 2006). Notice, however, that we do not make any assumption on the individual distributions of each random variable; this will be determined by the inference procedure. The optimal solution is found by minimizing the Kullback-Leibler divergence between the approximate $q(\Theta)$ and true posterior $p(\Theta|\mathbf{Y})$. This optimization has a well-known optimum given by

$$\log q(\boldsymbol{\theta}_i) = \langle \log p(\mathbf{Y}, \Theta) \rangle_{\Theta \setminus \boldsymbol{\theta}_i} + \text{const.}, \quad (5)$$

where $\langle \cdot \rangle_{\Theta \setminus \boldsymbol{\theta}_i}$ denotes the expectation taken w.r.t. all approximating variables $\boldsymbol{\theta}_j \in \Theta$, with $j \neq i$. In the case of degenerate distributions $q(\boldsymbol{\theta}_i)$, this calculation simplifies to finding the maximum w.r.t. $\boldsymbol{\theta}_i$ of the same expectation as in (5). Notice that this implies that we will not find a proper distribution for $\boldsymbol{\theta}_i$, but only its mean with zero variance. We will assume degenerate posterior distributions on \mathbf{M} , Γ and $\boldsymbol{\lambda}$, which will simplify the calculation of the expectations w.r.t. these random variables since $\langle f(\boldsymbol{\theta}_i) \rangle_{\boldsymbol{\theta}_i} = f(\hat{\boldsymbol{\theta}}_i)$, where $\hat{\boldsymbol{\theta}}_i := \langle \boldsymbol{\theta}_i \rangle_{\boldsymbol{\theta}_i}$. In contrast, we will obtain full distributions for the noise precision and the stain concentration sparse vectors.

After careful derivation using (5) on \mathbf{C} , (Serra et al., 2017) for the details, we find that each \mathbf{c}_q follows a Gaussian distribution with mean and covariance matrix given by

$$\hat{\mathbf{c}}_q = \hat{\beta}\boldsymbol{\Sigma}_{\mathbf{c}_q}\hat{\mathbf{M}}^T\mathbf{y}_q, \quad (6)$$

$$\boldsymbol{\Sigma}_{\mathbf{c}_q} = \left(\hat{\beta}\hat{\mathbf{M}}^T\hat{\mathbf{M}} + \hat{\Gamma}_q^{-1} \right)^{-1}. \quad (7)$$

We can now find the optimal estimations for the associated hyperparameters of the hierarchical prior γ_q and λ_q by maximization of the right-hand side of (5) as described above, obtaining

$$\hat{\gamma}_{sq} = -\frac{1}{2\hat{\lambda}_q} + \sqrt{\frac{1}{4\hat{\lambda}_q^2} + \frac{\hat{c}_{sq}^2 + \boldsymbol{\Sigma}_{\mathbf{c}_q}(s, s)}{\hat{\lambda}_q}}, \quad (8)$$

$$\hat{\lambda}_q = \frac{2N_s}{\sum_{s=1}^{N_s} \hat{\gamma}_{sq}}. \quad (9)$$

It is interesting to study the effect of the uncertainty on the estimates of c_{sq} given by $\Sigma_{\mathbf{c}_q}(s, s)$ in (8). As our uncertainty in the estimation grows, so will γ_{sq} , which models the variance of c_{sq} , and, therefore, it will increase the uncertainty on this parameter.

The optimal $\mathbf{m}_s \in \mathbf{M}$ can be found assuming column independence and degenerate approximate posteriors $q(\mathbf{m}_s)$ on a point on the unit sphere, $\|\mathbf{m}_s\| = 1$. Following the inference procedure described above, we have

$$\hat{\mathbf{m}}_s \propto \left[\mathbf{Y} - \sum_{i \neq s} \hat{\mathbf{m}}_i \hat{\mathbf{c}}_{i,:} \right] \hat{\mathbf{c}}_{s,:}^T - \sum_{i \neq s} \sum_q \sigma_{isq} \hat{\mathbf{m}}_i, \quad (10)$$

where σ_{isq} denotes $\Sigma_{\mathbf{c}_q}(i, s)$ and defines the influence of the uncertainty of the estimation of the coefficient vectors. The actual estimate of \mathbf{m}_s is obtained by normalizing (10).

Finally, applying (5) for β results in a gamma-distributed posterior with mean given by

$$\hat{\beta} = \frac{3Q + 2a^\beta}{\|\mathbf{Y} - \hat{\mathbf{M}}\hat{\mathbf{C}}\|_{\mathbb{F}}^2 + \sum_{q=1}^Q \text{tr}(\hat{\mathbf{M}}^T \hat{\mathbf{M}} \Sigma_{\mathbf{c}_q}) + 2b^\beta}. \quad (11)$$

Once more, note here how the uncertainty in the estimation of the coefficient vectors \mathbf{c}_q given by $\Sigma_{\mathbf{c}_q}$ impacts the estimation of the noise precision, resulting in lower precision (higher variance) when this uncertainties grow.

The procedure to obtain the estimated $\hat{\mathbf{M}}$ and $\hat{\mathbf{C}}$ using the above presented modelling and inference is summarized in **Algorithm 1**.

Algorithm 1 Pseudocode for BKSVD BCD algorithm

Input: Observed image \mathbf{I} , initial normalized $\underline{\mathbf{M}}$, no. stains N_s .

Output: Estimated stain color-vector matrix, $\hat{\mathbf{M}}$, and concentrations, $\hat{\mathbf{C}}$,

- 1: Obtain the OD image \mathbf{Y} from \mathbf{I} and set $\hat{\mathbf{m}}_s = \underline{\mathbf{m}}_s$, $\Sigma_{\mathbf{c}_s} = \mathbf{0}$, $\hat{\mathbf{C}} = \underline{\mathbf{M}}^+ \mathbf{Y}$, with $\underline{\mathbf{M}}^+$ the Moore-Penrose pseudo-inverse of $\underline{\mathbf{M}}$ and $\Gamma = \mathbf{1}$
 - 2: **while** $\hat{\mathbf{C}}$ has not converged **do**
 - 3: **for** q in $1, \dots, Q$ **do**
 - 4: Update $\hat{\lambda}_q$ using (9)
 - 5: Update $\hat{\gamma}_{sq}$ using (8), for all s in $1, \dots, N_s$
 - 6: Update $\Sigma_{\mathbf{c}_q}$ and $\hat{\mathbf{c}}_q$ using (7) and (6), respectively
 - 7: **end for**
 - 8: **for** s in $1, \dots, N_s$ **do**
 - 9: Update $\hat{\mathbf{m}}_s$ using (10)
 - 10: **end for**
 - 11: Update $\hat{\beta}$ using (11)
 - 12: **end while**
 - 13: **return** $\hat{\mathbf{M}}$ and $\hat{\mathbf{C}}$
-

4. Sequential Inference for Sparse Bayesian Models

The previous section introduced a mathematically sound inference procedure. However, the sparse values in the columns \mathbf{c}_q are not guaranteed to be zero. Since most of the pixels

in the image should be stained only by one stain, higher sparsity is desired. To increase the sparsity of the obtained solution we use Empirical Bayes (Tipping and Faul, 2003; Babacan et al., 2010; Serra et al., 2017) to obtain a new inference procedure. This approach was first presented in (Tipping and Faul, 2003) for Sparse Bayesian Learning (SBL) and later in (Babacan et al., 2010) and (Serra et al., 2017) for recovery of sparse signals. In this paper, we introduce the necessary adaptation for the application to histological blind color deconvolution.

In particular, for each \mathbf{c}_q , we use a constructive approach for identifying the locations where it takes non-zero values, i.e., its support. At these non-zero locations, we use Maximum *A Posteriori* (MAP) estimation to obtain the values of the hyperparameters. Therefore, sparsity makes the effective problem dimensions to be drastically reduced. The estimated values of the columns \mathbf{c}_q in its support are obtained using (6).

The main idea behind this inference scheme consists on replacing the variational inference of hyperparameters γ_q with direct maximization of the (log) marginal likelihood

$$\mathcal{L}(\gamma_q) = \log \left[p(\gamma_q | \hat{\lambda}_q) \int p(\mathbf{y}_q | \mathbf{c}_q, \hat{\beta}) p(\mathbf{c}_q | \gamma_q) d\mathbf{c}_q \right], \quad (12)$$

where $p(\mathbf{y}_q | \mathbf{c}_q, \hat{\beta}) = \mathcal{N}(\mathbf{y}_q | \hat{\mathbf{M}}\mathbf{c}_q, \hat{\beta}^{-1}\mathbf{I})$, following the observation model, and $\hat{\mathbf{M}}$, $\hat{\beta}$ and $\hat{\lambda}_q$ are estimated as shown in Sec. 3. The marginal likelihood $\mathcal{L}(\gamma_q)$ has interesting properties that allow for a highly efficient maximization thereof. Concretely, its functional form allows us to separate the contribution of a single γ_{sq} so that $\mathcal{L}(\gamma_q) = \mathcal{L}(\{\gamma_{iq}\}_{i \neq s}) + l(\gamma_{sq})$. A closed form solution of the maximization of $\mathcal{L}(\gamma_q)$, when only its s -th component is changed, can be found by holding the other hyperparameters fixed, taking its derivative with respect to γ_{sq} and setting it equal to zero. Note that this derivative will be different from zero only for $l(\gamma_{sq})$. Analysis of $l(\gamma_{sq})$ (see Appendix A) shows that the marginal likelihood has a unique maximum w.r.t. γ_{sq} and allows us to efficiently estimate the increase in log-likelihood that changing this parameter will introduce.

The Empirical Bayesian K-SVD (EBKSVD) in **Algorithm 2** is initialized by including only one color vector, the one that produces the highest increase in log-likelihood, and the corresponding γ_{sq} ; the remaining $\{\gamma_{iq}\}_{i \neq s}$ are set to 0. At each iteration of the algorithm we will be able to add a new color vector, and its corresponding γ_{sq} , to our current model if the previous value of the γ_{sq} that produces the greatest increase of $\mathcal{L}(\gamma_q)$ was zero; we will remove the element from the model if the optimal value of γ_{sq} is 0; or, finally, reestimate (update) γ_{sq} if \mathbf{m}_s was already part of the model. In all three cases we are able to make incremental changes to the model structure while guaranteeing an increase of log-likelihood. Finally, the updates for \mathbf{c}_q and Σ_q will be done using only the γ_{sq} included in the model, which will guarantee the sparsity of this inference method. See details in Appendix A.

To conclude this section, let us briefly compare the variational and empirical approaches. As previously discussed, the variational inference in section 3 achieves a softer sparsity, where the concentrations will include residual non-zero values. The combination of residual and non-sparse values might influence the final estimation of $\hat{\mathbf{M}}$. The empirical approach reduces this effect by calculating only the values where \mathbf{c}_q takes non-zero values. Empirical Bayes is usually used to reduce the computational burden of Bayesian methods, as the calculation of the covariance matrix in (7) require to calculate the inverse matrix at each step and might be expensive for big matrices. Note that this is not the case for BCD of histological images,

Algorithm 2 Pseudocode for Empirical BKSVD BCD algorithm

Input: Observed image \mathbf{I} , initial normalized $\underline{\mathbf{M}}$, no. stains N_s .

Output: Estimated stain color-vector matrix, $\hat{\mathbf{M}}$, and concentrations, $\hat{\mathbf{C}}$,

- 1: Obtain the OD image \mathbf{Y} from \mathbf{I} and set $\hat{\mathbf{m}}_s = \underline{\mathbf{m}}_s$, $\underline{\Sigma}_{\mathbf{c}_s} = \mathbf{0}$, $\hat{\mathbf{C}} = \underline{\mathbf{M}}^+ \mathbf{Y}$, with $\underline{\mathbf{M}}^+$ the Moore-Penrose pseudo-inverse of $\underline{\mathbf{M}}$, $\underline{\Gamma} = \mathbf{0}$, and $\underline{\lambda} = \mathbf{0}$
 - 2: **while** $\hat{\mathbf{C}}$ has not converged **do**
 - 3: **for** q in $1, \dots, Q$ **do**
 - 4: Choose a $s \in \{1, \dots, n_s\}$ (or equivalently choose a γ_{sq})
 - 5: Find the optimal value of $\hat{\gamma}_{sq}$ using (A.9)
 - 6: Update $\underline{\Sigma}_{\mathbf{c}_q}$ and $\hat{\mathbf{c}}_q$ using (7) and (6), respectively
 - 7: Update g_{sq} and h_{sq} using (A.12) and (A.14), respectively, for all s in $1, \dots, N_s$
 - 8: Update $\hat{\lambda}_q$ using (9)
 - 9: **end for**
 - 10: **for** s in $1, \dots, N_s$ **do**
 - 11: Update $\hat{\mathbf{m}}_s$ using (10)
 - 12: **end for**
 - 13: Update $\hat{\beta}$ using (11)
 - 14: **end while**
 - 15: **return** $\hat{\mathbf{M}}$ and $\hat{\mathbf{C}}$
-

where the number of stains is usually $N_s = 2$ and the inversion of 2×2 matrices is not costly. Although the computational saving is reduced, the additional sparsity induced by the empirical method is useful in the estimation of \mathbf{M} , as we will make clear in the following sections.

5. Application of Bayesian K-SVD for WSI analysis

Bayesian methods are usually computationally expensive as they require to take into account the uncertainties of the coefficients at each element in the image. While previous applications as denoising or inpainting (Serra et al., 2017) were carried out on small 256×256 grayscale images (64Kpixels), its application to blind color deconvolution problem is hindered by the massive size of WSI images. WSIs are RGB images in the Gigapixels order which makes their processing challenging. Therefore, it is extremely necessary to introduce additional adaptations that make the BKSVD and EBKSVD more suitable for WSI images.

First, during training and reconstruction of the histological images, the highest computational cost is the computation of the sparse representation of the concentrations for each pixel. However, the reduced amount of stains suggests that it is not required to use all WSI pixels to learn $\hat{\mathbf{M}}$. We here reformulate (1) as

$$\mathbf{Y}_B = \mathbf{M}\mathbf{C}_B + \mathbf{N}_B, \quad (13)$$

where \mathbf{Y}_B is a representative subset of the pixels in \mathbf{Y} and \mathbf{C}_B its associated concentration matrix. To find the representative set of pixels, we first look at those that can be discarded. Large background areas are typically removed upon patching for most applications. However background pixels can also appear on lumens or tissue borders. Since those low stained

pixels do not provide information on the stain’s color, following [Vahadane et al. \(2016\)](#), we can remove them for the estimation of \mathbf{M} . The optical density of those pixels is close to zero making it easy to filter them. The removal of low stained pixels accelerates the procedure of estimating $\hat{\mathbf{M}}$ and eliminates the influence of background pixels.

Despite considering only tissue pixels, usually there are still too many pixels for practical application of the algorithms. WSI images often include several resolutions. While using the smaller images obtained at lower magnifications could be tempting, we should avoid them in the estimation of \mathbf{M} . Pixels values at lower resolutions, when interpolated linearly, are a weighted average of a set of pixels at a higher resolution. Note, however, that this average takes place in the RGB space. Then, obtaining the OD image requires the use of the non-linear logarithmic transformation. As the logarithm is a concave function, for a single pixel at a lower resolution, we have for non-negative weights $\{\tau_q\}$ in a neighborhood that add up to one

$$\mathbf{y} = -\log\left(\sum_q \tau_q \frac{\mathbf{i}_q}{i^0}\right) \leq -\sum_q \tau_q \log\left(\frac{\mathbf{i}_q}{i^0}\right) = -\sum_q \tau_q (\mathbf{M}\mathbf{c}_q) \quad (14)$$

where $\mathbf{i}_q = [\mathbf{i}_1, \dots, \mathbf{i}_N]$ are the high resolution pixels contributing to the averaged pixel. Although the linearity in the RGB space is not preserved in OD, we can expect the assumption that most pixels are stained by a single stain to be less satisfied as resolution decreases. Therefore, as the proposed methods are based on the sparsity assumption, it is preferred to extract a subset of pixels from the WSI at the higher magnification available, typically $40\times$.

Therefore we need find another method to reduce the amount of pixels to be considered. Patching is the most common way of dealing with the massive size of WSIs during preprocessing or classification. This approach allows to take into account local tissue structures, which are important for WSI interpretation. However, it is not a suitable solution for obtaining $\hat{\mathbf{M}}$ as local tissue structures may not correctly represent both stains. Note that the proposed framework assumes that each pixel stain’s concentrations are independent, thus eliminating spatial constraints. This modelling allows us to select individual pixels in the image, independently of their neighbours. Therefore, we can obtain a representative subset \mathbf{Y}_B within the image using an uniform random sampling of the stained pixels. This allows, on the one side, to accurately sample the whole WSI the image and, on the other, dramatically to reduce the number of pixels used to estimate the stain-color matrix \mathbf{M} .

For a given subset \mathbf{Y}_B , the color vector matrix $\hat{\mathbf{M}}$ can be estimated using BKSVD or EBKSVD in [Alg. 1](#) and [Alg. 2](#), respectively. To avoid overfitting to a given subset \mathbf{Y}_B , once the chosen method converges, a new batch of pixels \mathbf{Y}_B is selected and the estimation procedure is repeated until the matrix $\hat{\mathbf{M}}$ converges. Notice that we do not use complete epochs as our objective is to ensure that the obtained $\hat{\mathbf{M}}$ faithfully represents the colors in the WSI without using all pixels in the image.

Once the color vectors of the image are estimated using \mathbf{Y}_B , we still need to obtain the stain concentrations \mathbf{C} for the whole image. We could consider to execute [Alg. 1](#) or [Alg. 2](#) for the whole image keeping $\hat{\mathbf{M}}$ fixed. However, this still requires to iterate in order to estimate the model parameters and concentrations at each pixel, which is time prohibitive for the whole image. Then, assuming that $\hat{\mathbf{M}}$ is an accurate estimation of \mathbf{M} , the final values of the concentrations, $\hat{\mathbf{C}}$, for the whole image will be computed as $\hat{\mathbf{C}} = \hat{\mathbf{M}}^+ \mathbf{Y}$ ([Ruifrok and Johnston, 2001](#); [Alsubaie et al., 2017](#)), with $\hat{\mathbf{M}}^+$ the Moore-Penrose pseudo-inverse of $\hat{\mathbf{M}}$.

Note that, for a fixed $\hat{\mathbf{M}}$, this is also the minimum squared error estimator of \mathbf{C} from (1).

Finally, the described multibatch procedure is summarized in **Algorithm 3**.

Algorithm 3 Multibatch Bayesian KSVD

Input: Observed image \mathbf{I} , initial normalized $\underline{\mathbf{M}}$, no. stains N_s , batch size B .

Output: Estimated stain color-vector matrix, $\hat{\mathbf{M}}$, and concentrations, $\hat{\mathbf{C}}$,

- 1: Obtain the OD image \mathbf{Y} from \mathbf{I}
 - 2: Remove low stained pixels from \mathbf{Y}
 - 3: **while** $\hat{\mathbf{M}}$ has not converged **do**
 - 4: Sample a batch \mathbf{Y}_B of B stained pixels from \mathbf{Y} .
 - 5: Estimate $\hat{\mathbf{M}}$ using BKSVD or EBKSVD
 - 6: **end while**
 - 7: **return** $\hat{\mathbf{M}}$ and $\hat{\mathbf{C}} = \hat{\mathbf{M}}^+\mathbf{Y}$
-

6. Materials and methods

To assess its quality, the proposed BKSVD and EBKSVD were compared to the following methods frequently used in the literature: the classical non-blind CD method by Ruifrok and Johnston (2001) and the BCD methods by Macenko et al. (2009), Vahadane et al. (2016), Alsubaie et al. (2017), Hidalgo-Gavira et al. (2020), Pérez-Bueno et al. (2020), and Zheng et al. (2019). They will be denoted by RUI, MAC, VAH, ALS, HID, PER, and ZHE, respectively. All experiments in the following sections were conducted using the multibatch Bayesian K-SVD² in Alg. 3 with $N_s = 2$. As initial color-vector matrix, we used the standard H&E vectors proposed by Ruifrok and Johnston (2001). The proposed method was run until the criterion $\|\mathbf{M}^{(n)} - \mathbf{M}^{(n-1)}\|_F^2 < 5 \times 10^{-3}$ was met. Algorithms 1 and 2 were run until the criterion $\|\langle \mathbf{c}_s \rangle^{(n)} - \langle \mathbf{c}_s \rangle^{(n-1)}\|^2 / \|\langle \mathbf{c}_s \rangle^{(n)}\|^2 < 10^{-4}$ was met by both stains. All model parameters are automatically estimated. Using the obtained $\hat{\mathbf{M}}$ and $\hat{\mathbf{C}}$ it is possible to perform CN and SCA. Further details are provided in the following experimental section.

To test the performance and robustness of our algorithm in different scenarios related to digital histopathology (i.e., stain separation quality, color normalization, and stain color augmentation for cancer classification), we have selected data containing a variety of histopathological images from several types of tissue and laboratories. In this section we describe the details of the databases used in this paper.

6.1. Warwick Stain Separation Benchmark (WSSB)

WSSB dataset (Alsubaie et al., 2017) contains 24 H&E stained images of different tissues (breast, colon, and lung) from different laboratories which have been captured with different microscopes. For each image, its ground truth stain color-vector matrix, \mathbf{M}_{GT} , was manually obtained by expert pathologists as follows. The experts selected a set of pixels for each stain, based on biological structures: nuclei for hematoxylin and cytoplasm for eosin. Then, the median value of each set of pixels with a single stain was used as a measure of the

²The code used in the experiments will be made available at <https://github.com/vipgugr> upon acceptance of the paper.

Table 1: CAMELYON17 dataset labeling structure

Subset	WSI total	Stage label			
		Negative	ITC	Micro	Macro
Whole training set	500	318	36	59	87
Annotated	50	0	16	17	17
Not annotated	450	318	20	42	70

corresponding stain color-vector. Ground truth concentrations were obtained in [Alsubaie et al. \(2017\)](#) from the ground-truth color-vector matrix as

$$\mathbf{C}_{GT} = \mathbf{M}_{GT}^+ \mathbf{Y}. \quad (15)$$

From those ground-truth concentrations and color-vectors, a separate RGB image for each stain is obtained. This database will be used for BCD evaluation.

6.2. CAMELYON17

This database is part of the CAMELYON17 challenge ([Bándi et al., 2019](#)) for breast cancer metastasis detection in the lymph node sections. We will use it in CN and classification experiments including the use of SCA.

CAMELYON17 contains a total of 1000 WSIs from 5 medical centers. Only the training set, which contains 500 WSIs, was used since the annotations for the test WSIs are not available yet. The dataset comprises 20 patients per center and 5 slides per patient. Cancer regions were annotated by pathologists only on 50 WSIs, but the stage label: negative, isolated tumor cells (ITC), micrometastasis (Micro), macrometastasis (Macro), is available for all the slides in the training set. See Table 1 for details.

Following [Zheng et al. \(2019\)](#) the experiments on this dataset were performed using non-overlapping 224×224 pixel patches, with at least a 70% of tissue, sampled from each WSI.

7. Experimental results

We have carried out a set of experiments to evaluate the performance of the proposed framework on the most common histological color deconvolution related tasks: stain separation, image normalization, and CNN-based classification, where we include the use of SCA.

First, we evaluate the influence of the pixel batch size on the proposed methods. Then we assess the quality of the concentration and color-vector matrices obtained by the BCD algorithms. In a third experiment, we analyze the quality of the CN obtained by the algorithms when the color-vectors are substituted by those of a reference-image, keeping the concentration values. Finally, the deconvolved, normalized, and SCA images are evaluated on a histological classification scenario.

7.1. Influence of the batch size in the color vector estimation

The use of pixel sampling introduced in Section 5 requires to assess the influence of the pixel batch size on the similarity of the obtained color vector matrix $\hat{\mathbf{M}}_P$ (obtained using P

Table 2: Mean values required to estimate $\hat{\mathbf{M}}_P$ and $\hat{\mathbf{M}}_{all}$ using Alg. 3

BKSVD	batch size in pixels										full images		
	50	100	300	500	1000	2000	4000	10^4	$2 \cdot 10^4$	10^5	$2 \cdot 10^5$	$4 \cdot 10^6$	$16 \cdot 10^6$
no. batches	9.8	10	7.3	7.2	5.3	4.7	4.2	4.1	4.3	4.3	1	1	1
no. total iter.	97.34	80.07	43.87	41.53	31.20	29.07	27.87	27.34	27.6	27.67	18	13	17
time/iter. (s)	0.09	0.11	0.14	0.15	0.17	0.20	0.24	0.34	0.52	1.88	2.78	34.08	268.80
total time	8.68	8.86	6.00	6.05	5.37	5.86	6.8	9.31	14.47	52.09	50.06	443.06	4569.60
EBKSVD	batch size in pixels										full images		
	50	100	300	500	1000	2000	4000	10^4	$2 \cdot 10^4$	10^5	$2 \cdot 10^5$	$4 \cdot 10^6$	$16 \cdot 10^6$
no. batches	9.3	8.7	7.8	8	5.3	5.5	5.4	5.2	4.8	3.7	1	1	1
no. total iter.	154.47	135.27	109.33	98.73	85.00	79.13	76.2	67.8	49.4	43.00	34	8	15
time/iter. (s)	0.10	0.13	0.17	0.21	0.27	0.43	0.77	2.02	3.99	19.75	38.23	431.01	2338.30
total time	15.95	18.03	18.65	20.33	22.83	33.91	58.78	136.99	197.27	849.37	1300.15	3448.11	35074.50

pixels) to the $\hat{\mathbf{M}}_{all}$ obtained using all non-white pixels and the execution time required for the estimation. Unfortunately, it is not possible to use complete WSIs in this experiment due to the computational burden, therefore we use three different images of typical sizes 500×500 , 2000×2000 and 4000×4000 pixels and batch sizes from 50 to $1.6 \cdot 10^7$ pixels. Algorithm 3, using both BKSVD and EBKSVD, was run 5 times for each different batch size up to $2 \cdot 10^4$ pixels and only once for bigger ones.

Table 2 summarizes the mean number of batches, iterations, time per iteration and total time required by Alg. 3 on the three images tested when a different batch sizes are used. Analogous figures for Alg. 1 and Alg. 2 using the whole image are also reported. For both BKSVD and EBKSVD, the number of batches and the total number of iterations required to estimate $\hat{\mathbf{M}}_P$ decrease with the size of the batch P . The time per iteration grows with P reaching unaffordable values for higher values of P , which supports the idea of working with smaller batches. EBKSVD consumes more time, both with a larger number of iterations required to converge and a higher time per iteration. The times required by Alg. 1 and Alg. 2 are usually higher than those needed by Alg. 3, even when a large batch size is used. Although the tested images are far from the Gigapixel size of a WSI, the total time required to estimate the color vector matrix using the full images shows the importance of the adaptation introduced in section 5 for the use of Bayesian methods on the BCD problem for histological images.

Furthermore, the comparison plotted in Figure 1, depicting the time and convergence ratio for the different images and batch size, shows that the execution time grows linearly with P while the difference between $\hat{\mathbf{M}}_P$ and $\hat{\mathbf{M}}_{all}$ quickly converges to zero. Note that using only a batch size of 50 pixels we achieve a difference in norm of less than 0.05 in most cases. The EBKSVD method, plotted in dashed lines, requires a lower amount of pixels to converge but also requires more time since it needs to find the location of non-zero elements in each step. Using a batch size of 1000 pixels ensures an accurate estimation, with low variance and an affordable computational burden. Note that the time needed by EBKSVD grows significantly faster for batch sizes above 1000 pixels. The BKSVD method is significantly faster but requires more pixels to reach the same output as using the whole image. According to the three images tested and the above mentioned criteria, a batch size of 4000 pixels is the best choice for this method.

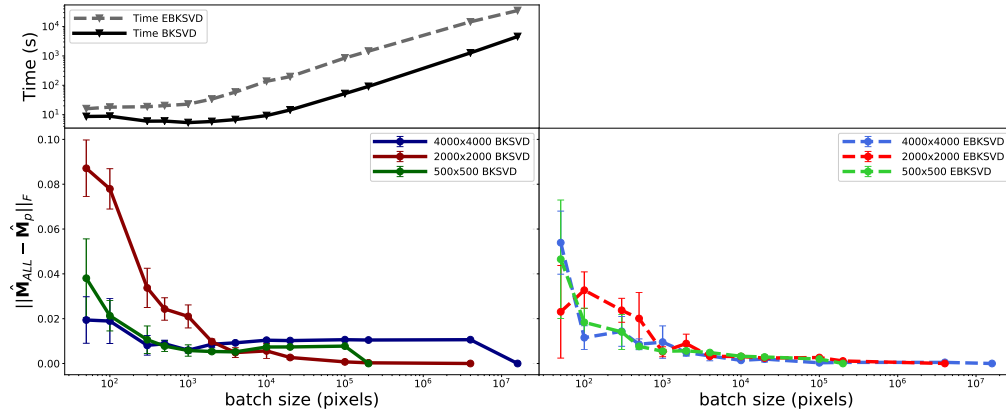


Figure 1: **Top:** Mean time required to obtain the dictionary for the proposed methods. **Bottom:** Difference between the obtained dictionary $\hat{\mathbf{M}}_P$ with a batch size P and the dictionary $\hat{\mathbf{M}}_{all}$ obtained using all pixels for the BKSVD (left) and EBKSVD (right) methods.

Note that both EBKSVD and BKSVD provide an accurate estimation with a batch size of 1000 and 4000 pixels, respectively, for all the image sizes tested. The results plotted in Fig. 1 suggest that these batch sizes will also provide an accurate estimation of \mathbf{M} for larger images in a similar time, making Alg. 3 an scalable solution for obtaining $\hat{\mathbf{M}}$ in WSIs. As a consequence, for the rest of the experiments in this paper, the batch size was fixed to 1000 pixels for EBKSVD and 4000 pixels for BKSVD.

7.2. BCD Stain Separation

Table 3: PSNR and SSIM for the different methods on the WSSB dataset (Alsubaie et al., 2017).

PSNR		RUI	MAC	VAH	ALS	HID	PER	ZHE	EBKSVD	BKSVD
Image	Stain									
Colon	H	22.27	23.91	25.83	21.11	28.57	28.62	17.89	32.12	34.08
	E	20.70	21.55	26.29	21.94	27.58	27.60	14.76	31.11	33.32
Breast	H	15.27	26.24	25.46	24.60	28.81	29.14	15.31	31.69	32.20
	E	17.66	23.62	27.68	25.92	26.60	26.76	14.99	28.81	29.43
Lung	H	22.47	19.52	25.87	20.62	32.91	33.10	19.51	33.06	32.67
	E	22.05	18.09	25.53	23.95	30.77	31.02	16.23	31.87	30.61
Mean	H	20.00	23.22	25.72	22.11	30.10	30.29	17.57	32.29	32.98
	E	20.14	21.08	26.50	23.94	28.32	28.46	15.33	30.60	31.12
SSIM		RUI	MAC	VAH	ALS	HID	PER	ZHE	EBKSVD	BKSVD
Image	Stain									
Colon	H	0.8141	0.8095	0.8851	0.7241	0.9542	0.9544	0.7894	0.9733	0.9826
	E	0.7456	0.6365	0.8904	0.8540	0.9139	0.9161	0.4625	0.9422	0.9646
Breast	H	0.6215	0.9552	0.9239	0.8068	0.9528	0.9560	0.6488	0.9845	0.9801
	E	0.7644	0.9336	0.9550	0.9380	0.9464	0.9492	0.7150	0.9717	0.9632
Lung	H	0.7987	0.7389	0.8912	0.5551	0.9763	0.9757	0.8116	0.9759	0.9764
	E	0.7734	0.5088	0.8195	0.8939	0.9306	0.9353	0.5390	0.9670	0.9461
Mean	H	0.7448	0.8345	0.9100	0.6953	0.9611	0.9621	0.7500	0.9779	0.9797
	E	0.7611	0.6930	0.8883	0.8953	0.9303	0.9336	0.5722	0.9603	0.9580

To evaluate the fidelity of the H&E separation obtained by the different BCD methods, we use the WSSB database (introduced in Section 6). A ground truth separation from WSSB

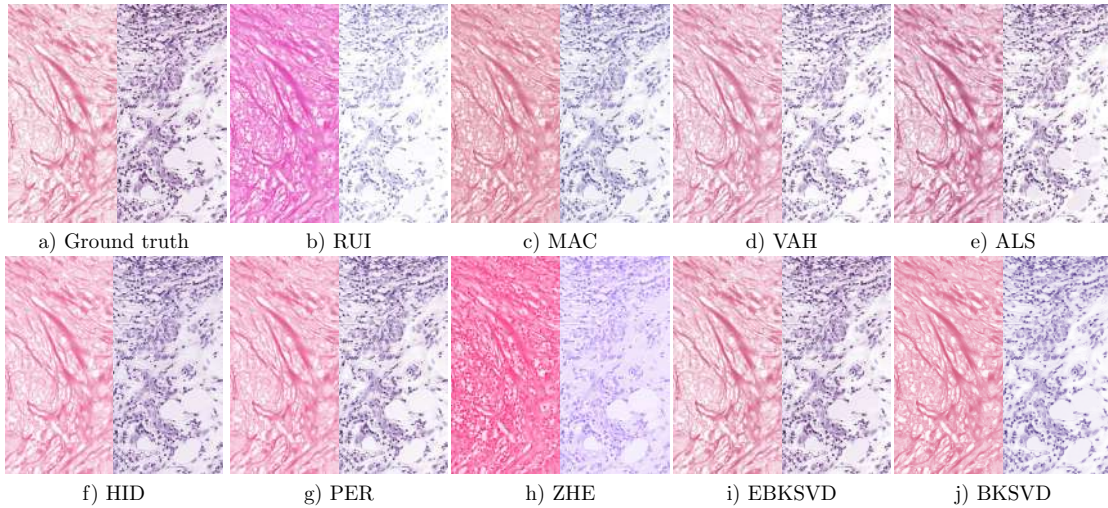


Figure 2: a) Ground truth separated E-only (left) and H-only (right) images from a Breast image of the WSSB dataset in [Alsubaie et al. \(2017\)](#) and results for the b)-h) competing and i-j) proposed methods.

is shown in Figure 2(a). Figures 2(b)–2(j) contain the separated images obtained by different BCD methods. RUI obtains highly contrasted images, but the fixed color vectors are far from those of the ground truth in Figure 2(a). Some nuclei are moved from the H to the E channel. MAC results are closer to the ground truth but the eosin channel still presents residual information from the nuclei. ALS creates artifacts in the flat zones of the H channel and over-saturates the colors. HID obtains colors slightly more saturated than the ground truth and smooths some details. ZHE colors seem unreal and it tends to mix the information of both channels with nuclei clearly appearing in the E channel and cytoplasm in the H channel. The proposed EBKSVD and BKSVD, VAH, as well as PER produce colors very similar to the ground truth separation in Figure 2(a). VAH obtains very similar colors with high differentiation between bands but some information is lost in the H channel, apparently moved to the E channel (see, for instance, the right side of the H channel and the center-left side of the E channel in Figure 2(d)). PER obtains a very good stain separation, although the E color is slightly more reddish than the ground truth. This is due to the prior on the color matrix. It imposes similarity to a reference color vector matrix manually selected for each tissue type. The proposed EBKSVD and BKSVD produce sharp edges, and automatically estimate the color vector matrix without manually selecting a reference. EBKSVD obtains a better mean estimation for the eosin and hematoxylin channels, while BKSVD obtains a slightly darker eosin and a bluish hematoxylin color. Both methods obtain richer details, and a stain separation closer to the ground truth than the competing methods.

The quantitative comparison, based on the Peak Signal to Noise Ratio (PSNR) and Structural Similarity (SSIM), is presented in Table 3. The proposed BKSVD outperforms the rest of methods obtaining a higher mean PSNR (+2.69dB in H and +2.66dB in E) and a higher SSIM than the closest competitor (PER). The proposed EBKSVD obtains the second best mean performance just behind BKSVD, and is able to obtain better values for some tissue types (i.e., lung tissue). For SSIM, both BKSVD and EBKSVD methods are close and the best choice depends on the tissue type.

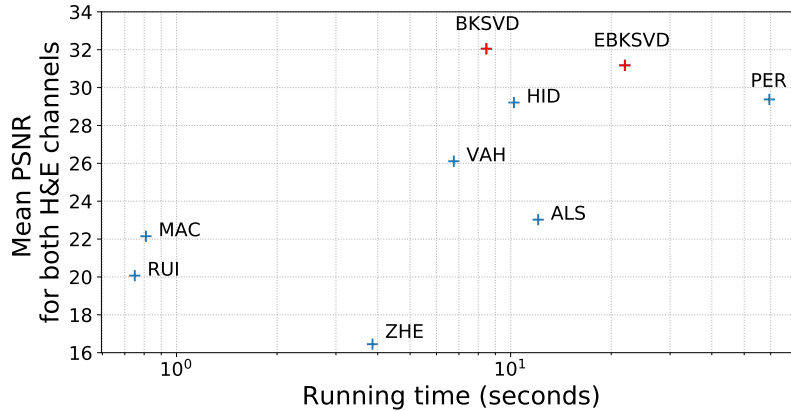


Figure 3: Mean PSNR in dB vs running time in seconds for deconvolving a 2000×2000 image.

The obtained results indicate that the proposed EBKSVD and BKSVD correctly separate the structural information in the image for all tested tissue types. BKSVD obtains the best estimation in mean, mainly due to its higher performance in the colon images. Since colon images are obtained at a lower magnification ($20\times$), this suggests that BKSVD performs better than EBKSVD when a lower magnification is used, that is when a lower sparsity is expected. This is consistent with the results obtained in [Serra et al. \(2017\)](#) where the performance of the EBKSVD is affected by a lower sparsity.

In both cases, the high quality stain separation obtained by the proposed methods guarantees the fidelity to the tissue in CN and SCA transformations detailed in the following sections.

7.2.1. Time Comparison

One important issue with BCD methods is that the required time to perform deconvolution needs to be low enough for practical use. Figure 3 shows the time needed by each BCD method vs. PSNR for the WSSB dataset. The RUI method is the fastest since no color estimation is performed. The computational time increases with the complexity of the method. The proposed BKSVD method outperforms the rest obtaining a significantly higher PSNR while requiring a similar time to HID and VAH methods. EBKSVD obtains the second highest mean PSNR but requires a higher computational time to obtain the sparser solution. Note that the proposed EBKSVD and BKSVD methods are scalable, requiring a similar time for larger images (see Section 7.1).

7.3. Color Normalization

This section compares the color distribution in the original data and the CN obtained by the competing methods. CN is the most extended procedure to deal with stain color variations because CNN based CAD systems usually work with the observed RGB image. CN aims to reduce the impact of color variations on those systems. With the use of BCD, CN can be easily achieved as an additional step, as the stain color information and stain concentrations are separated and can be modified independently. CN based on BCD ensures

Table 4: NMI values for the centers in CAMELYON17.

Method	Center 0		Center 1		Center 2		Center 3		Center 4		All centers	
	SD	CV	SD	CV	SD	CV	SD	CV	SD	CV	SD	CV
Original	0.0403	0.0527	0.0464	0.0667	0.0574	0.0792	0.0601	0.0867	0.0377	0.0441	0.0774	0.1036
MAC	0.0474	0.0734	0.0585	0.1035	0.0855	0.1559	0.0812	0.1489	0.0577	0.0771	0.1032	0.1689
VAH	0.0535	0.0868	0.0658	0.1236	0.0929	0.1787	0.0818	0.1582	0.0638	0.0892	0.1058	0.1823
ALS	0.0512	0.0855	0.0632	0.1303	0.0641	0.1267	0.0841	0.1740	0.0554	0.0821	0.0993	0.1806
HID	0.0413	0.0637	0.0363	0.0576	0.0587	0.0868	0.0463	0.0718	0.0478	0.0636	0.0635	0.0948
PER	0.0405	0.0626	0.0359	0.0570	0.0561	0.0832	0.0454	0.0706	0.0471	0.0628	0.0629	0.0941
ZHE	0.0345	0.0434	0.0277	0.0365	0.0449	0.0608	0.0428	0.0566	0.0311	0.0375	0.0489	0.0632
EBKSVD	0.0243	0.0313	0.0331	0.0440	0.0292	0.0379	0.0327	0.0436	0.0252	0.0323	0.0320	0.0418
BKSVD	0.0202	0.0258	0.0239	0.0317	0.0304	0.0398	0.0280	0.0372	0.0258	0.0329	0.0290	0.0378

fidelity to the image structures, while reducing color variations. Following [Vahadane et al. \(2016\)](#) we normalize an input image to a reference image using

$$\hat{\mathbf{Y}}^{norm} = \sum_{s=1}^{n_s} \hat{\mathbf{m}}_s^{ref} \hat{\mathbf{c}}_{s,:}^{norm}, \quad (16)$$

where

$$\hat{\mathbf{c}}_{s,:}^{norm} = \hat{\mathbf{c}}_{s,:} \frac{P_{99}(\hat{\mathbf{c}}_{s,:}^{ref})}{P_{99}(\hat{\mathbf{c}}_{s,:})}, \quad (17)$$

and $\hat{\mathbf{m}}_s^{ref}$ and $\hat{\mathbf{c}}_s^{ref}$ are the color vectors and concentrations obtained from the reference image. $P_{99}(\cdot)$ represents the pseudo-maximum at 99%. Note that the color vectors $\hat{\mathbf{m}}_s$ are replaced by $\hat{\mathbf{m}}_s^{ref}$ corresponding to the reference image, and the dynamic range of $\hat{\mathbf{c}}_s$ is corrected to be the same as that of $\hat{\mathbf{c}}_s^{ref}$. Therefore, $\hat{\mathbf{Y}}^{norm}$ is the normalized OD image and the normalized RGB image is obtained as $\hat{\mathbf{I}}^{norm} = \exp(-\hat{\mathbf{Y}}^{norm})$.

To measure the quality of the CN, we used the normalized median intensity (NMI) ([Basavanahally and Madabhushi, 2013](#)), defined as

$$NMI(\mathbf{I}) = median(\mathbf{u})/P_{95}(\mathbf{u}), \quad (18)$$

where \mathbf{I} denotes a WSI and \mathbf{u} is a vector where each component u_i is the mean value of the R, G, and B channels at the i th pixel, ([Bejnordi et al., 2016](#)). The NMI value was obtained for each image in a given dataset, and the standard deviation (NMI SD) and coefficient of the variation (NMI CV), i.e., NMI SD divided by the mean, were used as metrics. Lower values of NMI SD and NMI CV indicate a more consistent normalization.

CN tests are carried out on the CAMELYON17 dataset, introduced in Sect. 6, which includes images from 5 different centers. Following [Zheng et al. \(2019\)](#), 500 patches of size 224×224 pixels were sampled from each WSI in the dataset for CN and classification purposes. To avoid the influence of large background regions, only patches with at least 70% tissue were considered. The patch size is motivated for its use in the classification experiments in section 7.4 and does not affect the measurement of the normalization quality.

The results of the proposed and competing CN algorithms for each center and the whole dataset are reported numerically in Table 4 and graphically in Figure 4 where the NMI information for each center and method is plotted as a violin plot. MAC, VAH and ALS transform the images in each center to a similar distribution, but with a larger inter and

intra-center variance than the original images’ distribution. Bayesian methods HID and PER strongly reduce the intra-center differences, but are not able to completely reduce inter-center differences. They have a similar behavior as they share the same similarity prior on the color vector matrix. ZHE significantly reduces intra-center differences but does not completely eliminate inter-center variance. The proposed methods outperform all competitors. Figure 4 and Table 4 show that BKSVD obtains the most consistent normalization, with the lowest intra-center variance and the most similar median values for all the centers in the dataset. EBKSVD closely follows, obtaining the best values for two out of five centers, but with slightly more variation than BKSVD, as can be seen in Figure 4(h) and 4(i).

The CN results were also compared in terms of fidelity to the original observed image using PSNR and SSIM. Although it is important to keep the structure of the original image, notice that fidelity and CN could be conflicting goals as the best fidelity is obtained by not modifying the image. PSNR and SSIM values are shown in Table 5. ZHE obtains the highest fidelity, followed by the proposed BKSVD and EBKSVD. Except for ZHE, that was optimized for its use in CN, the results obtained by the other methods are consistent with those presented in Section 7.2. The better the fidelity to the H&E GT, the better the fidelity after CN. As previously discussed, our methods guarantee fidelity to the H&E bands separately. Since the CN in (16) modifies the concentration dynamic range, it will reduce the similarity to the original image (e.g. by increasing the contrast between stains) but will not have a negative impact on the stain structure and, hence, the PSNR and SSIM values are not heavily affected.

Table 5: PSNR and SSIM for the normalized CAMELYON17 dataset.

	MAC	VAH	ALS	HID	PER	ZHE	EBKSVD	BKSVD
PSNR	13.80	12.74	11.16	17.77	17.73	22.20	19.29	19.54
SSIM	0.7265	0.6490	0.3132	0.8617	0.8644	0.9603	0.8594	0.8735

For a visual qualitative analysis, we depict in Figure 5 a sample patch for each center and the corresponding CN by the different methods. The first row shows the reference image and some 224×224 patches extracted from it showing the variance within the reference image at the same scale as the other patches. The remaining rows show, in the first column, the patch to be normalized and the rest of the columns the CN result with different methods. We notice that MAC and VAH normalize the images but do not obtain colors similar to the reference. ALS introduces color artifacts in most of the patches. ZHE, which is trained to reduce NMI, obtains good figures, but tends to over-brighten the images to reduce NMI variation. The Bayesian methods HID and PER also obtain a consistent normalization, but in some cases, they tend to over-estimate the presence of hematoxylin. The proposed BKSVD and EBKSVD obtain the images most similar to the reference image for all centers, producing high quality results and minimizing the inter-center color variations while maintaining clear differences between both stains. The difference between both methods is difficult to appreciate in this figure. Only in the image of the second center (third row), where hematoxylin and eosin are difficult to differentiate, EBKSVD clearly separates them although it introduces some artifacts, while BKSVD and the other methods do not correctly identify the eosin.

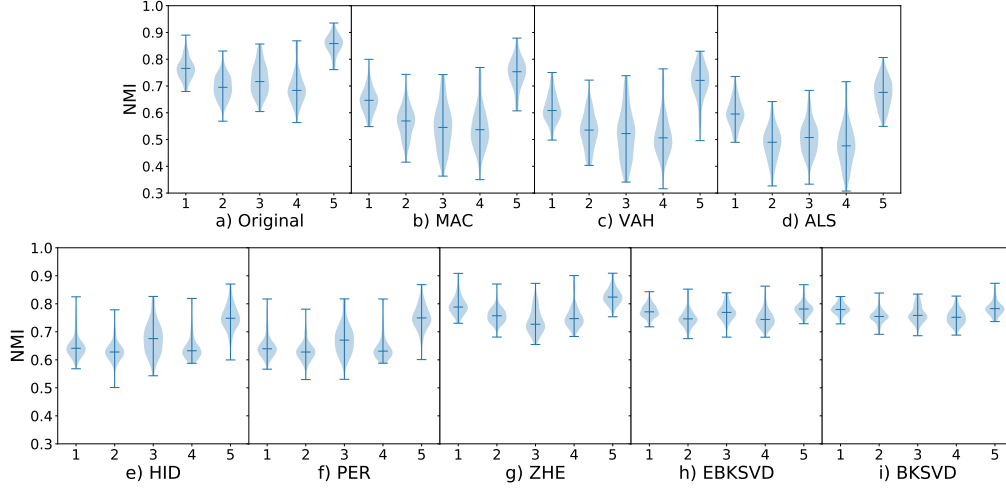


Figure 4: Violin plots of NMI values for each center in CAMELYON17. The blue shadow represents the histogram of NMIs for each plot, the maximum, median and minimum values for each plot are marked with bars. The x-axis indicates the center corresponding to a set of images.

7.4. Data augmentation and cancer classification

The main objective of BCD and CN is to improve the performance of CAD systems, usually based on patch classification systems (Esteban et al., 2019; Tellez et al., 2019). In this section we quantitatively assess the effect of BCD, CN and SCA on a breast cancer detection task (CAMELYON17). For that, we train a VGG19 (Simonyan and Zisserman, 2015) classifier, commonly used in cancer detection (Esteban et al., 2019), on the original, color normalized, and color augmented patches, both from RGB images and OD concentrations.

As previously discussed, using the original WSIs implies dealing with inter-center staining variations that produce generalization errors to unseen stain color variations. BCD and CN aim to reduce the generalization error by reducing color-variation in the input data. However, it is also possible to reduce the generalization error by simulating realistic variations of the training data. The SCA approach is a specific technique of data augmentation for histopathological images that produces realistic variations of the stain colors of the available data. As CN, SCA can also be obtained as an additional step after BCD. While Tellez et al. (2018) applies SCA on the concentrations obtained from Ruifrok and Johnston (2001), we propose to use a combination of both CN and SCA as to obtain an augmented OD image $\hat{\mathbf{Y}}^{aug}$ as follows:

$$\hat{\mathbf{Y}}^{aug} = \sum_{s=1}^{n_s} \hat{\mathbf{m}}_s^{ref} \hat{\mathbf{c}}_{s,:}^{aug}, \quad (19)$$

where the augmented concentrations $\hat{\mathbf{c}}_{s,:}^{aug}$ are synthesized as

$$\hat{\mathbf{c}}_{s,:}^{aug} = \alpha_s \hat{\mathbf{c}}_{s,:}^{norm} + \beta_s \cdot \mathbf{1}, \quad (20)$$

being $\hat{\mathbf{c}}_{s,:}^{norm}$ the normalized concentrations obtained using (17) and α_s, β_s random values following uniform distributions $U(1-\sigma, 1+\sigma)$ and $U(-\sigma, \sigma)$, respectively. This procedure leads to augmentation on the objective reference domain, allowing us to combine the advantages

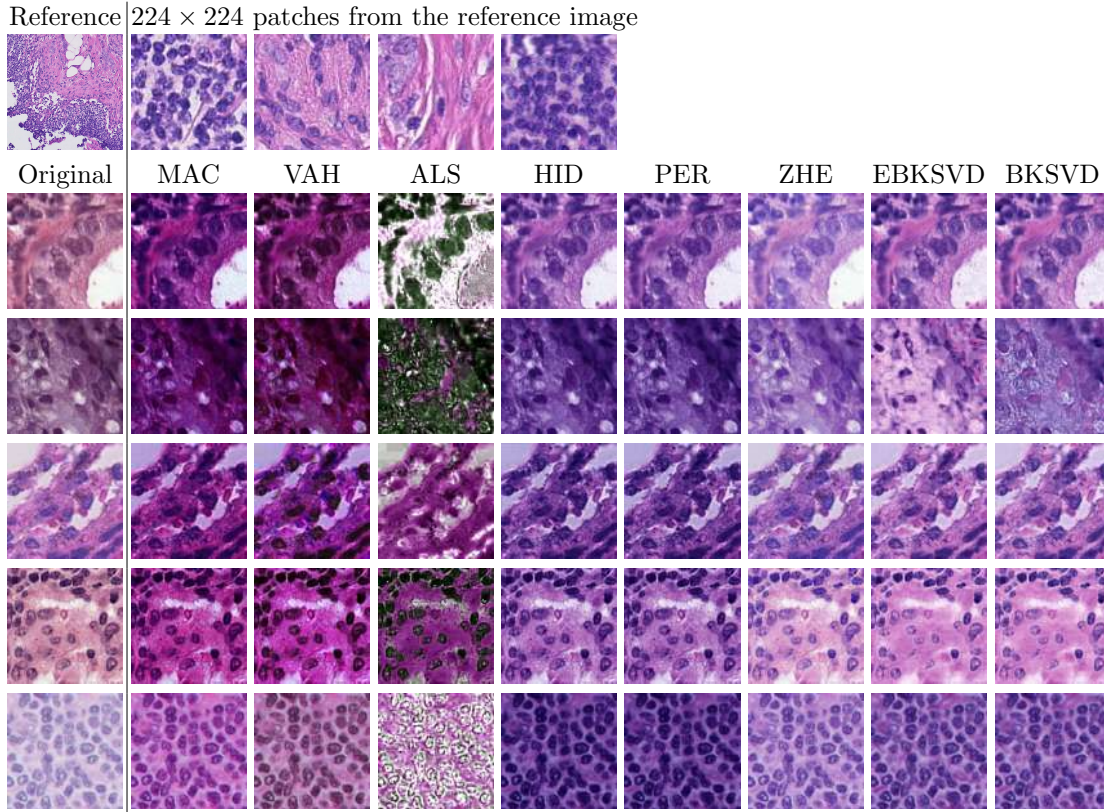


Figure 5: Example 224×224 patches from different centers in CAMELYON17. The first row shows the reference image and some 224×224 patches extracted from the reference. Rows 2–6 correspond to the different centers in CAMELYON17. The original patch is shown in the first column and the other columns show the CN results with different methods.

of both CN and SCA approaches. CN will reduce the variation between centers and SCA will cover the variations that were not completely captured by the CN.

We train the network with the RGB normalized images, OD concentrations obtained by the BCD methods, the SCA in [Tellez et al. \(2018\)](#), denoted by TEL, and the SCA using [\(19\)](#), denoted by $BKSVD_{aug}$ and $EBKSVD_{aug}$ depending on whether we use the BKSVD or EBKSVD concentrations. Following [Tellez et al. \(2018\)](#), $\sigma = 0.05$ and $\sigma = 0.2$ were used for light and strong augmentation, respectively.

From CAMELYON17, four centers were used for training and the 5th center, which showed a bigger color difference in the previous section, was used as test set. From the 50 tumor annotated WSIs in CAMELYON17, approximately 55,000 positive patches were sampled for training and 12,500 for testing. Negative patches were sampled from negative WSIs only, obtaining 55,000 for training and 12,500 for testing.

VGG19 was trained from scratch for 100 epochs in each case using 64 sample batches with batch normalization. The learning rate was initially set to 0.01, which is halved every 30 epochs. When using OD concentrations, the architecture was modified to use 2 input channels (H&E) instead of the RGB image. The area under the ROC curve (AUC), shown in

Tables 6 and 7, was calculated on the test set for the best performing epoch during training for each method.

Table 6: AUC performance of the VGG19 classifier for the proposed and competing methods using CN both on the RGB and OD spaces. Bold values indicate the highest performance for each space.

Input	Original	RUI	MAC	ALS	HID	VAH	PER	ZHE	EBKSVD	BKSVD
RGB	0.9491	NA	0.9499	0.9738	0.9479	0.7985	0.9305	0.9755	0.9817	0.9711
OD	NA	0.9417	0.9468	0.9725	0.9642	0.6614	0.9508	0.9864	0.9834	0.9672

Table 7: AUC performance of the VGG19 classifier for the proposed and competing methods using SCA both on the RGB and OD spaces. Bold values indicate the highest performance for each space.

Input	TEL ^{strong} _{aug}	TEL ^{light} _{aug}	EBKSVD ^{strong} _{aug}	EBKSVD ^{light} _{aug}	BKSVD ^{strong} _{aug}	BKSVD ^{light} _{aug}
RGB	0.9673	0.9601	0.9716	0.9647	0.9679	0.9650
OD	0.9654	0.9639	0.9865	0.9879	0.9728	0.9790

The results show that, when using RGB images, CN increased the AUC in most cases, increasing it from the original images (0.9491) up to 0.9817 with the proposed EBKSVD. The less sparse BKSVD approach, slightly increases the AUC without reaching the outperforming result of the EBKSVD. CN obtained by ZHE and ALS also increased the classification performance considerably despite the over-brightened images produced by ZHE and the artifacts produced by ALS.

Results using OD concentrations show that most methods increase AUC in comparison to the baseline RUI method. Also, the performance of HID, PER, ZHE and the proposed EBKSVD is better in OD than in normalized RGB space, showing that BCD is able to provide more useful information for the CNN. Separating the structures in the image from the color information, usually produces better results than using the RGB image since the network does not need to extract the structural information from colors. SCA improves the performance with respect to the original images, both using RGB and OD concentrations, obtaining the best performance with the latter. The highest AUC value was obtained using EBKSVD and light SCA in the OD space. Our results show that SCA benefits from the use of EBKSVD instead of the RUI method used by TEL. The difference between light and strong augmentation is minor both in TEL and the proposed augmentation. Our results show that CN and BCD have a bigger impact on classification than SCA when RGB images are used. However, the proposed combination of CN and SCA improves the results on the OD space.

8. Conclusions

In this paper, we have proposed a novel Bayesian approach for blind color deconvolution of histopathological images, based on K-SVD with two possible inference approaches: variational and empirical Bayes. We utilize a hierarchical prior on the concentrations that enforces sparsity in the same way as a Laplacian prior while allowing for a tractable Bayesian inference. The framework presented automatically estimates the stain concentrations, the

color-vector matrix, and all model parameters. The proposed BKSVD and EBKSVD methods guarantee fidelity to the tissue structure on different relevant histopathological tasks such as color normalization, stain color augmentation, and classification of histological images.

The proposed method is designed to work at the highest magnification available. Although the proposed approach has shown a good performance at $20\times$ and $40\times$, it is unclear how magnification affects the estimation of the color-vector matrix and has never been explored in the literature. This is an interesting topic to be addressed in future research, specially if hierarchical model are to be used.

The proposed approach solves the dependency on the reference color-vector matrix of previous Bayesian approaches. However, this also exposes a limitation that affects to many other BCD and CN methods: the common assumption that colors on the image come exclusively from H&E stains might not hold in some scenarios. Although the proposed Bayesian approach and the pixel sampling provide a certain robustness to variations, large areas of blood, cauterized tissue (e.g. bladder samples) or other anomalies in the WSIs can affect the BCD results and therefore the CN or SCA performance. This issue, that also affects CNN-based CN methods, has never been explored in the BCD or CN fields and needs to be addressed in future research.

The proposed BKSVD and EBKSVD methods outperform classical and state-of-the-art methods on all the performed experiments obtaining higher fidelity to the tissue structure, a more consistent normalization, and a stain specific color augmentation that improves classification on VGG19. The optimal approach, BKSVD or EBKSVD, varies depending on the task.

We have analyzed the effect of using color normalized images or OD concentrations to feed a CNN classifier. The carried out experiments indicate that using OD concentrations for H&E achieves higher classification performance than feeding the network with RGB images. The dependency on a reference image is a well-known issue for BCD-based CN. The choice of a proper reference image also have an impact on the classification performance. The relevance of this choice needs to be quantified in future research. However, it can be avoided with the use of OD concentrations directly for classification.

Finally, we have shown that stain color augmentation techniques are more beneficial when using high-quality stain concentrations that better represent the real structure of the stains in the image. The use of the OD concentrations as input for the network is also useful when working with augmentation techniques.

Appendix A. Derivation of the Sequential Inference for Sparse Bayesian Models

We detail now the maximization of the marginal likelihood in (12), which we introduce here for the sake of completeness

$$\mathcal{L}(\boldsymbol{\gamma}_q) = \log \left[p(\boldsymbol{\gamma}_q | \hat{\lambda}_q) \int p(\mathbf{y}_q | \mathbf{c}_q, \hat{\beta}) p(\mathbf{c}_q | \boldsymbol{\gamma}_q) d\mathbf{c}_q \right], \quad (\text{A.1})$$

where $p(\mathbf{y}_q | \mathbf{c}_q, \hat{\beta}) \sim \mathcal{N}(\hat{\mathbf{M}}\mathbf{c}_q, \hat{\beta}^{-1}\mathbf{I})$, which is clear from the observation model in (1); $p(\mathbf{c}_q | \boldsymbol{\gamma}_q) \sim \mathcal{N}(\mathbf{c}_q | \mathbf{0}, \boldsymbol{\Gamma}_q)$ as defined in Sec. 3; and, the remaining variables, \mathbf{M} , β and λ_q , are fixed to the values estimated with variational inference. The marginal integral in (A.1) is a well-known

result:

$$\mathbb{P}(\mathbf{y}_q | \hat{\beta}, \hat{\mathbf{M}}, \boldsymbol{\gamma}_q) := \int \mathbb{P}(\mathbf{y}_q | \mathbf{c}_q, \hat{\beta}) \mathbb{P}(\mathbf{c}_q | \boldsymbol{\gamma}_q) d\mathbf{c}_q = \mathcal{N}(\mathbf{y}_q | \mathbf{0}, \mathbf{X}_q), \quad (\text{A.2})$$

with covariance matrix

$$\mathbf{X}_q = \hat{\beta}^{-1} \mathbf{I}_3 + \hat{\mathbf{M}} \boldsymbol{\Gamma}_q \hat{\mathbf{M}}^T. \quad (\text{A.3})$$

Now we can rewrite the marginal likelihood as

$$\begin{aligned} \mathcal{L}(\boldsymbol{\gamma}_q) &= \log \mathbb{P}(\boldsymbol{\gamma}_q | \hat{\lambda}_q) \mathbb{P}(\mathbf{y}_q | \hat{\beta}, \hat{\mathbf{M}}, \boldsymbol{\gamma}_q) \\ &= -\frac{1}{2} \left[\log |\mathbf{X}_q| + \mathbf{y}_q^T \mathbf{X}_q^{-1} \mathbf{y}_q + \hat{\lambda}_q \sum_s \gamma_{sq} \right] + \text{const.}, \end{aligned} \quad (\text{A.4})$$

where the constant includes all terms not depending on $\boldsymbol{\gamma}_q$.

Notice that we can easily find the posterior distribution of \mathbf{c}_q , using (6) and (7), once $\hat{\boldsymbol{\gamma}}_q$ has been calculated. In addition, if $\gamma_{sq} = 0$, then the posterior distribution of c_{sq} will be degenerate at zero.

The marginal likelihood $\mathcal{L}(\boldsymbol{\gamma}_q)$ has interesting properties that result in a sequential maximization strategy which will allow us to add, update or remove a single γ_{sq} in order to increase $\mathcal{L}(\boldsymbol{\gamma}_q)$. Concretely, see how we can isolate the contribution of a single γ_{sq} in the covariance matrix \mathbf{X}_q writing

$$\mathbf{X}_q = \left[\hat{\beta}^{-1} \mathbf{I}_3 + \sum_{i \neq s} \gamma_{iq} \hat{\mathbf{m}}_i \hat{\mathbf{m}}_i^T \right] + \gamma_{sq} \hat{\mathbf{m}}_s \hat{\mathbf{m}}_s^T =: \tilde{\mathbf{X}}_q + \gamma_{sq} \hat{\mathbf{m}}_s \hat{\mathbf{m}}_s^T, \quad (\text{A.5})$$

where, clearly, $\tilde{\mathbf{X}}_q$ has no dependence on γ_{sq} . Using the determinant identity and the matrix inversion lemma on \mathbf{X}_q we can write

$$\mathbf{X}_q^{-1} = \tilde{\mathbf{X}}_q^{-1} - \frac{\tilde{\mathbf{X}}_q^{-1} \hat{\mathbf{m}}_s \hat{\mathbf{m}}_s^T \tilde{\mathbf{X}}_q^{-1}}{\gamma_{sq}^{-1} + \hat{\mathbf{m}}_s^T \tilde{\mathbf{X}}_q^{-1} \hat{\mathbf{m}}_s}, \quad (\text{A.6})$$

$$|\mathbf{X}_q| = |\tilde{\mathbf{X}}_q| \cdot |1 + \gamma_{sq} \hat{\mathbf{m}}_s^T \tilde{\mathbf{X}}_q^{-1} \hat{\mathbf{m}}_s|. \quad (\text{A.7})$$

The previous equations allow us to rewrite (A.4) as

$$\begin{aligned} \mathcal{L}(\boldsymbol{\gamma}_q) &= -\frac{1}{2} \left[\log |\tilde{\mathbf{X}}_q| + \mathbf{y}_q^T \tilde{\mathbf{X}}_q^{-1} \mathbf{y}_q + \hat{\lambda}_q \sum_{n \neq s} \gamma_{nq} \right] + \frac{1}{2} \left[\log \frac{1}{1 + \gamma_{sq} g_{sq}} + \frac{h_{sq}^2 \gamma_{sq}}{1 + \gamma_{sq} g_{sq}} - \hat{\lambda}_q \gamma_{sq} \right] \\ &=: \mathcal{L}(\{\gamma_{iq}\}_{i \neq s}) + l(\gamma_{sq}), \end{aligned} \quad (\text{A.8})$$

where $g_{sq} = \hat{\mathbf{m}}_s^T \tilde{\mathbf{X}}_q^{-1} \hat{\mathbf{m}}_s$ and $h_{sq} = \hat{\mathbf{d}}_s^T \tilde{\mathbf{X}}_q^{-1} \mathbf{y}_q$ and the constant has been omitted as it plays no role in the optimization. Notice that the quantities g_{sq} and h_{sq} do not depend on γ_{sq} . Therefore, the terms related to a single hyperparameter γ_{sq} are now separated from the rest. A closed form solution of the maximization of $\mathcal{L}(\boldsymbol{\gamma}_q)$, when only its s th component is changed, can be found by holding the other hyperparameters fixed, taking its derivative with respect to γ_{sq} and setting it equal to zero, obtaining a unique maximum at

$$\hat{\gamma}_{sq} = \begin{cases} \frac{-(g_{sq} + 2\hat{\lambda}_q) + \sqrt{g_{sq}^2 - 4\hat{\lambda}_q h_{sq}^2}}{2\hat{\lambda}_q g_{sq}}, & h_{sq}^2 - g_{sq} \geq \hat{\lambda}_q \\ 0, & \text{otherwise.} \end{cases} \quad (\text{A.9})$$

In order to effectively reduce the computational burden, this calculation must be performed efficiently. To explain how to carry them out, let us overload slightly the notation. The current $(^c)$ covariance matrix of the marginal of the observations is rewritten as

$$\mathbf{X}_q^c = \hat{\beta}^{-1} \mathbf{I}_3 + \sum_{i \in \mathcal{A}} \gamma_{iq}^c \hat{\mathbf{m}}_i \hat{\mathbf{m}}_i^T + \sum_{i \in \bar{\mathcal{A}}} \gamma_{iq}^c \hat{\mathbf{m}}_i \hat{\mathbf{m}}_i^T, \quad (\text{A.10})$$

where $\mathcal{A} = \{i | \gamma_{iq}^c > 0\}$ and $\bar{\mathcal{A}} = \{i | \gamma_{iq}^c = 0\}$. Notice that, the last term on the right hand side of (A.10) is equal to zero and has been included for clarity. Then, applying the Woodbury identity, we obtain

$$\hat{\mathbf{m}}_s^T \mathbf{X}_q^{c-1} \hat{\mathbf{m}}_s = \hat{\beta} \hat{\mathbf{m}}_s^T \hat{\mathbf{m}}_s - \hat{\beta}^2 \hat{\mathbf{m}}_s^T \hat{\mathbf{M}}^c \Sigma_{\mathbf{c}_q}^c (\hat{\mathbf{M}}^c)^T \hat{\mathbf{m}}_s =: G_{sq} \quad (\text{A.11})$$

where $\Sigma_{\mathbf{c}_q}^c$ is obtained from $\Sigma_{\mathbf{c}_q}$ by keeping only the columns and rows associated to the indices in \mathcal{A} . We apply the same restriction to the columns of $\hat{\mathbf{M}}^c$, that is, we keep in $\hat{\mathbf{M}}^c$ the columns associated to $\gamma_{iq}^c > 0$. From (A.6), for $s \in \mathcal{A} \cup \bar{\mathcal{A}}$, we have

$$g_{sq} = \frac{G_{sq}}{1 - \gamma_{sq}^c G_{sq}}. \quad (\text{A.12})$$

Furthermore

$$\hat{\mathbf{m}}_s^T \mathbf{X}_q^{c-1} \mathbf{y}_q = \hat{\beta} \hat{\mathbf{m}}_s^T \mathbf{y}_q - \hat{\beta}^2 \hat{\mathbf{m}}_s^T \hat{\mathbf{M}}^c \Sigma_{\mathbf{c}_q}^c (\hat{\mathbf{M}}^c)^T \mathbf{y}_q =: H_{sq} \quad (\text{A.13})$$

Using an analogous procedure we can write

$$h_{sq} = \frac{H_{sq}}{1 - \gamma_{sq}^c G_{sq}}. \quad (\text{A.14})$$

Given $\Sigma_{\mathbf{c}_q}^c$ we can now efficiently check whether we should add γ_{sq} , $s \in \bar{\mathcal{A}}$, or update, or remove γ_{sq} , $s \in \mathcal{A}$. Moreover, the amount the marginal log likelihood is improved by each single addition, update, or removal is easily calculated from (A.8). Finally, we notice that $\Sigma_{\mathbf{c}_q}^c$ and $\hat{\mathbf{c}}_q^c$ can be updated very efficiently considering only a single coefficient γ_{sq} , see [Tipping and Faul \(2003\)](#).

Acknowledgements

This work was supported by project PID2019-105142RB-C22 funded by MCIN / AEI / 10.13039 / 501100011033 and project P20_00286 funded by FEDER/Junta de Andalucía-Consejería de Transformación Económica, Industria, Conocimiento y Universidades. The work by Fernando Pérez-Bueno was sponsored by Ministerio de Economía, Industria y Competitividad under FPI contract BES-2017-081584.

References

Aharon, M., Elad, M., Bruckstein, A., 2006. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing* 54, 4311–4322.

- Alsubaie, N., Raza, S.E.A., Rajpoot, N., 2016. Stain deconvolution of histology images via independent component analysis in the wavelet domain, in: 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI), pp. 803–806.
- Alsubaie, N., Trahearn, N., Raza, S.E.A., Snead, D., Rajpoot, N., 2017. Stain deconvolution using statistical analysis of multi-resolution stain colour representation. *PLOS ONE* 12, e0169875.
- Astola, L., 2016. Stain separation in digital bright field histopathology, in: 2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA), pp. 1–6.
- Babacan, S.D., Luessi, M., Molina, R., Katsaggelos, A.K., 2012. Sparse Bayesian methods for low-rank matrix estimation. *IEEE Transaction on Signal Processing* 60, 3964–3977.
- Babacan, S.D., Molina, R., Katsaggelos, A.K., 2010. Bayesian compressive sensing using Laplace priors. *IEEE Transactions on Image Processing* 19, 53–63.
- Basavanahally, A., Madabhushi, A., 2013. EM-based segmentation-driven color standardization of digitized histopathology, in: *Progress in Biomedical Optics and Imaging - Proceedings of SPIE*, p. 86760G.
- Bejnordi, B.E., Litjens, G., Timofeeva, N., Otte-Höller, I., Homeyer, A., Karssemeijer, N., Laak, J.A.v.d., 2016. Stain specific standardization of whole-slide histopathological images. *IEEE Transactions on Medical Imaging* 35, 404–415.
- Bentaieb, A., Hamarneh, G., 2018. Adversarial stain transfer for histopathology image analysis. *IEEE Transactions on Medical Imaging* 37, 792–802.
- Bishop, C., 2006. *Pattern Recognition and Machine Learning*. Springer. pp. 454–455.
- Bándi, P., Geessink, O., Manson, Q., Van Dijk, M., Balkenhol, M., Hermsen, M., Ehteshami Bejnordi, B., Lee, B., Paeng, K., Zhong, A., Li, Q., Zanjani, F.G., Zinger, S., Fukuta, K., Komura, D., Ovtcharov, V., Cheng, S., Zeng, S., Thagaard, J., Dahl, A.B., Lin, H., Chen, H., Jacobsson, L., Hedlund, M., Çetin, M., Halıcı, E., Jackson, H., Chen, R., Both, F., Franke, J., Küsters-Vandeveld, H., Vreuls, W., Bult, P., van Ginneken, B., van der Laak, J., Litjens, G., 2019. From detection of individual metastases to classification of lymph node status at the patient level: The CAMELYON17 challenge. *IEEE Transactions on Medical Imaging* 38, 550–560.
- Duggal, R., Gupta, A., Gupta, R., Mallick, P., 2017. SD-Layer: Stain Deconvolutional Layer for CNNs in Medical Microscopic Imaging, in: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (Eds.), *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017*. Lecture Notes in Computer Science. Springer, Cham, pp. 435–443.
- Esteban, A.E., Lopez-Perez, M., Colomer, A., Sales, M.A., Molina, R., Naranjo, V., 2019. A new optical density granulometry-based descriptor for the classification of prostate histological images using shallow and deep Gaussian processes. *Computer Methods and Programs in Biomedicine* 178, 303–317.

- Fischer, A.H., Jacobson, K.A., Rose, J., Zeller, R., 2008. Hematoxylin and Eosin Staining of Tissue and Cell Sections. Cold Spring Harbor Protocols.
- Gavrilovic, M., Azar, J.C., Lindblad, J., Wählby, C., Bengtsson, E., Busch, C., Carlbom, I.B., 2013. Blind color decomposition of histological images. *IEEE Transactions on Medical Imaging* 32, 983–994.
- Hidalgo-Gavira, N., Mateos, J., Vega, M., Molina, R., Katsaggelos, A.K., 2018. Blind color deconvolution of histopathological images using a variational Bayesian approach, in: *International Conference on Image Processing (ICIP)*, pp. 983–987.
- Hidalgo-Gavira, N., Mateos, J., Vega, M., Molina, R., Katsaggelos, A.K., 2020. Variational Bayesian blind color deconvolution of histopathological images. *IEEE Transactions on Image Processing* 29, 2026–2036.
- Hoque, M.Z., Keskinarkaus, A., Nyberg, P., Seppänen, T., 2021. Retinex model based stain normalization technique for whole slide image analysis. *Computerized Medical Imaging and Graphics* 90, 101901. URL: <https://www.sciencedirect.com/science/article/pii/S0895611121000495>, doi:<https://doi.org/10.1016/j.compmedimag.2021.101901>.
- Janowczyk, A., Basavanthally, A., Madabhushi, A., 2017. Stain normalization using sparse autoencoders (StaNoSA): Application to digital pathology. *Computerized Medical Imaging and Graphics* 57, 50–61.
- Khan, A.M., Rajpoot, N., Treanor, D., Magee, D., 2014. A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution. *IEEE Transactions on Biomedical Eng.* 61, 1729–1738.
- Lan, J., Cai, S., Xue, Y., Gao, Q., Du, M., Zhang, H., Wu, Z., Deng, Y., Huang, Y., Tong, T., Chen, G., 2021. Unpaired stain style transfer using invertible neural networks based on channel attention and long-range residual. *IEEE Access* 9, 11282–11295. doi:[10.1109/ACCESS.2021.3051188](https://doi.org/10.1109/ACCESS.2021.3051188).
- Liu, Y., Gadepalli, K.K., Norouzi, M., Dahl, G., Kohlberger, T., Venugopalan, S., Boyko, A.S., Timofeev, A., Nelson, P.Q., Corrado, G., Hipp, J., Peng, L., Stumpe, M., 2017. Detecting cancer metastases on gigapixel pathology images. arXiv Also presented at the 2017 MICCAI tutorial, Deep Learning for Medical Imaging: <https://sites.google.com/view/miccai2017-deeplearning>.
- Macenko, M., Niethammer, M., et al., 2009. A method for normalizing histology slides for quantitative analysis, in: *International Symposium on Biomedical Imaging (ISBI)*, pp. 1107–1110.
- Mairal, J., Bach, F., Ponce, J., Sapiro, G., 2009. Online dictionary learning for sparse coding, in: *Proceedings of the 26th Annual International Conference on Machine Learning*, Association for Computing Machinery, New York, NY, USA. p. 689–696.

- McCann, M.T., Majumdar, J., et al., 2014. Algorithm and benchmark dataset for stain separation in histology images, in: International Conference on Image Processing (ICIP), pp. 3953–3957.
- Mpinda Ataky, S.T., de Matos, J., Britto, A.d.S., Oliveira, L.E.S., Koerich, A.L., 2020. Data augmentation for histopathological images based on gaussian-laplacian pyramid blending, in: 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. doi:[10.1109/IJCNN48605.2020.9206855](https://doi.org/10.1109/IJCNN48605.2020.9206855).
- Pérez-Bueno, F., López-Pérez, M., Vega, M., Mateos, J., Naranjo, V., Molina, R., Katsaggelos, A.K., 2020. A TV-based image processing framework for blind color deconvolution and classification of histological images. *Digital Signal Processing* 101, 102727.
- Pérez-Bueno, F., Vega, M., Sales, M.A., Aneiros-Fernández, J., Naranjo, V., Molina, R., Katsaggelos, A.K., 2021. Blind color deconvolution, normalization, and classification of histological images using general super gaussian priors and bayesian inference. *Computer Methods and Programs in Biomedicine* 211, 106453. URL: <https://www.sciencedirect.com/science/article/pii/S0169260721005277>, doi:<https://doi.org/10.1016/j.cmpb.2021.106453>.
- Rabinovich, A., Agarwal, S., Laris, C., Price, J.H., Belongie, S.J., 2004. Unsupervised color decomposition of histologically stained tissue samples, in: *Advances in Neural Information Processing Systems*, pp. 667–674.
- Reinhard, E., Adhikhmin, M., Gooch, B., Shirley, P., 2001. Color transfer between images. *IEEE Computer Graphics and Applications* 21, 34–41.
- Ruifrok, A.C., Johnston, D.A., 2001. Quantification of histochemical staining by color deconvolution. *Analytical and quantitative cytology and histology* 23, 291–299.
- Runz, M., Rusche, D., Schmidt, S., Weihrauch, M.R., Hesser, J., Weis, C.A., 2021. Normalization of HE-stained histological images using cycle consistent generative adversarial networks. *Diagnostic Pathology* 16, 71. URL: <https://doi.org/10.1186/s13000-021-01126-y>, doi:[10.1186/s13000-021-01126-y](https://doi.org/10.1186/s13000-021-01126-y).
- Salehi, P., Chalechale, A., 2020. Pix2pix-based stain-to-stain translation: A solution for robust stain normalization in histopathology images analysis, in: 2020 International Conference on Machine Vision and Image Processing (MVIP), pp. 1–7. doi:[10.1109/MVIP49855.2020.9116895](https://doi.org/10.1109/MVIP49855.2020.9116895).
- Salvi, M., Michielli, N., Molinari, F., 2020. Stain color adaptive normalization (SCAN) algorithm: Separation and standardization of histological stains in digital pathology. *Computer Methods and Programs in Biomedicine* 193, 105506.
- Serra, J., Testa, M., Molina, R., Katsaggelos, A.K., 2017. Bayesian K-SVD using fast variational inference. *IEEE Transactions on Image Processing* 26, 3344–3348.
- Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. [arXiv:1409.1556](https://arxiv.org/abs/1409.1556).

- Tellez, D., Balkenhol, M., Otte-Höller, I., van de Loo, R., Vogels, R., Bult, P., Wauters, C., Vreuls, W., Mol, S., Karssemeijer, N., Litjens, G., van der Laak, J., Ciompi, F., 2018. Whole-slide mitosis detection in H&E breast histology using PHH3 as a reference to train distilled stain-invariant convolutional networks. *IEEE Transactions on Medical Imaging* 37, 2126–2136.
- Tellez, D., Litjens, G., Bándi, P., Bulten, W., Bokhorst, J.M., Ciompi, F., van der Laak, J., 2019. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical Image Analysis* 58, 101544.
- Tipping, M., Faul, A., 2003. Fast marginal likelihood maximisation for sparse Bayesian models, in: *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, pp. 3–6.
- Tosta, T.A.A., de Faria, P.R., Servato, J.P.S., Neves, L.A., Roberto, G.F., Martins, A.S., do Nascimento, M.Z., 2019a. Unsupervised method for normalization of hematoxylin-eosin stain in histological images. *Computerized Medical Imaging and Graphics* 77, 101646.
- Tosta, T.A.A., de Faria, P.R., Neves, L.A., do Nascimento, M.Z., 2019b. Computational normalization of H&E-stained histological images: Progress, challenges and future potential. *Artificial Intelligence in Medicine* 95, 118 – 132.
- Trahearn, N., Snead, D., Cree, I., Rajpoot, N., 2015. Multi-class stain separation using independent component analysis, in: *Medical Imaging 2015: Digital Pathology*, p. 94200J.
- Vahadane, A., Peng, T., Sethi, A., Albarqouni, S., Wang, L., Baust, M., Steiger, K., Schlitter, A.M., Esposito, I., Navab, N., 2016. Structure-preserving color normalization and sparse stain separation for histological images. *IEEE Transactions on Medical Imaging* 35, 1962–1971.
- Vicory, J., Couture, H.D., Thomas, N.E., Borland, D., Marron, J., Woosley, J., Niethammer, M., 2015. Appearance normalization of histology slides. *Computerized Medical Imaging and Graphics* 43, 89–98.
- Vijh, S., Saraswat, M., Kumar, S., 2021. A new complete color normalization method for H&E stained histopathological images. *Applied Intelligence* 51, 7735–7748. URL: <https://doi.org/10.1007/s10489-021-02231-7>, doi:10.1007/s10489-021-02231-7.
- Wei, J., Suriawinata, A., Vaickus, L., Ren, B., Liu, X., Wei, J., Hassanpour, S., 2020. Generative image translation for data augmentation in colorectal histopathology images, in: *Proceedings of the Machine Learning for Health NeurIPS Workshop*, PMLR 116:10-24, p. 16.
- Xiang, Y., Chen, J., Liu, Q., Liang, Y., 2020. Disentangled representation learning based multidomain stain normalization for histological images, in: *2020 IEEE International Conference on Image Processing (ICIP)*, pp. 360–364. doi:10.1109/ICIP40778.2020.9190757.

- Xu, J., Xiang, L., Wang, G., Ganesan, S., Feldman, M., Shih, N.N., Gilmore, H., Madabhushi, A., 2015. Sparse non-negative matrix factorization (SNMF) based color unmixing for breast histopathological image analysis. *Computerized Medical Imaging and Graphics* 46, 20–29.
- Zanjani, F.G., Zinger, S., Bejnordi, B.E., van der Laak, J.A.W.M., de With, P.H.N., 2018. Stain normalization of histopathology images using generative adversarial networks, in: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI), pp. 573–577.
- Zheng, Y., Jiang, Z., Zhang, H., Xie, F., Hu, D., Sun, S., Shi, J., Xue, C., 2021. Stain standardization capsule for application-driven histopathological image normalization. *IEEE Journal of Biomedical and Health Informatics* 25, 337–347. doi:[10.1109/JBHI.2020.2983206](https://doi.org/10.1109/JBHI.2020.2983206).
- Zheng, Y., Jiang, Z., Zhang, H., Xie, F., Shi, J., Xue, C., 2019. Adaptive color deconvolution for histological WSI normalization. *Computer Methods and Programs in Biomedicine* 170, 107–120.
- Zhou, M., Chen, H., Ren, L., Sapiro, G., Carin, L., Paisley, J., 2009. Non-parametric Bayesian dictionary learning for sparse image representations, in: Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C., Culotta, A. (Eds.), *Advances in Neural Information Processing Systems*, Curran Associates, Inc.. pp. 2295–2303.
- Zhu, J.Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks, in: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2242–2251. doi:[10.1109/ICCV.2017.244](https://doi.org/10.1109/ICCV.2017.244).

CHAPTER 5

Pansharpening of Multispectral Images Using Probabilistic Models

5.1 JCR Publication Details

Authors: Fernando Pérez-Bueno, Miguel Vega, Javier Mateos, Rafael Molina, Aggelos K. Katsaggelos

Title: Variational Bayesian Pansharpening with Super-Gaussian Sparse Image Priors

Reference: *Sensors*, 2020, 20, 5308

Status: Published

DOI: <https://doi.org/10.3390/s20185308>

Quality indices:






- Impact Factor (JCR 2020): 3.576
 - Rank: 14/64 (Q1) in Instruments and Instrumentation
 - Rank: 82/273 (Q2) in Engineering, Electrical and Electronic
- Journal Citation Indicator (JCR 2021): 0.89
 - Rank: 14/72 (Q1) in Instruments and Instrumentation
 - Rank: 100/319 (Q2) in Engineering, Electrical and Electronic

5.2 Main Contributions

- We propose a variational Bayesian methodology for pansharpening of multispectral images based on the use of SG priors, with fully automatic estimation of the model parameters.

- The model is evaluated with two representative members of the SG distributions, those corresponding to l_p and log energy functions.
- The proposed approach was evaluated using real and synthetic data from three different satellites.

Variational Bayesian Pansharpening with Super-Gaussian Sparse Image Priors

Fernando Pérez-Bueno ^{1,*}, Miguel Vega ², Javier Mateos ¹, Rafael Molina ¹ and Aggelos K. Katsaggelos ³

¹ Dpto. de Ciencias de la Computación e I. A., Universidad de Granada, Spain; {fpb,jmd,rms}@decsai.ugr.es

² Dpto. de Lenguajes y Sistemas Informáticos, Universidad de Granada, Spain; mvega@ugr.es

³ Dept. of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL, USA; aggk@eecs.northwestern.edu

* Correspondence: fpb@decsai.ugr.es

Abstract: Pansharpening is a technique that fuses a low spatial resolution multispectral image and a high spatial resolution panchromatic one to obtain a multispectral image with the spatial resolution of the latter while preserving the spectral information of the multispectral image. In this paper we propose a variational Bayesian methodology for pansharpening. The proposed methodology uses the sensor characteristics to model the observation process and super-Gaussian sparse image priors on the expected characteristics of the pansharpened image. The pansharpened image, as well as all model and variational parameters, are estimated within the proposed methodology. Using real and synthetic data, the quality of the pansharpened images is assessed both visually and quantitatively and compared with other pansharpening methods. Theoretical and experimental results demonstrate the effectiveness, efficiency, and flexibility of the proposed formulation.

Keywords: Pansharpening; Variational Bayesian ; image fusion; super-Gaussians.

1. Introduction

Remote sensing sensors capture, simultaneously, a multispectral (MS) low resolution image along with a single band high resolution image of the same area, referred to as panchromatic (PAN) image. However, MS high resolution images are needed by many applications, such as land use and land cover analyses or change detection. Pansharpening is a technique that fuses the MS and PAN images into an MS high resolution image that has the spatial resolution of the PAN image and the spectral resolution of the MS one.

In this paper we formulate the pansharpening problem following the Bayesian framework. Within this framework, we use the sensor characteristics to model the observation process as a conditional probability distribution. The observation process describes both the MS high resolution image to MS low resolution image relationship and how the PAN image is obtained from the MS high resolution one. This probability distributions provides fidelity to the observed data in the pansharpened image reconstruction process. together with from fidelity to the data, Bayesian methods incorporate prior knowledge on the MS high resolution image in the form of prior probability distributions. Crisp images, such as high resolution MS images, are expected to have Super-Gaussian (SG) statistics while upsampled images suffer from blur that smooths out sharp gradients making them to become more Gaussian in their statistics [1]. Our goal is to integrate the sharp edges of the PAN image into the pansharpened image, leading to less Gaussian statistics which makes SG priors a suitable choice. SG priors have been successfully applied to other image processing tasks, such as compressed sensing [2], blind deconvolution [1,3] and blind color deconvolution [4] and so it is also expected to produce good results in pansharpening. However, the form of the SG prior does not allow us to obtain the posterior distribution in an analytical way, making full Bayesian inference intractable. Hence, in this paper, we

use the variational Bayesian inference to estimate the distribution of the pansharpened image as well as the model parameters from the MS low resolution and PAN images.

The rest of the paper is organized as follows: a categorization and short review of related pansharpening methods is presented in section 2. In section 3 the pansharpening problem is mathematically formulated. Following the Bayesian modelling and inference, in section 4 we propose a fully Bayesian method for the estimation of all the problem unknowns and model parameters. In section 5, the quality of the pansharpened images is assessed both visually and quantitatively and compared with other classic and state-of-the-art pansharpening methods. In section 6 we discuss the obtained results and finally, section 7 concludes the paper.

2. Related work

Early pansharpening techniques, such as in the Brovey method [5], substituted some bands for image visualization or performed simple arithmetic transformations. Other classical methods included the transformation of the MS image and the substitution of one of its components by the high spatial resolution PAN image. Examples of this strategy are PCA substitution [6], Brovey Transform [7] and IHS [8] methods. A review of those early methods, among others, can be found in [9].

Over the past 20 years, numerous methods have been presented and, in an attempt to bring some order to the diversity of approaches, different reviews, comparisons and classifications have been proposed in the literature (see, for instance, [10–17]) each one with different criteria and, therefore, with a different categorization. Nevertheless, in the last years there seems to be a consensus in three main categories, namely Component Substitution (CS), Multi-Resolution Analysis (MRA) and Variational Optimization (VO) [15–17]. Additionally, the increasing number of Deep Learning (DL) based pansharpening methods proposed in recent years can be regarded as a new category.

The Component Substitution (CS) category includes the most widely used pansharpening methods. CS methods [12] usually upsample the MS image to the size of the PAN image and transform it to another space that separates the spatial and spectral image components. Then, the transformed component containing the spatial information is substituted by the PAN image (possibly, after histogram matching). Finally, the backward transform is applied to obtain the pansharpened image. Examples of these methods include the already mentioned PCA substitution [6], IHS methods [8,18,19], the Gram-Schmidt (GS) methods [20] and Brovey transform [7]. In [21], the transformation is replaced by any weighted average of the MS bands. It is shown that this approach generalizes any CS image fusion method. Determination of the weights has been carried out in different ways. For instance, in [22] the weights are optimally estimated to minimize the mean squared error while in [23] they are set to the correlation coefficient between a single band low resolution image (obtained from the MS image) and each MS band. A local criterion, based on the belonging of a given pixel to a fuzzy cluster, was applied in [24] to estimate weights that are different for each pixel of the image. To obtain a crisper MS high-resolution image, in [25] a Wiener deconvolution of the upsampled MS bands was performed before fusion.

In general, CS-based methods produce spectral distortions due to the different statistics of the PAN image and the transformed component containing the spatial details. To tackle this issue, Multi-Resolution Analysis (MRA) methods decompose the MS and PAN images to different levels, extract spatial details from the decomposed PAN image, and inject them into the finer scales of the MS image. This principle is also known as the ARSIS concept [10]. The High-pass filtering (HPF) algorithm in [11,18], can be considered to be the first approach in this category where only two levels are considered. Multi-scale decompositions, such as the wavelet transform (WT) [26–28], the generalized Laplacian pyramid (GLP) [29–31] or the non-subsampled contourlet transform (NSCT) [32–34], were used to bring more precision to the methods. The “a trous” wavelet transform (AWT) was the preferred decomposition technique [26,28] until the publication of [31] showed the advantages of GLP over AWT. This was later corroborated in [14] where a comparison of different methods based on decimated and undecimated WT, AWT, GLP and NSCT concluded that GLP outperforms AWT

because it better removes aliasing. MRA category also includes the Smoothing Filter Based Intensity Modulation (SFIM) method [35,36], which first upsamples the MS image to the size of the PAN one and then uses a simplified solar radiation and land surface reflection model to increase its quality, and the Indusion method [37] in which upscaling and fusion steps are carried out together.

Deep Learning (DL) techniques have gained prominence in the last years and several methods have been proposed for pansharpening. As far as we know, the use of Deep Neural Networks (DNN) for pansharpening were first introduced in [38] where a Modified Sparse Denoising Autoencoder (MSDA) algorithm was proposed. For the same task, a Coupled Sparse Denoising Autoencoder (CSDA) was used in [39]. Convolutional neural networks were introduced in [40] and also used, for instance, in [41]. Instead of facing the difficult task of learning the whole image, residual networks [42,43] learn, from upsampled MS and PAN patches, only the details of the MS high-resolution image that are not already in the upsampled MS image and add them to it to obtain the pansharpened image. To adjust the size of the MS image to the size of the PAN one in a coarse-to-fine manner, two residual networks in cascade were set in the so called Progressive Cascade Deep Residual Network (PCDRN) [44]. In [45] a multi-scale approach is followed by learning a DNN to upsample each NSCT directional sub-band from the MS and PAN images. In general, the main weaknesses of the DL techniques are the high computational resources needed for training, the need of a huge amount of training data, which, in the case of pansharpening, might not be available, and the poor generalization to satellite images not used during training. The absence of ground-truth MS high-resolution images, needed for training these DL methods, is a problem pointed-out by [46] where a non-supervised generative adversarial network (Pan-GAN) was proposed. The GAN aims to generate pansharpened images that are consistent with the spectral information of the MS image while maintaining the spatial information of the PAN image. However, the generalization of this technique to satellite images different from the ones used for training is not clear. The adaptation of general image fusion methods, like the U2Fusion method in [47], to the pansharpening problem is a promising research area.

From a practical perspective, Variational Optimization (VO) based methods present advantages both from a theoretical as well as a computational points of view [48]. VO-based methods mathematically model the relation between the observed images and the original MS high resolution image, building an energy functional based on some desired properties of the original image. The pansharpened image is obtained as the image that minimizes this energy functional [49]. This mathematical formulation allows to rigorously introduce and process features that are visually important into the energy functional. Variational optimization can be considered as a particular case of the Bayesian approach [50], where the estimated image is obtained by maximizing the posterior probability distribution of the MS high resolution image. Bayesian methods for pansharpening formulate the relations between the observed images and the original MS high resolution image as probability distributions, model the desired properties as prior distributions and use Bayes' theory to estimate the pansharpened image based on the posterior distribution of the original MS high resolution image.

Following the seminal P+Xs method [51], the PAN image is usually modelled as a combination of the bands of the original high resolution multispectral image. However, in [49] this model was generalized by substituting the intensity images by their gradients. Note that while the P+Xs method [51] preserves spectral information, it produces blurring artifacts. To remove blur while preserving spectral similarity, other restrictions are introduced as reasonable assumptions or prior knowledge about the original image such as Laplacian prior [52], total variation [53,54], sparse representations [55], band correlations [56,57], non-local priors [58,59], etc. Spectral information is also preserved by enforcing the pansharpened image to be close to the observed MS one when downsampled to the size of the latter [52,60,61]. A special class of VO-based methods are the super-resolution methods which model pansharpening as the inverse problem of recovering the original high-resolution image by fusing the MS image and the PAN (see [52], and [62] for a recent review and [63] for a recent work). Deconvolution methods, such as [64], also try to solve the inverse problem but the upsampling of the

MS image to the size of the PAN one is performed prior to the pansharpening procedure. Registration and fusion are carried out simultaneously in [65].

Note that the variational Bayesian approach, also followed in this paper, is more general than variational optimization. While VO-based methods aim at obtaining a single estimate of the pansharpened image, the variational Bayesian approach estimates the whole posterior distribution of the pansharpened images and the model parameters, given the observations. When a single image is needed, the mode of the distribution is usually selected but other solutions can be obtained, for instance, by sampling the estimated distribution. Even more, the proposed approach allows us to simultaneously estimate the model parameters along with the pansharpened image using the same framework.

3. Problem Formulation

Let us denote by \mathbf{y} the MS high-resolution image hypothetically captured with an ideal high-resolution sensor with B bands \mathbf{y}_b , $b = 1, \dots, B$, of size $p = m \times n$ pixels, that is, $\mathbf{y} = [\mathbf{y}_1^T, \dots, \mathbf{y}_B^T]^T$, where the superscript T denotes the transpose of a vector or matrix. Note that each band of the image is flattened into a column vector containing its pixels in lexicographical order. Unfortunately, this high-resolution image is not available in real applications. Instead, we observe an MS low-resolution image $\mathbf{Y} = [\mathbf{Y}_1^T, \dots, \mathbf{Y}_B^T]^T$ with B bands \mathbf{Y}_b of size $P = M \times N$ pixels with $M < m$, $N < n$. The bands in this image are flattened as well to express them as a column vector. The relation between each low-resolution band, \mathbf{Y}_b , and its corresponding high-resolution one, \mathbf{y}_b , is defined by

$$\mathbf{Y}_b = \mathbf{D}\mathbf{H}\mathbf{y}_b + \mathbf{n}_b = \mathbf{B}\mathbf{y}_b + \mathbf{n}_b, \quad (1)$$

where \mathbf{D} is $P \times p$ decimation operator, \mathbf{H} is a $p \times p$ blurring matrix, $\mathbf{B} = \mathbf{D}\mathbf{H}$, and the capture noise \mathbf{n}_b is modeled as additive white Gaussian noise with variance β_b^{-1} .

A single band high resolution PAN image covering a wide range of frequencies is also provided by the sensor. This PAN image \mathbf{x} of size $p = m \times n$ is modelled as an spectral average of the unknown high-resolution bands \mathbf{y}_b , as

$$\mathbf{x} = \sum_{b=1}^B \lambda_b \mathbf{y}_b + \mathbf{v}, \quad (2)$$

where $\lambda_b > 0$ are known quantities that depend on each particular satellite sensor, and the capture noise \mathbf{v} is modeled as additive white Gaussian noise with variance γ^{-1} .

Once the image formation is formulated, let us use the Bayesian formulation to tackle the problem of recovering \mathbf{y} , the MS high resolution image, using the observed \mathbf{Y} , its degraded MS low resolution and PAN \mathbf{x} .

4. Bayesian Modelling and Inference

We model the distribution of each low resolution image \mathbf{Y}_b , $b = 1, \dots, B$, following the degradation model in (1) as a Gaussian distribution with mean $\mathbf{B}\mathbf{y}_b$ and covariance matrix $\beta_b^{-1}\mathbf{I}$. Then, the distribution of the observed image \mathbf{Y} is modelled by

$$p(\mathbf{Y}|\mathbf{y}, \boldsymbol{\beta}) = \prod_{b=1}^B \mathcal{N}(\mathbf{Y}_b | \mathbf{B}\mathbf{y}_b, \beta_b^{-1}\mathbf{I}), \quad (3)$$

with $\boldsymbol{\beta} = \{\beta_1, \dots, \beta_b\}$.

Analogously, using the degradation model in (2), the distribution of the PAN image \mathbf{x} is given by

$$p(\mathbf{x}|\mathbf{y}, \gamma) = \mathcal{N}(\mathbf{x} | \sum_{b=1}^B \lambda_b \mathbf{y}_b, \gamma^{-1}\mathbf{I}). \quad (4)$$

Table 1. Some possible penalty functions

Label	$\rho(s)$	$\rho'(s)/ s $
$\ell_p, 0 < p \leq 1$	$\frac{1}{p} s ^p$	$ s ^{p-2}$
log	$\log(\epsilon + s)$	$(\epsilon + s)^{-1} s ^{-1}$

The starting point for Bayesian methods is to choose a prior distribution for the unknowns. In this paper, we use SG distributions as priors for the MS high resolution image as

$$p(\mathbf{y}|\boldsymbol{\alpha}) = \prod_{b=1}^B \prod_{v=1}^J p(\mathbf{y}_{bv}|\alpha_{bv}) = \prod_{b=1}^B \prod_{v=1}^J \prod_{i=1}^p Z(\alpha_{bv}) \exp[-\alpha_{bv}\rho(y_{bv}(i))], \quad (5)$$

with $\alpha_{bv} > 0$ and $\boldsymbol{\alpha} = \{\alpha_{11} \dots, \alpha_{LB}\}$ and $Z(\alpha_{bv})$ is a partition function. In (5), $\mathbf{y}_{bv} = \mathbf{F}_v \mathbf{y}_b$ is a filtered version of the b -th band, \mathbf{y}_b , where $\{\mathbf{F}_v\}_{v=1}^J$ is a set of J high-pass filters, $y_{bv}(i)$ is the i -th pixel value of \mathbf{y}_{bv} , and $\rho(\cdot)$ is a penalty function. The image priors are placed on the filtered image \mathbf{y}_{bv} . It is well known that the application of high-pass filters to natural images returns sparse coefficients. Most of the coefficients are zero or close to zero while only the edge related coefficients remain large. Sparse priors enjoy SG properties, heavier tails, more peaked and positive excess kurtosis compared to the Gaussian distribution. The distribution mass is located around zero, but large values have a higher probability than in a Gaussian distribution. For $p(\mathbf{y}_{bv}|\alpha_{bv})$ in (5) to be SG, $\rho(\cdot)$ has to be symmetric around zero and the function $\rho(\sqrt{s})$ increasing and concave for $s \in (0, \infty)$. This condition is equivalent to $\rho'(s)/|s|$ being decreasing on $(0, \infty)$, and allows ρ to be represented as

$$\rho(y_{bv}(i)) = \inf_{\eta_{bv}(i) > 0} \frac{1}{2} \eta_{bv}(i) y_{bv}^2(i) - \rho^*\left(\frac{1}{2} \eta_{bv}(i)\right) \quad (6)$$

$$\Rightarrow \rho(y_{bv}(i)) \leq L(y_{bv}(i), \eta_{bv}(i)) = \frac{1}{2} \eta_{bv}(i) y_{bv}^2(i) - \rho^*\left(\frac{1}{2} \eta_{bv}(i)\right), \quad (7)$$

where \inf denotes the infimum, $\rho^*(\cdot)$ is the concave conjugate of $\rho(\cdot)$ and $\boldsymbol{\eta}_{bv} = \{\eta_{bv}(i)\}_{i=1}^p$ are a set of positive parameters. The relationship dual to (6) is given by [66]

$$\rho^*\left(\frac{1}{2} \eta_{bv}(i)\right) = \inf_{y_{bv}(i)} \frac{1}{2} \eta_{bv}(i) y_{bv}^2(i) - \rho(y_{bv}(i)). \quad (8)$$

To achieve sparsity, the function ρ should suppress most of the coefficients in \mathbf{y}_{bv} and preserve a small number of key features. Table 1 shows some penalty functions, corresponding to SG distributions (see [1]).

From (3), (4) and (5), the joint probability distribution $p(\boldsymbol{\Theta}, \mathbf{Y}, \mathbf{x})$, with $\boldsymbol{\Theta} = \{\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}\}$ the set of all unknowns, is given by

$$p(\boldsymbol{\Theta}, \mathbf{Y}, \mathbf{x}) = p(\mathbf{Y}|\mathbf{y}, \boldsymbol{\beta}) p(\mathbf{x}|\mathbf{y}, \boldsymbol{\gamma}) p(\mathbf{y}|\boldsymbol{\alpha}) p(\boldsymbol{\beta}) p(\boldsymbol{\gamma}) p(\boldsymbol{\alpha}), \quad (9)$$

where flat hyperpriors $p(\boldsymbol{\beta})$, $p(\boldsymbol{\gamma})$ and $p(\boldsymbol{\alpha})$ on the model hyperparameters have been included.

Following the Bayesian paradigm, inference will be based on $p(\boldsymbol{\Theta}|\mathbf{Y}, \mathbf{x})$. Since this posterior distribution cannot be analytically calculated due to the form of the SG distribution, in this paper we use the mean-field variational Bayesian model [67] to approximate $p(\boldsymbol{\Theta}|\mathbf{Y}, \mathbf{x})$ by the distribution $q(\boldsymbol{\Theta})$ of the form $q(\boldsymbol{\Theta}) = \prod_{\theta \in \boldsymbol{\Theta}} q(\theta)$, that minimizes the Kullback-Leibler divergence [68] defined as

$$\mathbf{KL}(q(\boldsymbol{\Theta}) || p(\boldsymbol{\Theta}|\mathbf{Y}, \mathbf{x})) = \int q(\boldsymbol{\Theta}) \log \frac{q(\boldsymbol{\Theta})}{p(\boldsymbol{\Theta}|\mathbf{Y}, \mathbf{x})} d\boldsymbol{\Theta} + \log p(\mathbf{Y}) + \log p(\mathbf{x}). \quad (10)$$

The Kullback-Leibler divergence is always non negative and it is equal to zero if and only if $q(\boldsymbol{\Theta}) = p(\boldsymbol{\Theta}|\mathbf{Y}, \mathbf{x})$.

Even with this factorization, the SG prior for \mathbf{y} hampers the evaluation of this divergence, but the quadratic bound for ρ in (7) allows us to bound the prior in (5) with a Gaussian form such that

$$p(y_{bv}(i)|\alpha_{bv}) \geq Z(\alpha_{bv}) \exp[-\alpha_{bv}L(y_{bv}(i), \eta_{bv}(i))], \quad \forall \eta_{bv}(i) > 0. \quad (11)$$

We then define the lower bound of the prior $\mathcal{M}_v(\mathbf{y}_v, \boldsymbol{\eta}_v | \boldsymbol{\alpha}_v) = \prod_b \mathcal{M}_{vb}(\mathbf{y}_{bv}, \boldsymbol{\eta}_{bv} | \boldsymbol{\alpha}_{bv})$ where

$$\mathcal{M}_{vb}(\mathbf{y}_{bv}, \boldsymbol{\eta}_{bv} | \boldsymbol{\alpha}_{bv}) = \prod_{i=1}^p Z(\alpha_{bv}) \exp[-\alpha_{bv}L(y_{bv}(i), \eta_{bv}(i))] \quad (12)$$

and obtain the lower bound of the joint probability distribution

$$F(\Theta, \mathbf{Y}, \mathbf{x}, \boldsymbol{\eta}) = p(\mathbf{Y} | \mathbf{y}, \boldsymbol{\beta}) p(\mathbf{x} | \mathbf{y}, \boldsymbol{\gamma}) \prod_{v=1}^J \mathcal{M}_v(\mathbf{y}_v, \boldsymbol{\eta}_v | \boldsymbol{\alpha}_v) \quad (13)$$

to obtain the inequality $\log p(\Theta, \mathbf{Y}, \mathbf{x}) \geq \log F(\Theta, \mathbf{Y}, \mathbf{x}, \boldsymbol{\eta})$.

Utilizing the lower bound $F(\Theta, \mathbf{Y}, \mathbf{x}, \boldsymbol{\eta})$ for the posterior probability distribution in (10) we minimize $\mathbf{KL}(q(\Theta) || F(\Theta, \mathbf{Y}, \mathbf{x}, \boldsymbol{\eta}))$ instead of $\mathbf{KL}(q(\Theta) || p(\Theta | \mathbf{Y}, \mathbf{x}))$.

As shown in [67], for each unknown $\theta \in \Theta$, the estimated $q(\theta)$ will have the form

$$q(\theta) \propto \exp \langle \log F(\Theta, \mathbf{Y}, \mathbf{x}, \boldsymbol{\eta}) \rangle_{q(\Theta \setminus \theta)}, \quad (14)$$

where $\Theta \setminus \theta$ represents all the variables in Θ except θ and $\langle \cdot \rangle_{q(\Theta \setminus \theta)}$ denotes the expected value calculated using the distribution $q(\Theta \setminus \theta)$. When point estimates are required $\hat{\theta} = \langle \theta \rangle_{q(\theta)}$ is used.

For variables with a degenerate posterior approximation, that is, for $\theta \in \{\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}\}$, the value where the posterior degenerates is given by [67]

$$\hat{\theta} = \arg \max_{\theta} \langle \log F(\Theta, \mathbf{Y}, \mathbf{x}, \boldsymbol{\eta}) \rangle_{q(\Theta \setminus \theta)}. \quad (15)$$

Let us now obtain the analytic expressions for each unknown posterior approximation.

4.1. High Resolution Multispectral Image Update

Using (14) we can show in a straightforward way that the posterior distribution for the high resolution MS image will have the form

$$q(\mathbf{y}) = \mathcal{N}(\mathbf{y} | \langle \mathbf{y} \rangle, \boldsymbol{\Sigma}_{\mathbf{y}}), \quad (16)$$

where the inverse of the covariance matrix is given by

$$\begin{aligned} \boldsymbol{\Sigma}_{\mathbf{y}}^{-1} = & \text{diag}(\boldsymbol{\beta}) \otimes \mathbf{B}^T \mathbf{B} + \gamma (\boldsymbol{\lambda} \boldsymbol{\lambda}^T) \otimes \mathbf{I}_{p \times p} \\ & + \sum_v \begin{pmatrix} \alpha_{1v} \mathbf{F}_v^T \text{diag}(\boldsymbol{\eta}_{1v}) \mathbf{F}_v & \mathbf{0}_{p \times p} & \dots & \mathbf{0}_{p \times p} \\ \mathbf{0}_{p \times p} & \alpha_{2v} \mathbf{F}_v^T \text{diag}(\boldsymbol{\eta}_{2v}) \mathbf{F}_v & \dots & \mathbf{0}_{p \times p} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{p \times p} & \mathbf{0}_{p \times p} & \dots & \alpha_{Bv} \mathbf{F}_v^T \text{diag}(\boldsymbol{\eta}_{Bv}) \mathbf{F}_v \end{pmatrix}, \end{aligned} \quad (17)$$

with \otimes denoting the Kronecker product, $\text{diag}(\cdot)$ is a diagonal matrix formed from the elements of a vector and the mean is obtained as

$$\boldsymbol{\Sigma}_{\mathbf{y}}^{-1} \langle \mathbf{y} \rangle = \left(\text{diag}(\boldsymbol{\beta}) \otimes \mathbf{B}^T \right) \mathbf{Y} + \gamma (\text{diag}(\boldsymbol{\lambda}) \otimes \mathbf{I}_{p \times p}) \left(\mathbf{x}^T, \mathbf{x}^T, \dots, \mathbf{x}^T \right)^T. \quad (18)$$

4.2. Variational Parameters Update

To estimate the value of the variational parameters, $\boldsymbol{\eta}$ introduced in (7), we need to solve, for each band $b \in \{1, \dots, B\}$, filter $\nu \in \{1, \dots, J\}$, and pixel $i \in \{1, \dots, p\}$, the optimization problem

$$\begin{aligned} \hat{\eta}_{b\nu}(i) &= \arg \min_{\eta_{b\nu}(i)} \langle L(y_{b\nu}(i), \eta_{b\nu}(i)) \rangle_{q(\mathbf{y})} \\ &= \arg \min_{\eta_{b\nu}(i)} \frac{1}{2} \eta_{b\nu}(i) u_{b\nu}^2(i) - \rho^* \left(\frac{1}{2} \eta_{b\nu}(i) \right), \end{aligned} \quad (19)$$

where $u_{b\nu}(i) = \sqrt{\langle y_{b\nu}^2(i) \rangle}$. Since

$$\rho^* \left(\frac{\hat{\eta}_{b\nu}(i)}{2} \right) = \min_x \frac{1}{2} \hat{\eta}_{b\nu}(i) x^2 - \rho(x) \quad (20)$$

whose minimum is achieved at $x = u_{b\nu}(i)$, we have, differentiating the right hand side of (19) with respect to x ,

$$\hat{\eta}_{b\nu}(i) = \rho'(u_{b\nu}(i)) / u_{b\nu}(i). \quad (21)$$

4.3. Model Parameters Update

The estimates of the noise variance in the degradation models in (3) and (4) are obtained using (15) as

$$\hat{\beta}_b^{-1} = \frac{\text{tr} \langle (\mathbf{Y}_b - \mathbf{B}\mathbf{y}_b)(\mathbf{Y}_b - \mathbf{B}\mathbf{y}_b)^T \rangle_{q(\Theta)}}{P}, \quad b = 1, \dots, B, \quad (22)$$

$$\hat{\gamma}^{-1} = \frac{\text{tr} \langle (\mathbf{x} - \sum_{b=1}^B \lambda_b \mathbf{y}_b)(\mathbf{x} - \sum_{b=1}^B \lambda_b \mathbf{y}_b)^T \rangle_{q(\Theta)}}{p}, \quad (23)$$

where $\text{tr}(\cdot)$ represents the trace of the matrix.

From (14) we obtain the following distribution for the parameter $\alpha_{b\nu}$ of the SG prior in (5).

$$q(\alpha_{b\nu}) = \text{const} + \sum_{i=1}^p \log Z(\alpha_{b\nu}) \exp[-\alpha_{b\nu} \rho(y_{b\nu}(i))]. \quad (24)$$

The mode of this distribution can be obtained (see [69]) by solving

$$\frac{\partial Z(\hat{\alpha}_{b\nu})}{\partial \hat{\alpha}_{b\nu}} = \frac{\text{tr} \langle (\mathbf{F}_\nu^T \mathbf{F}_\nu) \langle \mathbf{y}_{b\nu} \mathbf{y}_{b\nu}^T \rangle \rangle}{p}. \quad (25)$$

The ℓ_p penalty function shown in Table 1 produces proper priors, for which the partition function can be evaluated, but the log penalty function produces an improper prior. We tackle this problem examining, for $\alpha_{b\nu} \neq 1$, the behavior of

$$Z(\alpha_{b\nu}, K)^{-1} = \int_{-K}^K \exp[-\alpha_{b\nu} \rho(t)] dt \quad (26)$$

and keeping in $\partial Z(\alpha_{b\nu}) / \partial \alpha_{b\nu}$ the term that depends on $\alpha_{b\nu}$. This produces for the log prior

$$\frac{\partial Z(\hat{\alpha}_{b\nu})}{\partial \hat{\alpha}_{b\nu}} = (\hat{\alpha}_{b\nu} - 1)^{-1}. \quad (27)$$

4.4. Calculating the Covariance Matrices

The matrix $\Sigma_{\mathbf{y}}$ in (17) must be explicitly computed to find its trace and also to calculate $\hat{\eta}_{bv}(i)$. However, since its calculation is very intense, we propose the following approximation. We first approximate $\text{diag}(\boldsymbol{\eta}_{bv})$ using

$$\text{diag}(\boldsymbol{\eta}_{bv}) \approx z(\boldsymbol{\eta}_{bv})\mathbf{I}_{p \times p}, \quad (28)$$

where $z(\boldsymbol{\eta}_{bv})$ is calculated as the mean of the values in $\boldsymbol{\eta}_{bv}$.

We then use the approximation

$$\Sigma_{\mathbf{y}}^{-1} \approx \begin{pmatrix} \Sigma_{\mathbf{y}_1}^{-1} & \mathbf{0}_{p \times p} & \cdots & \mathbf{0}_{p \times p} \\ \mathbf{0}_{p \times p} & \Sigma_{\mathbf{y}_2}^{-1} & \cdots & \mathbf{0}_{p \times p} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{p \times p} & \mathbf{0}_{p \times p} & \cdots & \Sigma_{\mathbf{y}_B}^{-1} \end{pmatrix}$$

with

$$\Sigma_{\mathbf{y}_b}^{-1} \approx \beta_b \mathbf{B}^T \mathbf{B} + \gamma \lambda_b^2 \mathbf{I}_{p \times p} + \sum_v \alpha_{bv} z(\boldsymbol{\eta}_{bv}) \mathbf{F}_v^T \mathbf{F}_v = \mathbf{C}_b, \quad b = 1, \dots, B.$$

Finally we have

$$\langle y_{bv}^2(i) \rangle \approx (\langle y_{bv}(i) \rangle)^2 + \frac{1}{p} \text{tr} [\mathbf{C}_b^{-1} \mathbf{F}_v^T \mathbf{F}_v].$$

4.5. Proposed Algorithm

Based on the previous derivations, we propose the Variational Bayesian SG Pansharpening Algorithm in Alg. 1. The linear equations problem in (18), used in step 4 of Alg. 1, has been solved using the Conjugate Gradient approach.

Algorithm 1 Variational Bayesian SG Pansharpening

Require: Observed multispectral image, \mathbf{Y} , panchromatic image \mathbf{x} , and λ parameter.

Set $\Sigma_{\mathbf{y}}^{(0)} = \mathbf{0}$ and $n = 0$. $\langle \mathbf{y}_b \rangle^{(0)}$ is obtained by bicubic interpolation of \mathbf{Y}_b , $\forall b = 1, \dots, B$.

while convergence criterion is not met **do**

1. Set $n = n + 1$.
2. Obtain $\beta^{(n)}$, $\gamma^{(n)}$ and $\alpha_{bv}^{(n)}$ from (22), (23) and (25) respectively.
3. Using $\langle \mathbf{y} \rangle^{(n-1)}$ and $\Sigma_{\mathbf{y}}^{(n-1)}$, update the variational parameters $\hat{\eta}_{bv}^{(n)}$, $\forall b, v$ from (21).
4. Using $\beta^{(n)}$, $\gamma^{(n)}$, $\alpha_{bv}^{(n)}$, and $\hat{\eta}_{bv}^{(n)}$, update $\Sigma_{\mathbf{y}}^{-1(n)}$ in (17) and solve (18) for $\langle \mathbf{y} \rangle^{(n)}$.

end while

Output the high resolution hyperspectral image $\hat{\mathbf{y}} = \langle \mathbf{y} \rangle^{(n)}$.

5. Materials and Methods

To test the performance of the proposed methodology on different kind of images, five satellite images have been used: three LANDSAT 7-ETM+ [70] images, a SPOT-5 [71] image and a FORMOSAT-2 [72] image. LANDSAT MS images have six bands and a ratio between PAN and MS images $p/P = 4$. Figures 1 and 2 show RGB color images formed by the bands B4, B3 and B2 of LANDSAT MS images, and their corresponding PAN images. Figure 1 corresponds to an area from Chesapeake Bay (US) while Fig. 2 depicts two areas from Neatherland. SPOT-5 MS images have four bands and two PAN

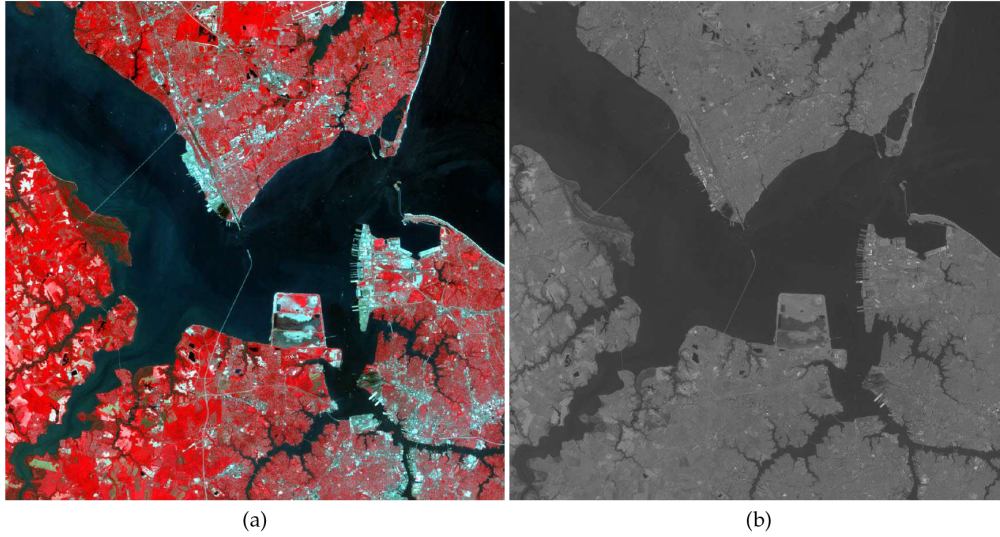


Figure 1. Observed LANDSAT 7-ETM+ Chesapeake Bay image: (a) 1024×1024 MS, (b) 2048×2048 PAN.

images, with resolution ratios of $p/P = 4$ and $p/P = 16$, are provided. FORMOSAT-2 MS images also have four bands and a ratio between PAN and MS images $p/P = 16$. Figures 3(a) & (c) show the RGB color images formed from bands B3, B2 and B1 bands of a SPOT-5 image from Roma (IT) and a FORMOSAT-2 MS image from Salon-de-Provence (FR) and Figures 3(b) & (d) their corresponding PAN images.

Both the observed \mathbf{Y} and \mathbf{x} images have been normalized to the range $[0, 1]$ before running Algorithm 1. The convergence criterion in the algorithm was $\| \langle \mathbf{y} \rangle^{(n)} - \langle \mathbf{y} \rangle^{(n-1)} \|^2 / \| \langle \mathbf{y} \rangle^{(n)} \|^2 \leq 10^{-6}$ or 50 iterations were reached, whatever occurs first. The relationship between the MS high resolution image and the panchromatic image in (2) is governed by the parameters λ that need to be set before pansharpening is carried out. If we knew the sensor spectral response characteristics, the values of λ could be estimated from them. For instance, for LANDSAT 7-ETM+, Fig. 4 shows the sensor spectral response curves for the MS bands B1-B6, shown in color, and the PAN band shown in black. For this sensor, the PAN band mainly overlaps B2-B4 MS bands, and λ coefficients could be obtained from this overlapping (see [52]). In this paper, however, a more general approach is followed to estimate λ from the observations. First, we define $\mathbf{X} = \mathbf{D}\mathbf{x}$, a version of the PAN image downsampled to the size of the MS image. Then, since the sensor spectral response is the same in high and low resolution, the parameters λ can be obtained by solving

$$\lambda = \underset{\lambda}{\operatorname{argmin}} \left\| \mathbf{X} - \sum_{b=1}^B \lambda_b \mathbf{Y}_b \right\|^2, \quad (29)$$

$$\text{subject to } \lambda_b \geq 0, \forall b, \sum_{b=1}^B \lambda_b = 1. \quad (30)$$

Table 2 shows the λ s associated to the different considered observed images. For the LANDSAT 7-ETM+ images only the first four bands are positive and λ_5 and λ_6 values are 0 since we know that bands B5 and B6 are not covered by the panchromatic sensor. For this process each band is normalized to the interval $[0, 1]$. Note that due to the normalization, the estimated λ values do not only depend on the sensor spectral response but also on the observed area characteristics. This explains the differences between the obtained λ values for the images in Figures 2(a) and 2(c). Although those images are from the same area of Netherlands, clouds in Figure 2(a) modify the estimation of the values of λ .

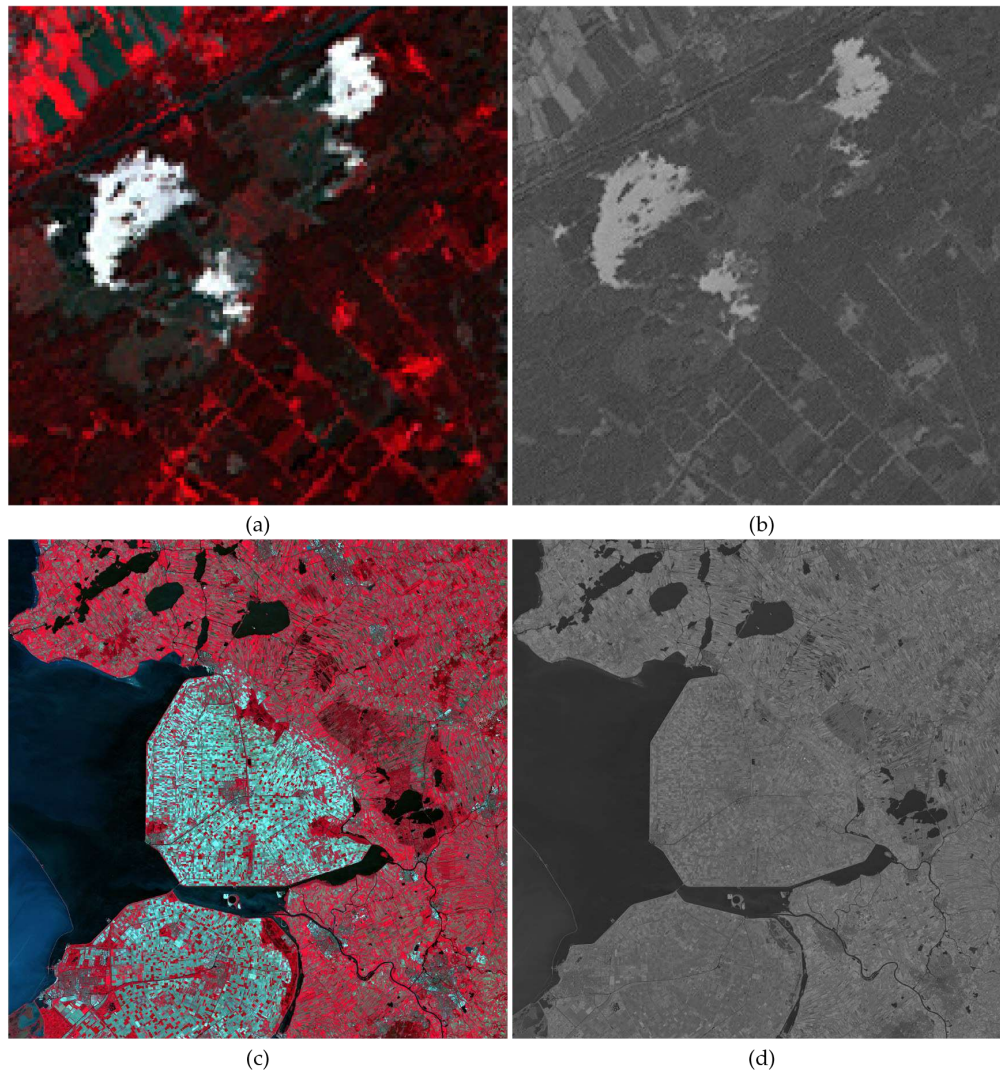


Figure 2. Observed LANDSAT 7-ETM+ Netherland images: (a) 128×128 MS, (b) 256×256 PAN, (c) 2048×2048 MS, (d) 4096×4096 PAN.

Table 2. Estimated λ values for the different sensors.

Sensor	Image	B1	B2	B3	B4	B5	B6
LANDSAT 7-ETM+	Fig. 1	0.0986	0.1011	0.2576	0.5427	0	0
LANDSAT 7-ETM+	Fig. 2(a)	0.0183	0.4243	0.0576	0.4998	0	0
LANDSAT 7-ETM+	Fig. 2(c)	0	0.2283	0.1611	0.6106	0	0
SPOT-5	Fig. 3(a)	0	0.2993	0.6897	0.0110	-	-
FORMOSAT-2	Fig. 3(c)	0.0384	0.5566	0	0.4051	-	-

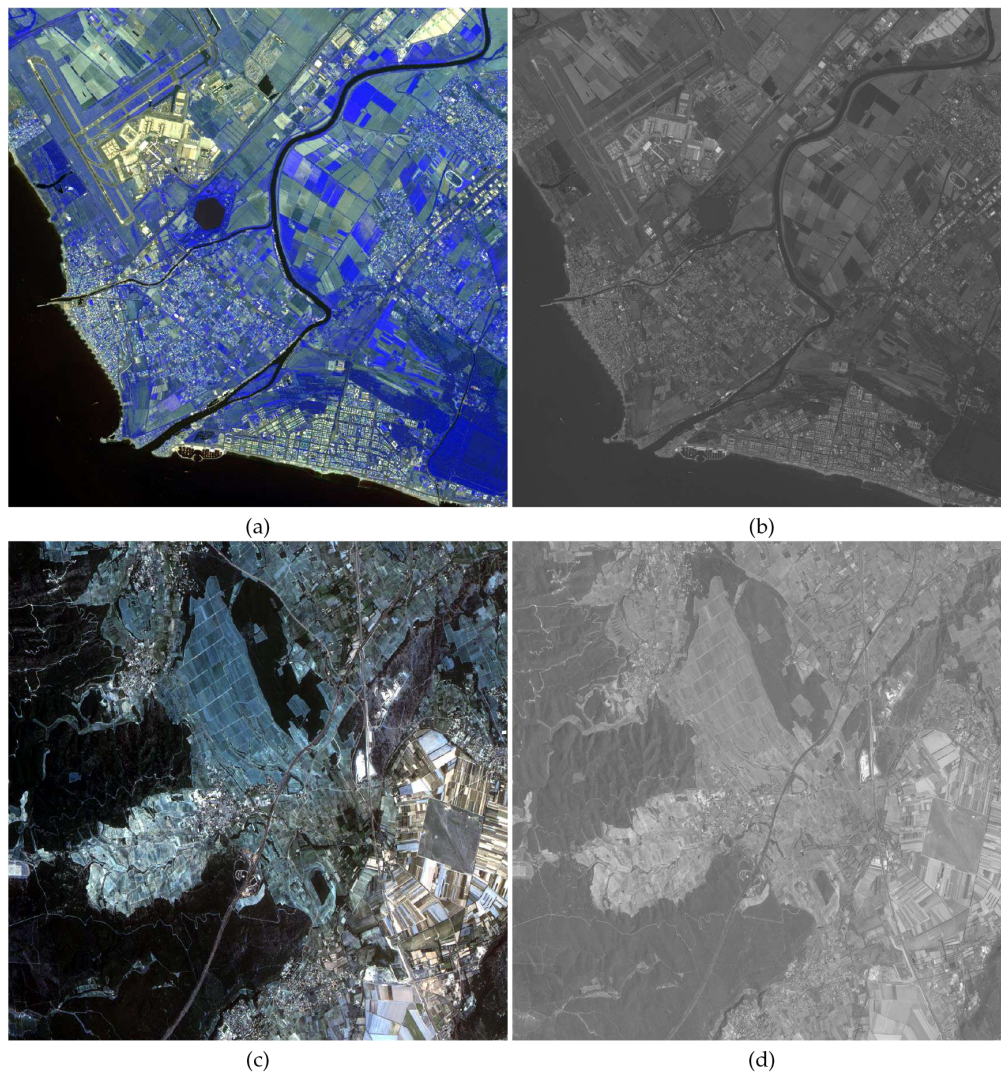


Figure 3. Observed SPOT-5 Roma image: (a) 1024×1024 MS, (b) 4096×4096 PAN. FORMOSAT-2 Salon-de-Provence image: (c) 1024×1024 MS, (d) 4096×4096 PAN.

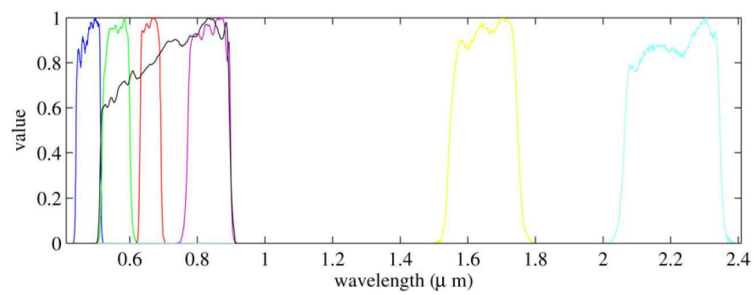


Figure 4. LANDSAT 7-ETM+ band spectral response normalized to one.

6. Discussion

Within the variational Bayesian methodology, two methods are proposed in this paper: one using the log penalty function (see Table 1), hence, named log method, and another using the ℓ_p penalty function, with $p = 1$, referred as ℓ_1 method. The proposed methods have been compared with the following classic and state-of-the-art pansharpening methods: the Principal Component Analysis (PCA) [6], the Intensity-Hue-Saturation (IHS) transform [19], the Brovey transform (Brovey) [7], the Band-Dependent Spatial-Detail (BDSD) method in [22], the Gram-Schmidt (GS) method in [20], the Gram-Schmidt adaptive (GSA) method in [21], the Partial Replacement Adaptive Component Substitution (PRACS) method in [23], the High Pass Filtering (HPF) algorithm in [18], the Smoothing Filter Based Intensity Modulation (SFIM) method [35,36], the Indusion method in [37], the Additive A Trouis Wavelet Transform (ATWT) in [26], the Additive Wavelet Luminance Proportional (AWLP) method in [28], the ATWT Model 2 (ATWT-M2) and ATWT Model 3 (ATWT-M3) methods in [10], the Generalized Laplacian Pyramid (GLP) based methods in [29], concretely the modulation transfer functions MTF-GLP, GLP with High Pass Modulation (MTF-GLP-HPM), and GLP with Context Based Decision (MTF-GLP-CBD) methods, and the pansharpening method using a Total Variation (TV) image model in [53]. We have used the implementation of the methods and measures provided by the Pansharpening Toolbox¹ [13]. For those methods not included in the toolbox we have used the code provided by the authors. The code of the proposed methods will be publicly available at <https://github.com/vipgugr>. We have also included the results of bilinear interpolating the MS image to the size of the PAN, marked as EXP, as a reference. Both quantitative and qualitative comparisons of the different methods have been performed.

6.1. Quantitative comparison

A common problem in pansharpening is the nonexistence of a MS high resolution ground-truth image to compare with. Hence we performed two kinds of quantitative comparisons. Firstly, the images obtained using the different methods have been compared following the Wald's protocol [73] as follows: the observed MS image, \mathbf{Y} , and the PAN image, \mathbf{x} , are downsampled by applying the operator \mathbf{D} to generate low resolution versions of them. Then, pansharpening is applied to those low resolution images and the obtained estimation of the MS image, $\hat{\mathbf{y}}$, is quantitatively compared with the observed MS image, \mathbf{Y} . Secondly, the different methods have been compared using Quality with No Reference (QNR) measures [13,74]. As previously stated, for the LANDSAT image in Fig. 1, the resolution ratio between MS and PAN images is $p/P = 4$. Since the SPOT-5 satellite provides two PAN images, two experiments were carried out on the image in Fig. 3, one with a decimation ratio of 4 and another with a ratio of 16. For the FORMOSAT-2 image the ratio is $p/P = 16$. However, for the sake of completeness, two experiments were also carried out, one assuming a decimation ratio of 4 and another with a ratio of 16.

Both spatial and spectral quality metrics have been used to compare the results obtained using the different methods. Details for the metrics used is shown below:

Spatial measures:

- Q
 - Universal Quality Index (UQI) [75] averaged on all MS bands.
 - Range: [-1, 1]
 - The higher the the better.
- Q4, Q8
 - Instances of the $Q2^n$ [76] index taking values. Suitable to measure quality for multiband images having an arbitrary number of spectral bands. Q4 is used for SPOT-5 and

¹ <https://rscl-grss.org/coderecord.php?id=541>

FORMOSAT-2 images which have four bands and Q8 for the LANDSAT image with 6 bands.

- Range: $[0, 1]$
- The higher the better.
- Spatial Correlation Coefficient (SCC) [77]
 - Measures the correlation coefficient between compared images after the application of a Sobel filter.
 - Range: $[0, 1]$
 - The higher the better.
- QNR spatial distortion (D_S) [78]
 - Measures the spatial distortion between MS bands and PAN image.
 - Range: $[0, 1]$
 - The lower the better.

Spectral measures:

- Spectral Angle Mapper (SAM) [79]
 - For spectral fidelity. Measures the mean angle between the corresponding pixels of the compared images in the space defined by considering each spectral band as a coordinate axis
 - Range: $[0, 180]$
 - The lower the better.
- Erreur Relative Globale Adimensionnelle de Synthese (ERGAS) [80]
 - Measures spectral consistency between compared images.
 - Range: $[0, \infty[$
 - The lower ERGAS value the better consistency, specially for values lower than the number of image bands B .
- QNR spectral distortion (D_λ) [78]
 - This measure is derived from the differences between the inter-band Q index values computed for HR and LR images.
 - Range: $[0, 1]$
 - The lower the better.

Spatial and spectral measures:

- Jointly Spectral and Spatial Quality Index (QNR) [78]
 - QNR is obtained as the product of $(1 - D_S)$ and $(1 - D_\lambda)$.
 - Range: $[0, 1]$
 - The higher the better.

Table 3 shows the obtained figures of merit using the Wald's protocol for the LANDSAT image in Fig. 1. As it is clear from the table, $\ell 1$ outperforms all the other methods both in spectral fidelity and the incorporation of spatial details. Note the high SCC value (meaning that the details in the PAN image have been successfully incorporated into the pansharpened image) while also obtaining the lowest spectral distortion as evidenced by the SAM and ERGAS values. The TV method obtains the second best results except for the SAM metric, for this metric the proposed log method has the second best value. This method also obtains the third best values for ERGAS and SCC measures. GLP based and PRACS methods also obtain high values for the Q, Q8 indices and low value for SAM. However, their ERGAS and SCC performance is worse. Table 4 shows the QNR quantitative results for the LANDSAT image in Fig. 1. In this table, the proposed methods achieve competitive results. Log obtains the best D_λ value and this method together with $\ell 1$ obtain second and third QNR scores, respectively. Note

that EXP obtained the highest score using QNR since bilinear interpolation of the observed MS low resolution image is used as the MS high resolution estimation to calculate D_S and D_λ calculations.

Tables 5 and 7 show the quantitative results using the Wald's protocol for the LANDSAT images in Figures 2(a) and 2(c), respectively. PRACS outperforms all other methods on the image in Fig. 2(a) (see Table 5) and the proposed ℓ_1 and log obtain the first and second best scores on the image in Fig. 2(c) (see Table 7). Tables 6 and 8 show the obtained QNR figures of merit for those two images. The proposed methods produce good D_S , D_λ and QNR values for both images, both above 0.9 which supports their good performance. Again the EXP results are the best in all the measures for Table 8 and provides the best D_λ , for the image associated to Table 6. The ℓ_1 method obtains the best D_S for this image and BSD the highest QNR.

Figures 5 and 6 show a zoomed in region of the RGB color images formed by bands B4, B3, and B2 of MS ground truth images used to apply the Wald's protocol and also the absolute error images for the methods in Tables 6 and 8. In those images, the darker the intensity the lower the absolute error. Figures 5 and 6 are consistent with the quantitative comparison shown in Tables 5 and 7, respectively. The best results for the image in Fig. 2(a) were obtained using PRACS while for the image in Fig. 2(c) the best performing method is ℓ_1 . Note that brighter areas in Figures 5(e) and 5(f) correspond to the borders of cloudy areas in Fig. 2(a). We argue that since clouds alter the weights of λ estimated using (30), the boundaries of clouds and land areas in Fig. 2(a) are not well resolved. This explains a worse behavior of the proposed methods in the cloudy areas of this image.

Tables 9 and 11 show, respectively, the quantitative results using the Wald's protocol for the SPOT-5 and the FORMOSAT-2 images in Fig. 3 for the decimation ratios $p/P = 4$ and $p/P = 16$. The proposed log obtains the best figures of merit for the SPOT image in Fig. 3(a) with $p/P = 4$ except for Q and Q4 metrics. The Q values obtained by log and ℓ_1 are slightly lower than those obtained by BSD. Note that BSD achieved the third best general figures just below the proposed log and ℓ_1 algorithms. With $p/P = 16$ the proposed log algorithm provides the best results except for Q, Q4 and SAM values, where competitive values are obtained. The proposed log achieves a slightly lower Q value than PRACS and a slightly higher SAM value than Brovey. In general, PRACS is the second best performing method for this image for $p/P = 16$. For the FORMOSAT-2 image in Fig. 3(c), the proposed ℓ_1 and log algorithms obtained the best numerical results for a $p/P = 4$ magnification. Both methods provide similar results, which are better than all the one provided by the competing methods. For a ratio $p/P = 16$, there is not a clear winner. The proposed methods are competitive in this image although they do not stand out in any of the measures. Tables 10 and 12 show, respectively, the QNR quantitative results for the SPOT-5 and the FORMOSAT-2 images in Fig. 3 for the decimation ratios $p/P = 4$ and $p/P = 16$. In Table 10, EXP achieves the best D_S , D_λ and QNR scores. In this table, the proposed methods obtain good scores. The log method obtains the second best D_λ and D_S values and very high QNR values for both decimation ratios. Results for the FORMOSAT image, shown in Table 12, are very similar although in this case, BSD obtains the best D_S and QNR values for $p/P = 4$ and ATWT-M3 for $p/P = 16$.

Table 13 shows the required CPU time in seconds on a 2.40GHz Intel[®] Xeon[®] CPU for the pansharpening of a MS image with 4 bands to a 1024×1024 size, for $p/P = 4$ and $p/P = 16$, using the different methods under comparison. Equation (18) has been solved using the Conjugate Gradient method which required, to achieve convergence, less than 30 iterations for the ℓ_1 prior and at least 1000 iterations for the log prior. This explains the differences between their required CPU time. Note that the proposed methods automatically estimate the model parameters which increases the running time but makes our methods parameter free.

6.2. Qualitative comparison

Figure 7 shows a small region of interest of the observed LANDSAT-7 images in Fig. 1 and the pansharpening results with $p/P = 4$ obtained by the proposed methods and the competing ones with the best quantitative performance, that is, PRACS, MTF-GLP-HPM, MTF-GLP-CBD and TV methods.

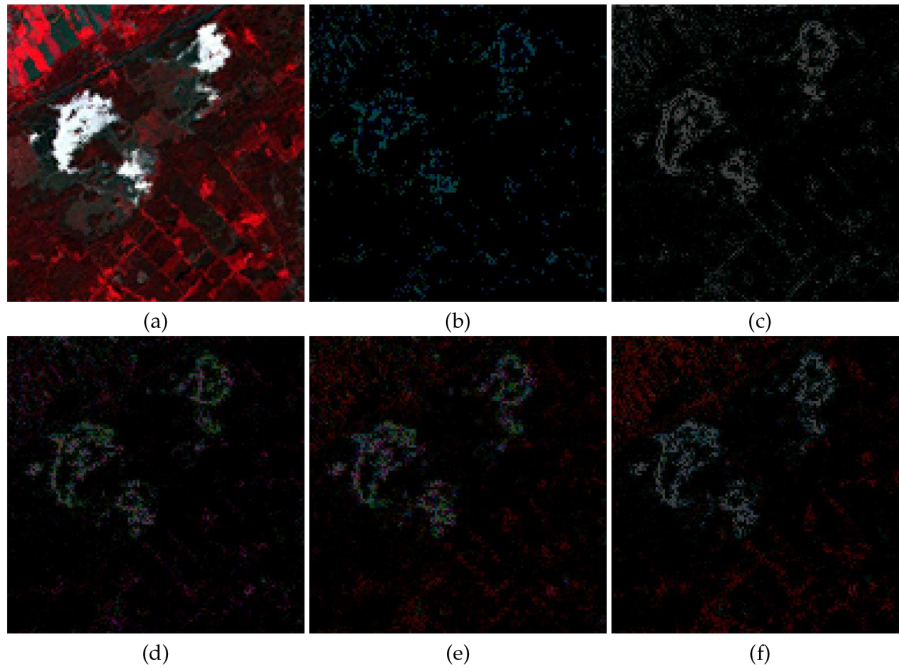


Figure 5. (a) Ground truth 128×128 image from Fig. 2(a). The normalized maximum absolute error minus the absolute error images for the following methods: (b) PRACS, (c) MTF-GLP-CBD, (d) TV, (e) ℓ_1 and (f) log.

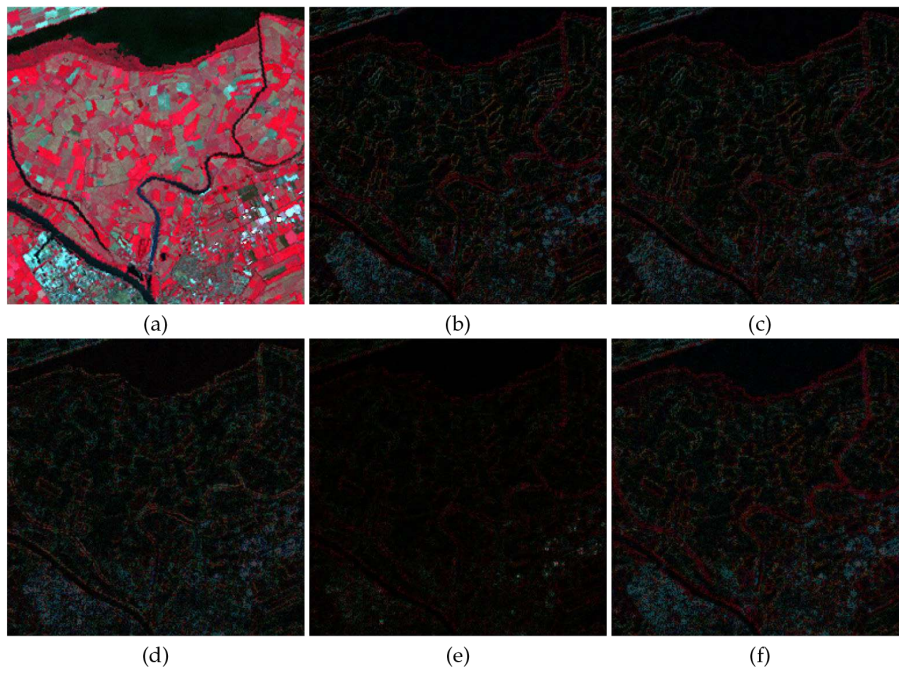


Figure 6. (a) Ground truth 256×256 image from Fig. 2(c). The normalized maximum absolute error minus the absolute error images for the following methods: (b) PRACS, (c) MTF-GLP-CBD, (d) TV, (e) ℓ_1 and (f) log.

Table 3. Quantitative results on the LANDSAT 7-ETM+ Chesapeake Bay image.

Method	Q8	Q	SAM	ERGAS	SCC
EXP	0.8335	0.8383	2.0223	5.1113	0.8718
PCA	0.7830	0.8074	2.7937	5.8086	0.8569
IHS	0.6795	0.6734	2.6640	7.8586	0.8223
Brovey	0.6798	0.6790	2.1605	7.5226	0.8148
BDSB	0.7599	0.7829	2.4662	5.9244	0.8539
GS	0.7447	0.7401	3.3052	8.7258	0.8040
GSA	0.8010	0.8202	2.4033	5.0688	0.8764
PRACS	0.8363	0.8423	2.0998	4.8655	0.8774
HPF	0.8137	0.8243	2.2221	5.1279	0.8834
SFIM	0.8008	0.8184	2.1857	5.0228	0.8905
Indusion	0.8052	0.8167	2.2906	5.5412	0.8495
ATWT	0.8100	0.8208	2.3482	5.3642	0.8809
AWLP	0.8267	0.8315	2.2516	5.3873	0.8734
ATWT-M2	0.7737	0.7802	2.6004	5.8133	0.8317
ATWT-M3	0.7884	0.7925	2.5720	5.8973	0.8367
MTF-GLP	0.8157	0.8258	2.2418	5.1307	0.8838
MTF-GLP-HPM	0.8037	0.8206	2.1878	5.0236	0.8918
MTF-GLP-CBD	0.8200	0.8292	2.2499	5.1015	0.8809
TV	0.8492	0.8606	2.1362	4.2505	0.9163
ℓ_1	0.8595	0.8694	1.8518	4.0954	0.9220
log	0.7951	0.7856	1.8839	4.4819	0.9007

Table 4. QNR Quantitative results on the LANDSAT 7-ETM+ Chesapeake Bay image.

	D_A	D_S	QNR
EXP	0.0096	0.0166	0.9740
PCA	0.0555	0.1459	0.8066
IHS	0.1049	0.2888	0.6366
Brovey	0.0963	0.2290	0.6968
BDSB	0.1055	0.1486	0.7616
GS	0.0716	0.1992	0.7435
GSA	0.0605	0.1167	0.8298
PRACS	0.0154	0.0823	0.9035
HPF	0.0857	0.1308	0.7947
SFIM	0.0834	0.1146	0.8116
Indusion	0.0595	0.0525	0.8911
ATWT	0.1044	0.1500	0.7612
AWLP	0.1045	0.1517	0.7596
ATWT-M2	0.1530	0.2327	0.6499
ATWT-M3	0.1227	0.1975	0.7041
MTF-GLP	0.0917	0.1378	0.7831
MTF-GLP-HPM	0.0890	0.1215	0.8004
MTF-GLP-CBD	0.0585	0.1081	0.8397
TV	0.0425	0.0919	0.8695
l1	0.0338	0.0527	0.9153
log	0.0090	0.0364	0.9549

Table 5. Quantitative results using Wald's protocol on the LANDSAT 7-ETM+ Netherland image in Fig. 2(a).

	Q8	Q	SAM	ERGAS	SCC
EXP	0.4727	0.8849	3.0445	7.7235	0.8686
PCA	0.3759	0.7617	3.8165	12.6420	0.8205
IHS	0.3322	0.7479	1.7633	10.7598	0.8570
Brovey	0.2892	0.7675	0.0000	10.7809	0.8492
BDSB	0.7205	0.9520	1.7296	4.6748	0.9777
GS	0.3860	0.7833	3.3486	11.7850	0.8323
GSA	0.5543	0.8474	2.5086	10.2990	0.8707
PRACS	0.8230	0.9720	0.9558	2.9097	0.9878
HPF	0.6420	0.9045	2.0699	7.4387	0.9458
SFIM	0.5950	0.9043	1.8898	8.1778	0.9379
Indusion	0.4108	0.8406	3.6963	9.6521	0.8269
ATWT	0.5582	0.8741	2.6859	9.6507	0.9267
AWLP	0.4715	0.8741	2.2059	10.1057	0.9195
ATWT-M2	0.3943	0.8436	3.7879	8.7289	0.8606
ATWT-M3	0.4861	0.8685	3.5274	7.7506	0.8829
MTF-GLP	0.5975	0.8946	2.2544	8.1279	0.9351
MTF-GLP-HPM	0.5659	0.8926	2.0133	9.0221	0.9272
MTF-GLP-CBD	0.6095	0.9001	2.1726	7.8290	0.9392
TV	0.4798	0.8906	3.4873	7.1207	0.8977
l1	0.4815	0.9044	3.3783	6.9422	0.9022
log	0.4931	0.8920	2.9187	6.9321	0.8952

Table 6. QNR quantitative results on the LANDSAT 7-ETM+ Netherland image in Fig. 2(a).

	D_λ	D_S	QNR
EXP	0.0104	0.0593	0.9309
PCA	0.2463	0.3998	0.4523
IHS	0.2632	0.4035	0.4394
Brovey	0.2182	0.3873	0.4790
BDSB	0.0159	0.0505	0.9344
GS	0.2335	0.4067	0.4548
GSA	0.2139	0.3240	0.5314
PRACS	0.0665	0.2106	0.7369
HPF	0.1711	0.2638	0.6102
SFIM	0.1610	0.2513	0.6282
Indusion	0.1354	0.1612	0.7252
ATWT	0.1968	0.2961	0.5654
AWLP	0.1977	0.2954	0.5653
ATWT-M2	0.1727	0.3127	0.5686
ATWT-M3	0.1187	0.2143	0.6924
MTF-GLP	0.1796	0.2791	0.5915
MTF-GLP-HPM	0.1688	0.2661	0.6100
MTF-GLP-CBD	0.1741	0.2778	0.5965
TV	0.0912	0.1127	0.8064
l1	0.0342	0.0386	0.9286
log	0.0157	0.0627	0.9225

Table 7. Quantitative results using Wald's protocol on the LANDSAT 7-ETM+ Netherland image in Fig. 2(c).

	Q8	Q	SAM	ERGAS	SCC
EXP	0.7874	0.7867	3.3362	5.9279	0.8521
PCA	0.6167	0.4854	5.7078	12.8127	0.5788
IHS	0.5343	0.3778	4.7417	12.9526	0.5918
Brovey	0.5465	0.4063	3.4605	13.0876	0.5960
BDSB	0.7575	0.7733	3.9348	6.8610	0.7856
GS	0.5899	0.4369	6.1158	13.9275	0.5522
GSA	0.7323	0.7444	4.0920	7.2078	0.7535
PRACS	0.7822	0.7829	3.4787	6.3244	0.8037
HPF	0.7167	0.7241	3.8135	7.3266	0.7603
SFIM	0.6908	0.7181	4.6693	9.0262	0.7261
Indusion	0.7230	0.7346	3.8229	7.2177	0.7631
ATWT	0.6948	0.6971	4.0402	8.0306	0.7240
AWLP	0.7110	0.7066	3.8124	8.1733	0.7165
ATWT-M2	0.6332	0.6288	4.5223	8.5542	0.6135
ATWT-M3	0.6993	0.6967	4.3028	7.4420	0.7201
MTF-GLP	0.7116	0.7172	3.8632	7.5134	0.7457
MTF-GLP-HPM	0.6875	0.7133	4.6780	9.1110	0.7161
MTF-GLP-CBD	0.7519	0.7595	3.7522	6.8698	0.7740
TV	0.7843	0.8065	3.8402	5.7351	0.8519
l1	0.8118	0.8196	3.1337	5.1831	0.8853
log	0.7750	0.7682	3.0810	5.3115	0.8818

Table 8. QNR quantitative results on the LANDSAT 7-ETM+ Netherland image in Fig. 2(c).

	D_λ	D_S	QNR
EXP	0.0067	0.0157	0.9778
PCA	0.1893	0.5019	0.4038
IHS	0.2040	0.5973	0.3205
Brovey	0.1979	0.5227	0.3829
BDSB	0.0134	0.0107	0.9761
GS	0.1999	0.5237	0.3811
GSA	0.0986	0.2337	0.6907
PRACS	0.0255	0.1176	0.8599
HPF	0.1594	0.2772	0.6075
SFIM	0.1124	0.2272	0.6859
Indusion	0.1352	0.2054	0.6871
ATWT	0.1822	0.3234	0.5534
AWLP	0.1765	0.3101	0.5682
ATWT-M2	0.1903	0.3467	0.5290
ATWT-M3	0.0917	0.1581	0.7647
MTF-GLP	0.1656	0.2933	0.5897
MTF-GLP-HPM	0.1164	0.2424	0.6694
MTF-GLP-CBD	0.0854	0.1953	0.7359
TV	0.0381	0.1186	0.8478
l1	0.0228	0.0704	0.9084
log	0.0073	0.0199	0.9730

Table 9. Quantitative results using Wald's protocol on the SPOT-5 Roma image.

p/P	4					16				
	Q4	Q	SAM	ERGAS	SCC	Q4	Q	SAM	ERGAS	SCC
EXP	0.8766	0.8859	1.7048	3.7857	0.8640	0.7325	0.7407	2.5071	2.8441	0.6049
PCA	0.4067	0.5360	5.1646	12.3346	0.2788	0.3927	0.5091	5.7208	6.2911	0.2443
IHS	0.4072	0.5238	3.9951	12.2342	0.2772	0.3973	0.5051	4.4508	6.1770	0.2520
Brovoy	0.4124	0.5337	1.8413	12.1960	0.2594	0.4019	0.5124	2.4000	6.1718	0.2482
BDSB	0.8559	0.8825	2.0565	4.2776	0.8235	0.5947	0.6231	3.0328	4.3050	0.2804
GS	0.4102	0.5364	5.0523	12.1272	0.2634	0.3985	0.5124	5.5849	6.1471	0.2405
GSA	0.4897	0.5384	2.9893	11.0363	0.1932	0.4997	0.5354	3.1189	5.3703	0.2164
PRACS	0.8220	0.8380	2.0536	4.8936	0.7298	0.7291	0.7458	2.5190	2.9647	0.5295
HPF	0.7488	0.7695	2.0632	6.3388	0.6250	0.5888	0.6124	2.8370	4.5969	0.2988
SFIM	0.7744	0.7860	1.9658	6.0694	0.6447	0.6052	0.6232	2.6792	4.4425	0.3079
Indusion	0.7473	0.7894	2.1609	5.9717	0.6761	0.5301	0.5935	3.5689	4.9656	0.3408
ATWT	0.6928	0.7171	2.2358	7.6433	0.5183	0.5849	0.6092	2.8932	4.7245	0.3099
AWLP	0.7066	0.7260	2.1798	7.6246	0.5075	0.5947	0.6173	2.8504	4.6963	0.3052
ATWT-M2	0.7229	0.7343	2.3084	6.3145	0.4678	0.6723	0.6822	2.7291	3.3277	0.3965
ATWT-M3	0.7837	0.7930	2.2394	5.1935	0.6714	0.6924	0.7019	2.7026	3.1094	0.4801
MTF-GLP	0.7289	0.7507	2.1201	6.7975	0.5766	0.5775	0.6023	2.9275	4.8355	0.3022
MTF-GLP-HPM	0.7553	0.7675	2.0005	6.4923	0.5972	0.5928	0.6111	2.7388	4.7066	0.3073
MTF-GLP-CBD	0.7718	0.7870	2.0188	6.0490	0.6215	0.6021	0.6217	2.7767	4.5632	0.3118
TV	0.7472	0.7893	3.1882	6.1393	0.6162	0.6480	0.6872	3.5109	3.9763	0.3265
ℓ_1	0.8617	0.8783	2.0688	4.1557	0.8196	0.6409	0.7017	3.7793	3.7117	0.3792
log	0.8636	0.8762	1.6053	3.3673	0.8923	0.7323	0.7395	2.4228	2.7072	0.6262

Table 10. QNR Quantitative results on the SPOT-5 Roma image.

p/P	4			16		
	D_λ	D_s	QNR	D_λ	D_s	QNR
EXP	0.0041	0.0150	0.9809	0.0001	0.0312	0.9687
PCA	0.2047	0.4094	0.4697	0.3094	0.5035	0.3429
IHS	0.2389	0.4158	0.4447	0.3574	0.5143	0.3121
Brovoy	0.1804	0.3799	0.5082	0.2890	0.4754	0.3730
BDSB	0.0108	0.0922	0.8980	0.0388	0.0344	0.9281
GS	0.1964	0.4100	0.4741	0.3045	0.5044	0.3447
GSA	0.2194	0.3421	0.5135	0.3267	0.4287	0.3846
PRACS	0.0325	0.1555	0.8171	0.0656	0.2162	0.7324
HPF	0.0851	0.1405	0.7864	0.2149	0.2556	0.5844
SFIM	0.0661	0.1256	0.8167	0.1949	0.2416	0.6107
Indusion	0.0580	0.0370	0.9072	0.2458	0.1587	0.6345
ATWT	0.1310	0.2119	0.6849	0.2398	0.3037	0.5293
AWLP	0.1030	0.1950	0.7221	0.2015	0.2814	0.5738
ATWT-M2	0.0728	0.2002	0.7416	0.0996	0.1691	0.7482
ATWT-M3	0.0162	0.0349	0.9494	0.0493	0.0328	0.9195
MTF-GLP	0.1040	0.1586	0.7539	0.2511	0.3000	0.5242
MTF-GLP-HPM	0.0858	0.1441	0.7825	0.2314	0.2868	0.5481
MTF-GLP-CBD	0.0657	0.1272	0.8154	0.1922	0.2681	0.5912
TV	0.3399	0.1830	0.5394	0.1866	0.1510	0.6906
ℓ_1	0.0277	0.0378	0.9356	0.1927	0.1500	0.6862
log	0.0056	0.0272	0.9674	0.0422	0.0380	0.9214

Table 11. Quantitative results using Wald's protocol on the FORMOSAT-2 Salon-de-Provence image.

p/P	4					16				
	Q4	Q	SAM	ERGAS	SCC	Q4	Q	SAM	ERGAS	SCC
EXP	0.8610	0.8617	1.8158	3.7257	0.8460	0.6918	0.6925	2.5057	2.6934	0.5906
PCA	0.7858	0.8089	2.2098	5.0711	0.6595	0.7415	0.7668	2.6373	2.7485	0.5870
IHS	0.7912	0.8159	1.9637	4.6293	0.6695	0.7445	0.7720	2.4527	2.5803	0.5894
Brovoy	0.7861	0.8146	1.9106	4.5332	0.6644	0.7385	0.7687	2.4345	2.5597	0.5793
BDS	0.8443	0.8545	2.0292	4.2608	0.7232	0.7701	0.7823	2.5460	2.7217	0.5821
GS	0.7979	0.8195	2.1375	4.9193	0.6619	0.7506	0.7753	2.5761	2.6860	0.5888
GSA	0.7859	0.8167	2.1855	5.2508	0.6508	0.7607	0.7772	2.5510	2.7698	0.5796
PRACS	0.8495	0.8592	1.9581	4.0842	0.7373	0.7947	0.8016	2.4240	2.3683	0.6230
HPF	0.8469	0.8523	1.9799	4.2210	0.7782	0.7813	0.7885	2.4609	2.5388	0.5900
SFIM	0.8470	0.8523	1.9821	4.2718	0.7755	0.7817	0.7888	2.4600	2.5658	0.5874
Indusion	0.8304	0.8371	2.0075	4.4361	0.7407	0.7519	0.7698	2.4986	2.6941	0.5758
ATWT	0.8391	0.8449	2.0374	4.5675	0.7663	0.7901	0.7965	2.4679	2.5399	0.6161
AWLP	0.8408	0.8473	1.8896	4.3549	0.7671	0.7909	0.7979	2.3689	2.4540	0.6150
ATWT-M2	0.8277	0.8326	2.2307	4.2164	0.6975	0.7672	0.7704	2.5499	2.3709	0.6325
ATWT-M3	0.8325	0.8356	2.2137	4.0586	0.7341	0.7639	0.7660	2.5659	2.3878	0.6384
MTF-GLP	0.8477	0.8533	1.9883	4.2705	0.7704	0.7901	0.7968	2.4714	2.5563	0.6176
MTF-GLP-HPM	0.8476	0.8532	1.9889	4.3313	0.7676	0.7904	0.7971	2.4666	2.5937	0.6158
MTF-GLP-CBD	0.8493	0.8548	2.0076	4.2575	0.7730	0.7846	0.7916	2.5270	2.6394	0.6123
TV	0.8696	0.8790	2.1437	3.7602	0.7876	0.7807	0.7878	2.7906	2.4374	0.5956
ℓ_1	0.8946	0.8974	1.9200	3.3526	0.8457	0.7691	0.7625	3.0726	2.6041	0.5503
log	0.8706	0.8683	1.7597	3.3686	0.8751	0.6889	0.6848	2.4975	2.6009	0.6050

Table 12. QNR Quantitative results on the FORMOSAT-2 Salon-de-Provence image.

p/P	4			16		
	D_λ	D_s	QNR	D_λ	D_s	QNR
EXP	0.0087	0.0837	0.9083	0.0086	0.0931	0.8990
PCA	0.1108	0.2190	0.6945	0.1503	0.3083	0.5877
IHS	0.1083	0.2127	0.7020	0.1508	0.3027	0.5921
Brovoy	0.0859	0.1964	0.7345	0.1215	0.2815	0.6312
BDS	0.0179	0.0150	0.9673	0.0264	0.1775	0.8008
GS	0.0970	0.2084	0.7149	0.1366	0.2971	0.6069
GSA	0.1254	0.2095	0.6914	0.1618	0.2963	0.5899
PRACS	0.0621	0.1656	0.7826	0.0878	0.2413	0.6921
HPF	0.0896	0.1130	0.8075	0.1147	0.1559	0.7473
SFIM	0.0877	0.1121	0.8101	0.1128	0.1556	0.7492
Indusion	0.0510	0.0174	0.9325	0.0938	0.0965	0.8187
ATWT	0.1227	0.1529	0.7432	0.1355	0.1951	0.6958
AWLP	0.1290	0.1500	0.7404	0.1393	0.1887	0.6983
ATWT-M2	0.1184	0.1550	0.7449	0.0967	0.1238	0.7915
ATWT-M3	0.0871	0.0951	0.8261	0.0586	0.0388	0.9049
MTF-GLP	0.1004	0.1277	0.7847	0.1437	0.1996	0.6854
MTF-GLP-HPM	0.0980	0.1269	0.7875	0.1410	0.1989	0.6881
MTF-GLP-CBD	0.0910	0.1222	0.7979	0.1317	0.1919	0.7017
TV	0.1114	0.0787	0.8186	0.0951	0.1811	0.7410
ℓ_1	0.0425	0.0514	0.9083	0.0594	0.0847	0.8609
log	0.0094	0.0917	0.8998	0.0174	0.1011	0.8832

Table 13. Elapsed CPU time in seconds for the different pansharpening methods on a 1024×1024 image and with different p/P ratios.

P	p	PCA	IHS	Brovey	BDSB	GS
512×512	1024×1024	0.9	0.04	0.04	0.9	0.4
256×256	1024×1024	0.3	0.03	0.03	0.8	0.3
P	p	GSA	PRACS	HPF	SFIM	Inclusion
512×512	1024×1024	1	2.2	0.2	0.5	0.5
256×256	1024×1024	1.6	1.2	0.2	0.18	0.3
P	p	ATWT	AWLP	ATWT-M2	ATWT-M3	MTF-GLP
512×512	1024×1024	11	14	10	10	0.9
256×256	1024×1024	3	3	7	7	0.6
P	p	MTF-HPM	MTF-CBD	TV	ℓ_1	log
512×512	1024×1024	0.8	2	$1.5 \cdot 10^3$	808	$1.1 \cdot 10^4$
256×256	1024×1024	0.6	0.6	$1.6 \cdot 10^3$	$3 \cdot 10^3$	$2.8 \cdot 10^4$

All color images in this figure are RGB images formed from the B4, B3 and B2 Landsat bands. Since we are using full resolution images, there is no ground truth to compare with, so a visual analysis of the resulting images is performed. The improved resolution of all the pansharpening results in Fig. 7(c)-(h) with respect to the observed MS image in Fig. 7(a) is evident. PRACS, MTF-GLP-HPM and MTF-GLP-CBD images in Fig. 7(c)-(e) have a lower detail level than TV and the proposed ℓ_1 method, see Fig. 7(f) and 7(g), respectively. See, for instance, the staircase effects in some diagonal edges not present in the TV and proposed ℓ_1 method results. The PRACS, MTF-GLP-HPM and MTF-GLP-CBD methods produce similar, but lower, spectral quality than the proposed method, which is consistent with the numerical results in Table 3 and discussion presented in section 6.1. The image obtained using the ℓ_1 method, Fig. 7(g), has colors closer to those of the observed MS image than the TV image, Fig. 7(f), which is also somewhat noisier. The log method is very good at removing noise in the image (see the sea area) but it tends to remove other fine details too.

Figure 8 shows a region of interest of the observed SPOT-5 images in Fig. 3(a) and 3(b) and the pansharpening results with $p/P = 16$ obtained using the competing methods with the best performance on this image, that is, Brovey, PRACS, ATWT-M3, and TV methods, and the proposed ℓ_1 and log methods. All color images in this figure are RGB images formed from the bands B3, B2 and B1 of the SPOT-5 image. The Brovey method (Fig. 8(c)) produces the highest spectral distortion, however, it also recovers more spatial details in the image (see the airport runway and plane). ATWT-M3 (Fig. 8(e)), on the other hand, produces a blurry image. PRACS produces a sharper image, see Fig. 8(d), but details in the PAN image do not seem to be well integrated. The TV method (Fig. 8(f)) and the proposed ℓ_1 and log methods (Fig. 8(g) and 8(h)) obtain the most consistent results, with high spatial details and low spectral distortion. However, TV introduces staircase artifacts on diagonal lines that are not noticeable in the ℓ_1 and log images. As with the LANDSAT-7 image, the log image in Fig. 8(h) lacks some small details, removed by the method along with noise.

7. Conclusions

A variational Bayesian methodology for the pansharpening problem has been proposed. In this methodology, we model the relation between the MS high resolution image and the PAN image as a linear combination of the MS bands whose weights are estimated from the available data. The observed MS image is modelled as a downsampled version of the original MS image. The expected characteristics of the pansharpened image are incorporated in the form of SG sparse image priors. Two penalty functions corresponding to SG distributions are used, ℓ_1 and log. All the unknowns and model parameters have been automatically estimated within the variational Bayesian modelling and inference, and an efficient algorithm has been obtained.

The proposed ℓ_1 and log methods have been compared to classic and state-of-the-art methods obtaining very good results both quantitative and qualitatively. In general, they have obtained the

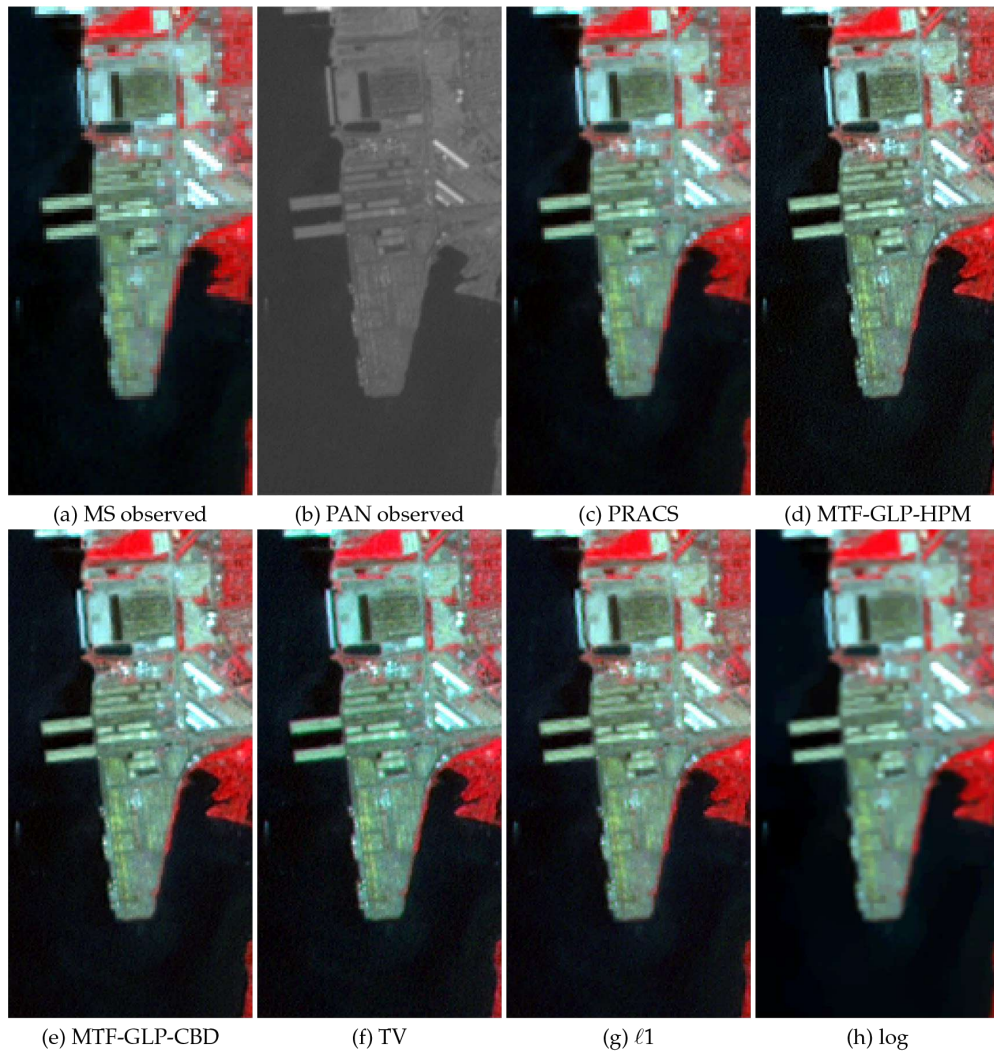


Figure 7. A region of interest of the LANDSAT 7-ETM+ Chesapeake Bay image in Fig. 1(a). Observed images: (a) 128×64 MS, (b) 256×128 PAN. 256×128 pansharpened images by: (c) PRACS, (d) MTF-GLP-HPM, (e) MTF-GLP-CBD, (f) TV, (g) ℓ_1 and (h) log methods.

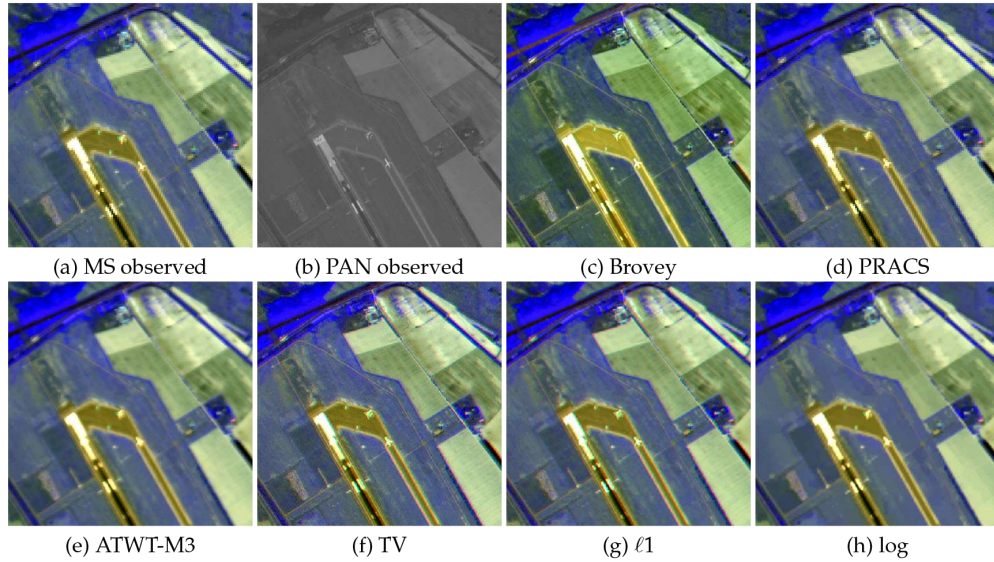


Figure 8. A region of interest of the SPOT-5 Roma image in Fig. 3(a). Observed images: (a) 128×128 MS, (b) 512×512 PAN. 512×512 pansharpened images by: (c) Brovey, (d) PRACS, (e) ATWT-M3, (f) TV, (g) ℓ_1 and (h) log methods.

best quantitative results for LANDSAT-7 ETM+, SPOT-5 and FORMOSAT-2 images with a resolution ratio of 4 and SPOT-5 with a resolution ratio of 16. Competitive results were also obtained for the FORMOSAT-2 image with a resolution ratio of 16. They stand out in terms of spectral consistency while improving the spatial resolution of pansharpened images. We argue that the superior spectral consistency of SG methods arises from the modelling of the PAN image which selectively incorporates PAN detailed information into the different MS high resolution bands without changing their spectral properties. Qualitatively, SG methods produce results consistent with the observed PAN and MS images and with the numerical results previously described. The log method is better at removing noise in the images, at the cost of removing some fine details.

Author Contributions: Conceptualization, M.V., R.M.; methodology, M.V.; software, F.P., M.V.; validation, F.P.; formal analysis, J.M., M.V.; investigation, F.P., M.V.; resources, J.M., M.V.; data curation, F.P.; writing—original draft preparation, F.P., M.V.; writing—review and editing, A.K.K., J.M. and R.M.; visualization, F.P., J.M. and M.V.; supervision, A.K.K., R.M.; project administration, J.M., R.M.; funding acquisition, J.M., R.M.. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Spanish Ministerio de Economía y Competitividad under contract DPI2016-77869-C2-2-R, by the Ministerio de Ciencia e Innovación under contract PID2019-105142RB-C22, and the Visiting Scholar Program at the University of Granada.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

MS	multispectral
PAN	panchromatic
PCA	principal components analysis
IHS	intensity-hue-saturation
CS	component substitution
MRA	multi-resolution analysis
VO	variational optimization
DL	deep learning
CS	component substitution
GS	Gram-Schmidt
HPF	high-pass filtering
WT	wavelet transform
GLP	generalized Laplacian pyramid
NSCT	non-subsampled contourlet transform
AWT	“a trous” wavelet transforms
SFIM	smoothing filter based intensity modulation
DNN	deep neural networks
MSDA	modified sparse denoising autoencoder
CSDA	coupled sparse denoising autoencoder
PCDRN	progressive cascade deep residual network
GAN	generative adversarial network
SG	super-Gaussian
Brovey	Brovey transform
BDSB	band-dependent spatial-detail
GSA	Gram-Schmidt adaptive
PRACS	partial replacement adaptive component substitution
ATWT	additive “a trous” wavelet transform
ATLP	additive wavelet luminance proportional
MTF	modulation transfer functions
HPM	high pass modulation
CBD	context based decision
TV	total variation
UQI	universal quality index
SCC	spatial correlation coefficient
SAM	spectral angle mapper
ERGAS	erreur relative globale adimensionnelle de synthese
QNR	quality with no reference

1. Babacan, S.D.; Molina, R.; Do, M.N.; Katsaggelos, A.K. Blind deconvolution with general sparse image priors. *ECCV*, 2012.
2. Rubio, J.; Vega, M.; Molina, R.; Katsaggelos, A.K. A general sparse image prior combination in Compressed Sensing. 21st European Signal Processing Conference (EUSIPCO 2013), 2013, pp. 1–5.
3. Zhou, X.; Vega, M.; Zhou, F.; Molina, R.; Katsaggelos, A.K. Fast bayesian blind deconvolution with Huber super Gaussian priors. *Digital Signal Processing* **2017**, *60*, 122–133.
4. Pérez-Bueno, F.; Vega, M.; Naranjo, V.; Molina, R.; Katsaggelos, A. Fully automatic blind color deconvolution of histological images using super Gaussians. 28th European Signal Processing Conference, EUSIPCO 2020. Amsterdam (Netherlands), 2021.
5. Ehlers, M. Multisensor image fusion techniques in remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing* **1991**, *46*, 19–30.
6. Chavez, P.S.; Kwarteng, A.Y. Extracting spectral contrast in Landsat thematic mapper image data using selective principal component analysis. *Photogrammetric Engineering and Remote Sensing* **1989**, *55*, 339–348.
7. Gillespie, A.R.; Kahle, A.B.; Walker, R.E. Color enhancement of highly correlated images. II. Channel ratio and “chromaticity” transformation techniques. *Remote Sensing of Environment* **1987**, *22*, 343–365.

8. Carper, W.J.; Lillesand, T.M.; Kiefer, R.W. The use of intensity-hue-saturation transformations for merging SPOT panchromatic and multispectral image data. *Phot. Eng. & Rem. Sens.* **1990**, *56*, 459–467.
9. Pohl, C.; Genderen, J.L.V. Multi-sensor image fusion in remote sensing: Concepts, methods, and applications. *International Journal of Remote Sensing* **1998**, *19*, 823–854.
10. Ranchln, T.; Wald, L. Fusion of high spatial and spectral resolution images: The ARSIS concept and its implementation. *Photogrammetric Engineering and Remote Sensing* **2000**, *66*, 49–61.
11. Schowengerdt, R.A. *Remote sensing: models and methods for image processing*; Elsevier, 2006.
12. Amro, I.; Mateos, J.; Vega, M.; Molina, R.; Katsaggelos, A. A survey of classical methods and new trends in pansharpening of multispectral images. *EURASIP Journal on Advances in Signal Processing* **2011**, *2011*, 79.
13. Vivone, G.; Alparone, L.; Chanussot, J.; Dalla Mura, M.; Garzelli, A.; Licciardi, G.A.; Restaino, R.; Wald, L. A critical comparison among pansharpening algorithms. *IEEE Transactions on Geoscience and Remote Sensing* **2015**, *53*, 2565–2586.
14. Alparone, L.; Baronti, S.; Aiazzi, B.; Garzelli, A. Spatial Methods for Multispectral Pansharpening: Multiresolution Analysis Demystified. *IEEE Transactions on Geoscience and Remote Sensing* **2016**, *54*, 2563–2576.
15. Duran, J.; Buades, A.; Coll, B.; Sbert, C.; Blanchet, G. A survey of pansharpening methods with a new band-decoupled variational model. *ISPRS Journal of Photogrammetry and Remote Sensing* **2017**, *125*, 78–105.
16. Kahraman, S.; Erturk, A. Review and performance comparison of pansharpening algorithms for RASAT images. *Istanbul University - Journal of Electrical & Electronics Engineering* **2018**, *18*, 109–120. doi:10.5152/iujeee.2018.1817.
17. Meng, X.; Shen, H.; Li, H.; Zhang, L.; Fu, R. Review of the pansharpening methods for remote sensing images based on the idea of meta-analysis: Practical discussion and challenges. *Information Fusion* **2019**, *46*, 102–113.
18. Chavez, P.S.; Sides, S.; Anderson, J. Comparison of three different methods to merge multiresolution and multispectral data: Landsat TM and SPOT panchromatic. *Phot. Eng. & Rem. Sens.* **1991**, *57*, 295–303.
19. Tu, T.M.; Su, S.C.; Shyu, H.C.; Huang, P.S. A new look at IHS-like image fusion methods. *Information fusion* **2001**, *2*, 177–186.
20. Laben, C.A.; Brower, B.V. Process for enhancing the spatial resolution of multispectral imagery using pan-sharpening, 2000. US Patent 6,011,875.
21. Aiazzi, B.; Baronti, S.; Selva, M. Improving component substitution pansharpening through multivariate regression of MS + Pan data. *IEEE Transactions on Geoscience and Remote Sensing* **2007**, *45*, 3230–3239.
22. Garzelli, A.; Nencini, F.; Capobianco, L. Optimal MMSE pan sharpening of very high resolution multispectral images. *IEEE Transactions on Geoscience and Remote Sensing* **2007**, *46*, 228–236.
23. Choi, J.; Yu, K.; Kim, Y. A new adaptive component-substitution-based satellite image fusion by using partial replacement. *IEEE Transactions on Geoscience and Remote Sensing* **2010**, *49*, 295–309.
24. Shahdoosti, H.R.; Javaheri, N. Pansharpening of clustered MS and Pan images considering mixed pixels. *IEEE Geoscience and Remote Sensing Letters* **2017**, *14*, 826–830.
25. Palsson, F.; Sveinsson, J.R.; Ulfarsson, M.O.; Benediktsson, J.A. MTF-based deblurring using a Wiener filter for CS and MRA pansharpening methods. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2016**, *9*, 2255–2269.
26. Nuñez, J.; Otazu, X.; Fors, O.; Prades, A.; Pala, V.; Arbiol, R. Multiresolution-Based Image fusion with additive wavelet decomposition. *IEEE Trans on Geosc. & Rem. Sens.* **1999**, *37*, 1204–1211.
27. King, R.; Wang, J. A wavelet based algorithm for pansharpening Landsat 7 imagery. Proc. of the Int. Geosc. and Rem. Sens. Symp., 2001, Vol. 2, pp. 849–851.
28. Otazu, X.; Gonzalez-Audicana, M.; Fors, O.; Nuñez, J. Introduction of sensor spectral response into image fusion methods. Application to wavelet-based methods. *IEEE Trans on Geosc. & Rem. Sens.* **2005**, *43*, 2376–2385.
29. Aiazzi, B.; Alparone, L.; Baronti, S.; Garzelli, A.; Selva, M. MTF-tailored multiscale fusion of high-resolution MS and Pan Imagery. *Phot. Eng. & Rem. Sens.* **2006**, *72*, 591–596.
30. Aiazzi, B.; Alparone, L.; Baronti, S.; Pippi, I.; Selva, M. Generalised Laplacian pyramid-based fusion of MS + P image data with spectral distortion minimisation. *ISPRS Internat. Archives Photogramm. Remote Sensing* **2002**, *34*, 3–6.

31. Aiazzi, B.; Alparone, L.; Baronti, S.; Garzelli, A.; Selva, M. Advantages of Laplacian pyramids over “à trous” wavelet transforms for pansharpening of multispectral images. *Image and Signal Processing for Remote Sensing XVIII. International Society for Optics and Photonics*, 2012, Vol. 8537, p. 853704.
32. Shah, V.P.; Younan, N.H.; King, R.L. An efficient pan-sharpening method via a combined adaptive PCA approach and contourlets. *IEEE Transactions on Geoscience and Remote Sensing* **2008**, *46*, 1323–1335.
33. Amro, I.; Mateos, J.; Vega, M. Parameter estimation in Bayesian super-resolution pansharpening using contourlets. *IEEE International Conference on Image Processing ICIP 2011*, 2011, pp. 1345–1348.
34. Upla, K.P.; Gajjar, P.P.; Joshi, M.V. Pan-sharpening based on non-subsampled contourlet transform detail extraction. *2013 Fourth National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG)*, 2013, pp. 1–4.
35. Liu, J. Smoothing filter-based intensity modulation: A spectral preserve image fusion technique for improving spatial details. *International Journal of Remote Sensing* **2000**, *21*, 3461–3472.
36. Wald, L.; Ranchin, T. Liu ‘Smoothing filter-based intensity modulation: A spectral preserve image fusion technique for improving spatial details’. *International Journal of Remote Sensing* **2002**, *23*, 593–597.
37. Khan, M.M.; Chanussot, J.; Condat, L.; Montanvert, A. Indusion: Fusion of multispectral and panchromatic images using the induction scaling technique. *IEEE Geoscience and Remote Sensing Letters* **2008**, *5*, 98–102.
38. Huang, W.; Xiao, L.; Wei, Z.; Liu, H.; Tang, S. A new pan-sharpening method with deep neural networks. *IEEE Geoscience and Remote Sensing Letters* **2015**, *12*, 1037–1041.
39. Cai, W.; Xu, Y.; Wu, Z.; Liu, H.; Qian, L.; Wei, Z. Pan-sharpening based on multilevel coupled deep network. *2018 IEEE International Geoscience and Remote Sensing Symposium*, 2018, pp. 7046–7049.
40. Masi, G.; Cozzolino, D.; Verdoliva, L.; Scarpa, G. Pansharpening by convolutional neural networks. *Remote Sensing* **2016**, *8*, 594.
41. Eghbalian, S.; Ghassemian, H. Multi spectral image fusion with deep convolutional network. *2018 9th International Symposium on Telecommunications*, 2018, pp. 173–177.
42. Wei, Y.; Yuan, Q.; Shen, H.; Zhang, L. Boosting the Accuracy of Multispectral Image Pansharpening by Learning a Deep Residual Network. *IEEE Geoscience and Remote Sensing Letters* **2017**, *14*, 1795–1799.
43. Li, N.; Huang, N.; Xiao, L. PAN-Sharpener via residual deep learning. *2017 IEEE International Geoscience and Remote Sensing Symposium*, 2017, pp. 5133–5136.
44. Yang, Y.; Tu, W.; Huang, S.; Lu, H. PCDRN: Progressive Cascade Deep Residual Network for Pansharpening. *Remote Sensing* **2020**, *12*, 676.
45. Huang, W.; Fei, X.; Feng, J.; Wang, H.; Liu, Y.; Huang, Y. Pan-sharpening via multi-scale and multiple deep neural networks. *Signal Processing: Image Communication* **2020**, *85*, 115850. doi:10.1016/j.image.2020.115850.
46. Ma, J.; Yu, W.; Chen, C.; Liang, P.; Guo, X.; Jiang, J. Pan-GAN: An unsupervised pan-sharpening method for remote sensing image fusion. *Information Fusion* **2020**, *62*, 110–120. doi:10.1016/j.inffus.2020.04.006.
47. Xu, H.; Ma, J.; Jiang, J.; Guo, X.; Ling, H. U2Fusion: A Unified Unsupervised Image Fusion Network. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2020**.
48. Chan, T.; Shen, J.; Vese, L. Variational PDE models in image processing. *Notices Amer. Math. Soc.* **2003**, *50*, 14–26.
49. Fang, F.; Li, F.; Shen, C.; Zhang, G. A variational approach for pan-sharpening. *IEEE Transactions on Image Processing* **2013**, *22*, 2822–2834.
50. Loncan, L.; de Almeida, L.B.; Bioucas-Dias, J.M.; Briottet, X.; Chanussot, J.; Dobigeon, N.; Fabre, S.; Liao, W.; Licciardi, G.A.; Simões, M.; Tourneret, J.; Veganzones, M.A.; Vivone, G.; Wei, Q.; Yokoya, N. Hyperspectral pansharpening: A review. *IEEE Geoscience and Remote Sensing Magazine* **2015**, *3*, 27–46.
51. Ballester, C.; Caselles, V.; Igual, L.; Verdera, J.; Rougé, B. A variational model for P+XS image fusion. *International Journal of Computer Vision* **2006**, *69*, 43–58. doi:10.1007/s11263-006-6852-x.
52. Molina, R.; Vega, M.; Mateos, J.; Katsaggelos, A. Variational posterior distribution approximation in Bayesian super resolution reconstruction of multispectral images. *Applied and Computat. Harmonic Analysis* **2008**, *24*, 251–267.
53. Vega, M.; Mateos, J.; Molina, R.; Katsaggelos, A. Super resolution of multispectral images using TV image models. *2th Int. Conf. on Knowledge-Based and Intelligent Information & Engineering Systems*, 2008, pp. 408–415.
54. Palsson, F.; Sveinsson, J.R.; Ulfarsson, M.O. A new pansharpening algorithm based on total variation. *IEEE Geoscience and Remote Sensing Letters* **2014**, *11*, 318–322.

55. Yang, X.; Jian, L.; Yan, B.; Liu, K.; Zhang, L.; Liu, Y. A sparse representation based pansharpening method. *Future Generation Computer Systems* **2018**, *88*, 385–399.
56. Vega, M.; Mateos, J.; Molina, R.; Katsaggelos, A. Super resolution of multispectral images using l1 image models and interband correlations. *Journal of Signal Processing Systems* **2011**, *65*, 509–523.
57. Zhang, M.; Li, S.; Yu, F.; Tian, X. Image fusion employing adaptive spectral-spatial gradient sparse regularization in UAV remote sensing. *Signal Processing* **2020**, *170*, 107434.
58. Duran, J.; Buades, A.; Coll, B.; Sbert, C. A nonlocal variational model for pansharpening image fusion. *SIAM Journal on Imaging Sciences* **2014**, *7*, 761–796.
59. Fu, X.; Lin, Z.; Huang, Y.; Ding, X. A variational pan-sharpening with local gradient constraints. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 10265–10274.
60. Li, W.; Hu, Q.; Zhang, L.; Du, J. Pan-sharpening with a spatial-enhanced variational model. *Journal of Applied Remote Sensing* **2018**, *12*, 035018.
61. Chen, Y.; Wang, T.; Fang, F.; Zhang, G. A pan-sharpening method based on the ADMM algorithm. *Frontiers of Earth Science* **2019**, *13*, 656–667.
62. Garzelli, A. A review of image fusion algorithms based on the super-resolution paradigm. *Remote Sensing* **2016**, *8*. doi:10.3390/rs8100797.
63. Tian, X.; Chen, Y.; Yang, C.; Gao, X.; Ma, J. A Variational Pansharpening Method Based on Gradient Sparse Representation. *IEEE Signal Processing Letters* **2020**, *27*, 1180–1184.
64. Vivone, G.; Simões, M.; Dalla Mura, M.; Restaino, R.; Bioucas-Dias, J.M.; Licciardi, G.A.; Chanussot, J. Pansharpening based on semiblind deconvolution. *IEEE Transactions on Geoscience and Remote Sensing* **2015**, *53*, 1997–2010.
65. Chen, C.; Li, Y.; Liu, W.; Huang, J. SIRF: Simultaneous Satellite Image Registration and Fusion in a Unified Framework. *IEEE Transactions on Image Processing* **2015**, *24*, 4213–4224.
66. Rockafellar, R. *Convex analysis*; Princeton University Press, 1996.
67. Bishop, C., Mixture models and EM. In *Pattern Recognition and Machine Learning*; Springer, 2006; pp. 454–455.
68. Kullback, S. *Information Theory and Statistics*; Dover Publications: Mineola, N.Y., 1959.
69. Vega, M.; Molina, R.; Katsaggelos, A. Parameter Estimation in Bayesian Blind Deconvolution with Super Gaussian Image Priors. *Signal Processing Conference (EUSIPCO)*, 2014 Proceedings of the 22nd European; IEEE., Ed. Lisbon (Portugal), 2014, pp. 1632–1636.
70. U.S. Geological Survey. Landsat Missions.
71. Satellite Imaging Corporation. SPOT-5 Satellite Sensor.
72. Satellite Imaging Corporation. FORMOSAT-2 Satellite Sensor.
73. Wald, L.; Ranchin, T.; Mangolini, M. Fusion of satellite images of different spatial resolutions: assessing the quality of resulting images. *Phot. Eng. Rem. Sens.* **1997**, *63*, 691–699.
74. Alparone, L.; Wald, L.; Chanussot, J.; Thomas, C.; Gamba, P.; Bruce, L. Comparison of pansharpening algorithms: Outcome of the 2006 GRS-S data-fusion contest. *IEEE Trans. on Geosc. & Rem. Sens.* **2007**, *45*, 3012–3020.
75. Wang, Z.; Bovik, A.C. A universal image quality index. *IEEE Sign. Proc. Lett.* **2002**, *9*, 81–84.
76. Garzelli, A.; Nencini, F. Hypercomplex quality assessment of multi/hyperspectral images. *IEEE Geoscience and Remote Sensing Letters* **2009**, *6*, 662–665. Publisher: IEEE.
77. Pratt, W.K. Correlation Techniques of Image Registration. *IEEE Transactions on Aerospace and Electronic Systems* **1974**, *AES-10*, 353–358.
78. Alparone, L.; Aiazzi, B.; Baronti, S.; Garzelli, A.; Nencini, F.; Selva, M. Multispectral and panchromatic data fusion assessment without reference. *Photogrammetric Engineering & Remote Sensing* **2008**, *74*, 193–200.
79. Yuhas, R.H.; Goetz, A.F.H.; Boardman, J.W. Discrimination among semi-arid landscape endmembers using the Spectral Angle Mapper (SAM) algorithm. *Proceeding Summaries 3rd Annual JPL Airborne Geoscience Workshop*, 1992, pp. 147–149.
80. Wald, L. Quality of high resolution synthesized images: Is there a simple criterion? *Proc. Int. Conf. Fusion of Earth Data* **2000**, *1*, 99–105.

CHAPTER 6

Network Security Anomaly Detection Using Probabilistic Models

6.1 JCR Publication Details

Authors: Fernando Pérez-Bueno, Luz García, Gabriel Maciá-Fernández, Rafael Molina

Title: Leveraging a Probabilistic PCA model to Understand the Multivariate Statistical Network Monitoring Framework for Network Security Anomaly Detection

Reference: IEEE/ACM Transactions on Networking (Early Access), 2022, 1-13

Status: Published

DOI: <https://doi.org/10.1109/TNET.2021.3138536>

Quality indices:

- Impact Factor (JCR 2021): 3.796
 - Rank: 26/109 (Q1) in Computer Science, Theory and Methods
- Journal Citation Indicator (JCR 2021): 1,03
 - Rank: 24/142 (Q1) in Computer Science, Theory and Methods
 - Rank: 83/344 (Q1) in Engineering, Electrical and Electronic

6.2 Main Contributions

- We connect the generative Probabilistic PCA model with the previously presented Multivariate Statistical Network Monitoring (MSNM) for anomaly detection, showing that the MSNM is a particular case of PPCA. We then develop a mathematical framework to explain from a probabilistic point

of view, the meaning of the anomaly detection statistics proposed in the MSNM approach.

- We use the generative model to circumvent limitations of the MSNM model, such as the weighting of the regularization and reconstruction terms and the automatic estimation of the model parameters.
- The generative PPCA model is used to better understand the relationship with more complex generative models such as Variational Autoencoders.
- The mathematical model was validated on synthetic and real-traffic datasets for network anomaly detection.

Leveraging a Probabilistic PCA model to Understand the Multivariate Statistical Network Monitoring Framework for Network Security Anomaly Detection

Fernando Pérez-Bueno, Luz García, Gabriel Maciá-Fernández, Rafael Molina

Abstract

Network anomaly detection is a very relevant research area nowadays, especially due to its multiple applications in the field of network security. The boost of new models based on variational autoencoders and generative adversarial networks has motivated a reevaluation of traditional techniques for anomaly detection. It is, however, essential to be able to understand these new models from the perspective of the experience attained from years of evaluating network security data for anomaly detection. In this paper, we revisit anomaly detection techniques based on PCA from a probabilistic generative model point of view, and contribute a mathematical model that relates them. Specifically, we start with the probabilistic PCA model and explain its connection to the Multivariate Statistical Network Monitoring (MSNM) framework. MSNM was recently successfully proposed as a means of incorporating industrial process anomaly detection experience into the field of networking. We have evaluated the mathematical model using two different datasets. The first, a synthetic dataset created to better understand the analysis proposed, and the second, UGR'16, is a specifically designed real-traffic dataset for network security anomaly detection. We have drawn conclusions that we consider to be useful when applying generative models to network security detection.

Index Terms

Anomaly Detection; PPCA; Generative Models; Network Security.

I. INTRODUCTION

Network security constitutes mainly a research and development focus nowadays, with a forecasted market of \$170.4 billion in 2022 according to Gartner [1], and a constant flow of worrying news concerning security incidents, like data breaches, denial of service, data exfiltration, privacy issues, advanced persistent threats, or even government cyberwar issues [2].

Fernando Pérez-Bueno (corresponding author) and Rafael Molina are with Dpto. de Ciencias de la Computación e I. A., Universidad de Granada, Spain. E-mail: fpb@ugr.es; rms@decsai.ugr.es

Luz García and Gabriel Maciá-Fernández are with Dpto. de Teoría de la Señal, Telemática y Comunicaciones, Universidad de Granada, Spain. E-mail: luzgm@ugr.es; gmacia@ugr.es

Preprint submitted to IEEE Transactions on Networking

Many security technologies have been developed in recent years to deal with these relevant problems. The in-depth security design paradigm advocates the use of different layers of security protection to deal with such problems. Among these layers, the detection of network security incidents is a crucial part of developing effective response measures when attacks occur.

There are two main approaches to the detection of network security incidents. One, *signature detection* works with rules defined by experts that identify known attacks. This approach works efficiently on known attacks but lacks the flexibility to detect unseen attack patterns or multi-stage attacks. An alternative approach, known as *anomaly detection*, based on building a normality model and detecting deviations from it, has gained popularity. However, the main limitation of anomaly detection technologies is related to the appearance of false positives or negatives, due to the lack of accuracy in the normality models obtained.

In this context, many different approaches to anomaly detection have been proposed in the literature [3]. Recent research in the deep learning area is generating high expectations with regard to the promising results that could be obtained in the field of network anomaly detection [4] [5] [6]. Preliminary results on the use of generative models, like variational autoencoders or generative adversarial networks, show high performance. There is, however, a concern that the internals of these models are not completely understood and that they might only show good results for specific datasets. In summary, there is a need to comprehensively understand the evolution from widely used models to these new deep generative models.

A family of models formed by those based on PCA are currently being successfully applied in the network anomaly detection problem. The first proposal to use PCA was that of Lakhina et al. [7] [8] in 2004. A training data matrix was built with different features extracted from network flows. The latent PCA model was then extracted from it and subsequently used to evaluate new network samples to decide on their abnormality.

Later in 2016, Camacho *et al.* [9] proposed the use of a framework based on PCA called MSNM (*Multivariate Statistical Network Monitoring*). MSNM essentially adapts a methodology known as MSPC (Multivariate Statistical Process Control), which has been extensively and successfully used in the field of anomaly detection in industrial process control to deal with the specificities of the network anomaly detection problem. In brief, the framework suggests the construction of a PCA model¹ for normal traffic patterns, and the use of two statistics, termed D and Q , which are thresholded in order to determine if an anomaly occurs. While the statistic Q is similar to the one used by Lakhina, the use of D , a Hotelling's T² statistic [10], is a novelty that also captures possible deviations from the latent model, rather than only in the observation model (Q statistic).

Despite the effectiveness of MSNM in anomaly detection (see the comparative analysis in [11]), two problems appear in the application of this model. First, there is a lack of understanding on how D and Q statistics behave in different scenarios and why. This implies that obtaining good results in certain scenarios could lead to false conclusions being drawn. Second, there is a need to relate these PCA based anomaly detection models to more novel approaches like generative models, e.g., VAEs, which are currently being used profusely in the anomaly detection area. Understanding this relationship allows well-known lessons from PCAs to be applied to generative models.

A better way to understand these models from a generative perspective was provided by Tipping and Bishop [12], in which the PCA model is derived from a probabilistic PCA (PPCA) model

¹The use of linear PCA techniques is justified by the advantages in the diagnosing of anomalies in posterior phases of an incident lifecycle.

when the variance of the latent distribution becomes zero. In addition, PPCA is at the basis of the definition of VAEs [13].

Starting from the PPCA model (a generative model that explains PCA), the focus of this paper is to derive an analytical model that elucidates why the use of the two MSNM statistics, D and Q , leads to effective anomaly detection and how this is understood in the framework of generative models. Note that the aim of this paper is not to contribute with a novel approach derived from MSNM, but rather to relate this model with generative models (analytically and empirically).

As it will be shown, while the Q statistic measures the quality of the reconstruction model, the D statistic will represent a regularization term in the generative model. As previously discussed, these conclusions should help to understand how to soundly employ generative models, like VAEs and GANs, for network anomaly detection. As an example, some of the current proposals for anomaly detection using VAEs, *e.g.*, [14], use only the reconstruction model, thus discarding the contributions of the regularization term that could be relevant, as we will show in what follows.

In summary, the contribution of this paper is threefold: *i)* A PPCA model is leveraged to understand the MSNM framework from a generative point of view. We then develop a mathematical framework to explain, from a probabilistic point of view, the meaning of the Q and D statistics. *ii)* Using the generative model, we show how some limitations of the MSNM model are circumvented. Specifically, the authors of [15] propose the use of a combined weighted statistic for both Q and D (called the t-Score). Then, we obtain a probabilistic interpretation for this weighted combination with PPCA-MSNM. *iii)* We test the generative model on both a synthetic dataset and a real network traffic dataset to show how detection results stay coherent with both models.

The rest of this paper is structured as follows. In Section II, we present related work and explain the contributions of this paper in more detail. Then, we provide the basics for MSNM and PPCA in Sections III and IV, respectively. The mathematical model that connects MSNM and PPCA is proposed in Section V. This proposal is validated with the experiments shown in Section VI. Finally, conclusions are drawn in Section VII.

II. RELATED WORK

Many statistical strategies for anomaly detection on the basis of Lakhina's approximation [7] [8] have been profusely proposed. The benefits of PCA's unsupervised nature have motivated the appearance of a wide range of work, like the PCA-based traffic matrix estimation of [16], the network anomography proposed by [17], or the combination of distributed tracking and in-network PCA-based anomaly detection of [18] among many others. Limitations of these models, like the high sensitivity to calibration settings, ineffective detection of large anomalies or difficulties to capture temporal correlations have been reported [19]. In addition, the proposals to solve these limitations also use different frameworks. Robust PCA [20], and its variation [21] [22], or the Karhunen-Loève expansion used by [23], are examples of the achieved progress.

In 2016, Camacho *et al.* proposed the use of the Multivariate Statistical Network Monitoring (MSNM) framework [9] [15] as an improvement to previous PCA proposals. In essence, MSNM is an adaptation from a sibling framework traditionally used in the field of industrial process control, known as MSPC (Multivariate Statistical Process Control) [24] [25] [26]. In order to face the particularities of the networking field, MSNM adapted the MSPC methodology to introduce new data pre-processing strategies and processing steps, like the deparsing of network traces [11]. In addition, MSNM research has focused on the evaluation of its implementation in real networks, the optimization of its parameters with semi-supervised models, enabling big-data processing, its

application to hierarchical architectures for issuing privacy and traffic reduction [9], enhancing visualization of network anomalies or supporting authentication systems [27].

The use of deep generative models in the field of anomaly detection, is currently a hot research topic due to good performance achieved by the use of deep learning techniques. Many authors have followed this approach using variational autoencoders (VAEs) as a natural evolution of PCA in the frame of reconstruction approaches to detect anomalies. Despite the fact that image and video processing research was the initial promoter of these models [6], they are also being extensively evaluated in the field of network anomaly detection. A recent survey of applications and techniques can be found in [4], yet, it is surprising to see that in many of these studies, like [14], where they consider the time gradient effect, or in the conditional VAE implemented in [28], or in others like [29] or [30], anomaly scores are evaluated by only considering the reconstruction error provided by the VAE. That is, the regularization term present in the VAE marginal likelihood is not taken into account. Very few proposals, like [31], use both the regularization and the error reconstruction terms of the marginal likelihood to evaluate anomaly scores.

As will be shown in the rest of this paper, not using the regularization term in VAEs is similar to the well-known limitation in Lakhina's PCA approaches that do not consider the latent variable space deviations for anomaly detection. Thus, although evaluating the impact of the use of both terms (regularization and reconstruction) in VAEs is out of the scope of this paper, the goal here is to show the connection between these terms in a PPCA generative model (precursor of VAE) and MSNM detection statistics.

III. MSNM FOR ANOMALY DETECTION

Multivariate Statistical Network Monitoring (MSNM) [9] transfers the theory of Statistical Processes Control, which has been used for a long time in industrial applications, to network traffic analysis. Its goal is to jointly analyze several interrelated variables to differentiate common from special causes of variation called anomalies. The approach consists of five steps [11], which are explained in what follows.

First, raw network traffic data from different sources are parsed and transformed into a set of quantitative features, often using the feature-as-a-counter approach. Examples of such features are the number of times an event takes place, the count of a given word in a log, the number of times a given event takes place in a given time window, or the number of traffic flows with a given destination port in a NetFlow. The selection of the specific features and their parsing step require an effective comprehension of the data.

Second, all features are fused for the multivariate analysis, maintaining a common sampling rate. As a result, traffic flow matrices of N observations featured in M -dimensional vectors $N \times M$ are ready to be analyzed.

The *third* and main step of MSNM is the anomaly detection, which is based on PCA. The well-known technique is applied to the mean-centered and auto-scaled M -dimensional dataset of N observations ($N \times M$), and projected into a subspace of range $P < M$ that maximizes the variance. To do so, original features are transformed into Principal Components (PC) using the eigenvectors of the covariance matrix $\mathbf{X}^T \mathbf{X} / N$. As a result, a residual matrix \mathbf{E} is generated as the differential error between projections and real samples. The transformation follows Eq. (1), with \mathbf{T} ($N \times P$) and \mathbf{V} ($M \times P$) being the score and loading matrices, respectively :

$$X = T \cdot V^T + E \quad (1)$$

Based on such transformation, MSNM proposes the usage of two complementary statistics extracted from the PCA analysis:

- The Q -statistic, also called *Squared Prediction Error*, comprises the residuals in the n -th row of \mathbf{E} for a given observation x_n , following expression (2). As mentioned in Section I, Q evaluates the reconstruction error of the projection used:

$$Q_n = e_n e_n^t \quad (2)$$

- The D -statistic, or *Hotelling's T2 statistic*, is computed by applying Eq. (3) to PCA scores. As mentioned in Section I, D represents a regularization term that rates how close the observation is to the data prior distribution. For a given observation x_n , and as t_n is the score vector in the n -th row of \mathbf{T} :

$$D_n = t_n \Lambda^{-1} t_n^t \quad (3)$$

Intuitively, Q measures the capability of the model to recover a certain point in the data, while the regularization term D measures the similarity of the latent representation with respect to those in the calibration data. As each term focuses on different domains, they are able to capture different types of anomalies.

It is well known that the Q -statistic has a high anomaly detection capability, and its usage together with the D -statistic is a key feature of the MSNM approach that offers attractive improvements [9]. Once D and Q are calculated, they are used to model the normal operating conditions for the calibration of the MSNM system. In order to do so, *upper control limits* (UCL) for a given significance level are defined for both Q , termed UCL_Q , and D , termed UCL_D . Diverse combinations of the two statistics can be used to provide a final expression for the anomaly evaluation. The authors of [15] propose the following weighting of Q and D to generate the anomaly score for a given observation x_n :

$$t - Score_n = \frac{P \cdot D_n}{M \cdot UCL_D} + \frac{(M - P) \cdot Q_n}{M \cdot UCL_Q} \quad (4)$$

Once the anomaly is detected, the *fourth* step of MSNM approximation is the pre-diagnosis. Features associated with the anomalies are identified in order to make an initial guess on their root causes. Contribution plots or others tools like oMEDA [9] are commonly used to identify such features.

Finally, the *fifth* step consists of de-parsing the information pointed out during the detection (anomaly time-stamps) and pre-diagnosis (anomaly related features) phases. As a final result, raw information about the anomaly is extracted from specific logs or network traces.

In what follows, we will introduce PPCA as a generative model framework (described in Section IV) in order to derive the expressions for Q and D statistics, as well as the interpretation of Eq. (4) in Section V.

IV. PPCA FOR ANOMALY DETECTION

Probabilistic PCA (PPCA) provides a method to calculate the principal subspace of a set of data vectors using a generative point of view and a maximum-likelihood framework. In order to be able to understand MSNM from a generative model perspective, it is important to first understand how PPCA is related to PCA.

This section provides a description of this connection, following the presentation of probabilistic PCA provided in [32]. While PCA is based on a deterministic linear projection of the data on a lower dimensional subspace, PPCA is a linear-Gaussian framework that considers a latent distribution for the data. Therefore a whole distribution of possible latent candidates is available

for each observed data point. PPCA includes the measurement of deviations in the latent space, achieved in MSNM with the addition of the D statistic, among other advantages (see [12] for a complete list).

Let \mathbf{X} denote the $N \times M$ network traffic matrix whose i -th row, \mathbf{x}_i^T , corresponds to the i -th observed instance, $i = 1, \dots, N$. That is, $\mathbf{X}^T = [\mathbf{x}_1, \dots, \mathbf{x}_N]$. We assume that each column of \mathbf{X}^T has been centered and normalized by its standard deviation. We also assume that calibrated observations are used. For each instance (network traffic observation) an explicit latent variable \mathbf{z} with P components is introduced. As we will see, it corresponds to components in a principal-component subspace. Next, a Gaussian prior distribution $p(\mathbf{z})$ over the latent variable and a Gaussian observation distribution $p(\mathbf{x}|\mathbf{z})$ are introduced. Specifically:

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I}) \quad (5)$$

$$p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|\mathbf{W}\mathbf{z}, \sigma^2\mathbf{I}) \quad (6)$$

where \mathbf{W} is a $M \times P$ matrix whose columns span a linear (the principal-component) subspace within the data space, and the scale σ^2 governs the variance of the conditional distribution. Notice that, in what follows, we will omit the dependence of $p(\mathbf{x}|\mathbf{z})$ on \mathbf{W} and σ^2 for simplicity.

To estimate the values of the parameters \mathbf{W} and σ^2 , we use maximum likelihood, and the marginal distribution $p(\mathbf{x})$ is required. It can be easily calculated because the prior and the observation models in Eqs. (5) and (6) are both Gaussian. It follows that:

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z} = \mathcal{N}(\mathbf{x}|\mathbf{0}, \mathbf{C}), \quad (7)$$

where

$$\mathbf{C} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}. \quad (8)$$

The above likelihood requires the calculation of \mathbf{C} which may consume a lot of computational resources. This can be alleviated when N is larger than P (the dimension of the principal component subspace) and by utilizing the matrix inversion identity

$$\mathbf{C}^{-1} = \sigma^{-2}(\mathbf{I} - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^T), \quad (9)$$

where

$$\mathbf{M} = \mathbf{W}^T\mathbf{W} + \sigma^2\mathbf{I}. \quad (10)$$

Using our normalized and calibrated observations, the maximum likelihood estimates are calculated by solving

$$\begin{aligned} \mathbf{W}_{\text{ML}}, \sigma_{\text{ML}}^2 &= \arg \max_{\mathbf{W}, \sigma^2} \sum_{n=1}^N \ln p(\mathbf{x}_n|\mathbf{W}, \sigma^2) \\ &= \arg \max_{\mathbf{W}, \sigma^2} -\frac{NP}{2} \ln(2\pi) - \frac{N}{2} \ln |\mathbf{C}| - \frac{1}{2} \sum_{n=1}^N \mathbf{x}_n^T \mathbf{C}^{-1} \mathbf{x}_n \\ &= \arg \max_{\mathbf{W}, \sigma^2} -\frac{N}{2} \{P \ln(2\pi) + \ln |\mathbf{C}| + \text{tr}(\mathbf{C}^{-1}\mathbf{S})\} \end{aligned} \quad (11)$$

where \mathbf{S} is the data normalized and calibrated covariance matrix.

Maximizing the above equation with respect to \mathbf{W} and σ^2 is not easy. However, it can be shown —see [32]— that

$$\mathbf{W}_{\text{ML}} = \mathbf{U}(\mathbf{L} - \sigma^2\mathbf{I})^{1/2}\mathbf{R} \quad (12)$$

where \mathbf{U} is a $M \times P$ matrix whose columns correspond to the P eigenvectors associated with the P largest eigenvalues of the data normalized and calibrated covariance matrix \mathbf{S} , $\lambda_1, \dots, \lambda_P$. \mathbf{L} is a diagonal matrix with diagonal values these eigenvalues, and \mathbf{R} is an $P \times P$ orthonormal matrix that represents any rotation. Note that σ^2 is constrained to be smaller than the lowest eigenvalue λ_P (the minimum element in the diagonal matrix \mathbf{L} , thus avoiding a negative square root in Eq. (12)). Thus, $\sigma^2 \in [0, \lambda_P)$. Furthermore, the maximum likelihood for σ^2 is

$$\sigma_{\text{ML}}^2 = \frac{1}{M-P} \sum_{i=P+1}^M \lambda_i \quad (13)$$

Note that with the estimated \mathbf{W}_{ML} and σ_{ML}^2 we can calculate, for a given normalized sample \mathbf{x} , the quantity

$$\ln p(\mathbf{x} | \mathbf{W}_{\text{ML}}, \sigma_{\text{ML}}^2) = -\frac{M}{2} \ln(2\pi) - \frac{1}{2} \ln |\mathbf{C}| - \frac{1}{2} \mathbf{x}^T \mathbf{C}^{-1} \mathbf{x} \quad (14)$$

and use it to decide whether \mathbf{x} is an anomaly (we have omitted the dependency of \mathbf{C} on \mathbf{W}_{ML} and σ_{ML}^2 for simplicity). The whole process for anomaly detection in PPCA is represented in algorithm 1. First, the data need to be mean-centered around zero and scaled as PCA and PPCA are sensitive to feature scaling. Then, the parameters of the model are estimated using Eq. (13) and Eq (12) on trusted calibration data. Once the parameters are fixed, new data can be checked for anomalies.

Algorithm 1 PPCA for anomaly detection.

Require: Centered and normalized observations \mathbf{X} , predefined threshold thr .

Obtain σ_{ML}^2 using Eq. (13).

Obtain \mathbf{W}_{ML} using Eq. (12).

For a new sample \mathbf{x}^{new} , decide whether it is an anomaly by thresholding $\ln p(\mathbf{x}^{\text{new}} | \mathbf{W}_{\text{ML}}, \sigma_{\text{ML}}^2)$ defined in Eq. (14), that is if,

$$\frac{1}{2} \mathbf{x}^T \mathbf{C}^{-1} \mathbf{x} > thr. \quad (15)$$

Notice that Algorithm 1 can be used even with a value σ^2 that is different from σ_{ML}^2 . The only constraint on σ^2 is that it has to be smaller than λ_P , see Eq. (12).

Using $\frac{1}{2} \mathbf{x}^T \mathbf{C}^{-1} \mathbf{x} > thr$ to detect anomalies does not provide an insight into the roles played by the prior and the observation models in the detection process. This is crucial since it will allow us to relate PPCA and MSNM, as we will see in the following section.

V. RELATING PPCA TO MSNM

A. Revisiting PPCA for anomaly detection

In order to relate PPCA and MSNM we will evaluate other variance values apart from σ_{ML}^2 , in the expressions for PPCA. Thus, for clarity, we will make the formulations using a generic variance δ (smaller than λ_P). We will use the Laplace approximation [32] described in Appendix A as a starting point for the analysis. Given a marginal probability $p(\mathbf{x} | \mathbf{W}, \delta)$, with \mathbf{W} and δ being its score matrix and variance, and following Eq. (33), it follows that:

$$\ln p(\mathbf{x} | \mathbf{W}, \delta) = -\frac{1}{2} (\hat{\mathbf{z}}^T \hat{\mathbf{z}} + \frac{1}{\delta} \| \mathbf{x} - \mathbf{W} \hat{\mathbf{z}} \|^2) + \text{const.} \quad (16)$$

where $\hat{\mathbf{z}}$ is the mode of the posterior distribution $p(\mathbf{z}|\mathbf{x})$.

Therefore, leaving constants aside for the sake of clarity, we can equalise the thresholding expression of $\ln p(x|\mathbf{W}, \delta)$ for PPCA —Eq. (14)— to its Laplace approximation:

$$\frac{1}{2}\mathbf{x}^T\mathbf{C}^{-1}\mathbf{x} = \frac{1}{2}(\hat{\mathbf{z}}^T\hat{\mathbf{z}} + \frac{1}{\delta} \|\mathbf{x} - \mathbf{W}\hat{\mathbf{z}}\|^2) \quad (17)$$

In order to calculate the term $\hat{\mathbf{z}}$ in Eq. (17), the closed expression for $p(\mathbf{z}|x)$ derived in [32] can be used:

$$p(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}|\mathbf{M}^{-1}\mathbf{W}^T\mathbf{x}, \delta\mathbf{M}^{-1}) \quad (18)$$

and since for this Gaussian distribution the mode and the mean coincide, it follows that:

$$\hat{\mathbf{z}} = \mathbf{M}^{-1}\mathbf{W}^T\mathbf{x} \quad (19)$$

As seen in Eq. (12) (Section IV), \mathbf{W} depends on the variance $\delta \in [0, \lambda_P)$:

$$\mathbf{W}(\delta) = \mathbf{U}(\mathbf{L} - \delta\mathbf{I})^{1/2}, \quad (20)$$

where for the sake of simplicity, $\mathbf{R} = \mathbf{I}$ is used in (12). Introducing Eq. (20) in Eq. (10), and taking into account that $\mathbf{U}^T\mathbf{U} = \mathbf{I}$, it follows that:

$$\mathbf{M} = \mathbf{W}(\delta)^T\mathbf{W}(\delta) + \delta\mathbf{I} = (\mathbf{L} - \delta\mathbf{I}) + \delta\mathbf{I} = \mathbf{L} \quad (21)$$

Using Eqs. (20) and (21) in Eq. (19):

$$\hat{\mathbf{z}}(\delta) = \mathbf{M}^{-1}\mathbf{W}(\delta)^T\mathbf{x} = \mathbf{L}^{-1}(\mathbf{L} - \delta\mathbf{I})^{1/2}\mathbf{U}^T\mathbf{x}. \quad (22)$$

In short, the whole process for anomaly detection with this revisited PPCA is summarized in Algorithm 2. Note that it is equivalent to Algorithm 1, and allows arbitrary variance values to be used. The parameters of the model are estimated on trusted calibration data before checking new data for anomalies.

Algorithm 2 PPCA for anomaly detection (Revisited)

Require: Centered and normalized observations \mathbf{X} , predefined threshold thr .

Set a certain variance $\delta \in [0, \lambda_P)$.

Obtain $\mathbf{W}(\delta)$ using Eq. (20).

Obtain $\mathbf{M}(\delta)$ using Eq. (21).

Obtain $\hat{\mathbf{z}}(\delta)$ using Eq. (22).

For a new sample \mathbf{x}^{new} , decide whether it is an anomaly by thresholding $\ln p(\mathbf{x}^{new}|\mathbf{W}(\delta), \delta)$ defined in Eq. (16), that is if,

$$\begin{aligned} \frac{1}{2}(\hat{\mathbf{z}}^T(\delta)\hat{\mathbf{z}}(\delta) + \frac{1}{\delta} \|\mathbf{x} - \mathbf{W}(\delta)\hat{\mathbf{z}}(\delta)\|^2) \\ > thr \end{aligned} \quad (23)$$

B. Analysis of the variance: connecting PPCA and MSNM

Recall that Algorithm 2 can be used with a value of $\delta \in [0, \lambda_P)$ which is different from σ_{ML}^2 . We can observe the connection between PPCA and MSNM when studying the range of values that this variance δ might take. To do so, let us analyze the expression of $\ln p(\mathbf{x}|W, \delta)$ in Eq. (16). Leaving aside constant terms, Eq. (16) is formed using two terms, the first of which considers the influence of the latent space in the probability calculation, and somehow acts as a *regularization term*:

$$\hat{\mathbf{z}}(\delta)^T \hat{\mathbf{z}}(\delta) \quad (24)$$

while the second evaluates the difference between a sample \mathbf{x} and its reconstruction from the latent space. Thus, it plays the role of a *reconstruction error*:

$$\| \mathbf{x} - W\hat{\mathbf{z}}(\delta) \|^2 \quad (25)$$

Note that the reconstruction error contribution to the sample's probability is weighted by a factor $1/\delta$. At this point, we separate the influence of δ in this weighting factor from that of δ in $\mathbf{W}(\delta)$ and $\hat{\mathbf{z}}(\delta)$. Thus, we let the weighting factor take a fixed value $1/\alpha$ while δ keeps taking possible values in $[0, \lambda_P)$. Thus, a function $f_\alpha(\delta)$ can be defined to study the behavior of $\ln p(\mathbf{x}|W, \delta)$ for different values $\delta \in [0, \lambda_P)$:

$$f_\alpha(\delta) = \frac{1}{2}(\hat{\mathbf{z}}^T(\delta))\hat{\mathbf{z}}(\delta) + \frac{1}{\alpha} \| \mathbf{x} - \mathbf{W}(\delta)\hat{\mathbf{z}}(\delta) \|^2 \quad (26)$$

It can be seen that —see Appendix B and Figure 1— $f_\alpha(\delta)$ in Eq. (26) has 3 properties:

- (i) $f_\alpha(\delta)$ is convex,
- (ii) its minimum value is achieved at $\delta = \alpha/2$,
- (iii) $f_\alpha(0) = f_\alpha(\alpha)$.

Based on the fact that MSNM uses PCA for its modelling and PPCA converges to PCA when $\delta = 0$ — see [32]—, we have explored the value of $f_\alpha(\delta)$ for $\delta = 0$:

$$f_\alpha(0) = \frac{1}{2}(\mathbf{x}^T \mathbf{U} \mathbf{L}^{-1} \mathbf{U}^T \mathbf{x} + \frac{1}{\alpha} \| \mathbf{x} - \mathbf{U} \mathbf{U}^T \mathbf{x} \|^2) \quad (27)$$

Eq. (27) is exactly the quantity used by MSNM to detect anomalies since:

- $\mathbf{x}^T \mathbf{U} \mathbf{L}^{-1} \mathbf{U}^T \mathbf{x}$, the regularization part of the probability, matches Eq. (3) that defines the D -statistic.
- $\| \mathbf{x} - \mathbf{U} \mathbf{U}^T \mathbf{x} \|^2$, the reconstruction error part of the probability, matches Eq. (2) that defines the Q -statistic.
- Both terms, regularization and reconstruction error, contribute with a different weighting: *i)* $1/\sigma_{\text{ML}}^2$ in PPCA, and *ii)* the empirical value $(M - P/M)$ —see Eq. (4)— in MSNM.

The results obtained show that the conditions that make MSNM and PPCA coincide are:

$$\begin{aligned} \alpha &= \sigma_{\text{ML}}^2 \\ \delta &= \sigma_{\text{ML}}^2 \text{ or } \delta = 0 \end{aligned} \quad (28)$$

The behavior of the function $f_\alpha(\delta)$ is shown in Figure 1.

Notice also that in this *revisited version of PPCA* we could use a value of δ that is different from zero and σ_{ML}^2 . In that case, for a given anomaly probability threshold, the anomaly detection model would accept less observations (lower probability) if $\delta \in [0, \alpha]$ and more if $\alpha < \delta < \lambda_P$.

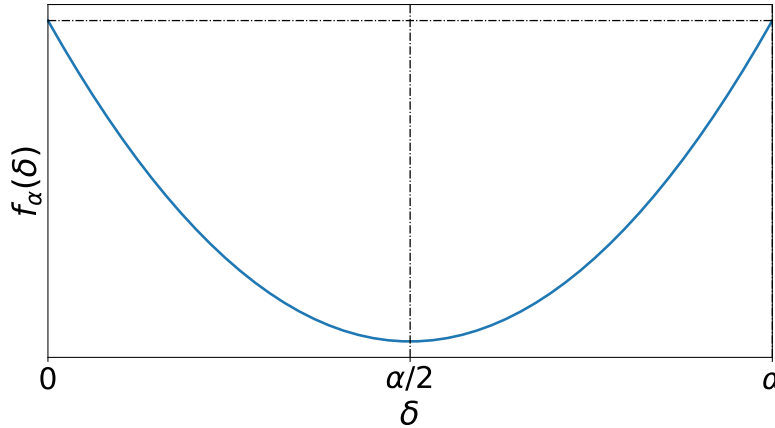


Figure 1. Analysis of $f_\alpha(\delta)$.

To end Section V, note that the conclusions obtained for PPCA are applicable to other generative models based on PPCA, like Variational Autoencoders (VAEs). In VAEs, the calculation of $p(\mathbf{z}|\mathbf{x})$ in a closed form is not possible and, thus, we must rely on the so called Evidence Lower Bound (ELBO), which comprises two different terms (as in PPCA): a reconstruction error and a regularization term. Our conclusions for PPCA lead to the same recommendations for VAEs, *i.e.*, both terms should be used in the calculation of anomaly scores. We consider this to be a relevant contribution of this paper, as we can still find examples in the state-of-the-art [14] [28] [29] [30] that only use the error reconstruction term for anomaly detection.

VI. EXPERIMENTAL VALIDATION

A. Datasets Description

To experimentally validate the theory introduced in the previous sections, we will make use of two datasets: one with synthetic data, and another with real network traffic information, as described below.

1) *Synthetic dataset*: The synthetic dataset is intended to provide a 2-dimensional plottable scenario that provides an easy interpretation of the anomalies and behaviour of the different models. Using the PCA observation model given by $\mathbf{x} = \mathbf{W}\mathbf{z} + \epsilon$, we obtain N samples of \mathbf{z} from a Gaussian distribution with zero mean and unit variance. Then we multiply \mathbf{z} by an arbitrary $\mathbf{W} = [0.707, 0.707]^T$ and add ϵ sampled from $\mathcal{N}(\mathbf{0}, 0.1 \cdot \mathbf{I})$ to obtain N bidimensional samples of \mathbf{x} . The samples will follow a Gaussian distribution along the selected \mathbf{W} . Using this method, we generate 1000 samples of clean data to calibrate our models. For the testing set, we mix 1000 new clean data samples with two different types of anomalies. The first type of anomaly is sampled from a distribution that is different from that of the above described samples. Specifically, we choose a multivariate Gaussian with zero mean and a diagonal variance matrix $5 \times \mathbf{I}$. Anomaly type 1 is not designed according to the linear model. The second type of anomaly is generated in the latent space, we sample z_{anom} using a Gaussian with mean 5 and unit variance. The values of z_{anom} are then randomly multiplied by -1 following a Bernoulli (0.5) distribution. This procedure provides anomalies that follow the generative model but whose latent distribution

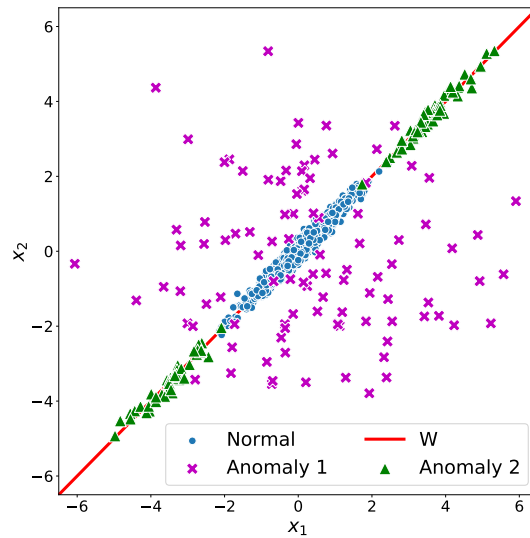


Figure 2. Synthetic dataset test bench. Normal data is presented in blue circles, Anomaly type 1 is presented in fuchsia with x-shaped markers, while the green triangles refer to Anomaly type 2. The arbitrary line \mathbf{W} used to create the data is plotted in red.

is different from that of the calibration data. For each type of anomaly mentioned, 100 data points are introduced. See Figure 2 for a visual representation of the data and anomalies.

2) *UGR'16 dataset*: The UGR'16 dataset [33] was designed for the evaluation of cyclostationary-based network IDSs. It contains real anonymized NetFlow traces captured over several months in a tier-3 ISP. The traces include legitimate traffic from many virtualized services in the cloud, like web servers with proprietary and standard configurations, and other hosted services like DNS, FTP, mail servers, etc. In addition, a set of malicious virtual machines was configured in the network to generate attack traffic. The traffic is captured from two border routers of the ISP network. Thus, the dataset includes both legitimate traffic and realistic attack scenarios, all of them labeled.

UGR'16 divides the data in two differentiated sets: calibration and test. The calibration set contains real background traffic data gathered from March to June in 2016 (4 months). The test set mixes real background traffic and synthetically generated malicious traffic, gathered from July to August 2016. Although the data was captured in 2016, it was published in 2018 and is still considered of interest in recent work [34], [35].

To train our models, we use the data gathered during working days in May, where less anomalies were detected after data obtained was analyzed [33], and no synthetically generated attacks were introduced. The test set uses data gathered on those working days when synthetically generated attacks were interlaced within background traffic. The attack types that were synthetically generated include:

- Low-rate Denial of Service (DoS): TCP SYN packets are sent to the victims port 80 using 1280-bit packets and a rate of 100 packets/s. The rate is not high enough to avoid the normal operation of the network being affected.
- Port scanning: A continuous SYN scan of common ports of victims. Two variants are implemented for this attack: Scan11 (one-to-one scan attack) and Scan44 (Four-to-four scan

Table I
SUMMARY OF DATA FLOWS FROM UGR'16 USED TO TRAIN AND TEST THE MODELS

Flow type	Calibration	Test
Background traffic	31680	8714
DoS	NA	299
Scan44	NA	65
Scan11	NA	66
Nerisbotnet	NA	488
UDPscan	NA	9
SSHscan	NA	9
Spam	NA	3616

attack).

- Botnet traffic: Obtained from the execution of the malware known as Neris in a controlled environment (See [36] for details about the malware and [33] for details about its injection in the data.).

The test set also included labels for a real UDP Scan campaign that was identified in the background traffic.

The labels are provided as a list of timestamps (in mins) of when the attacks were executed. Table I summarizes the traces in the train and test sets. The NetFlow logs cannot be directly used to feed PCA-based anomaly detection systems [8]. Thus, following [11], we use the FCParse² tool to extract 143 quantitative features from the NetFlow logs. Each feature counts the number of times that a given event takes place during each minute, *e.g.*, number of flows with a given destination port, number of flows with an accumulated payload size greater than a threshold, etc. Features were manually defined in [11] from domain knowledge using regular expressions. Therefore, they represent information that experts would use to manually identify anomalies in the data. In this paper, we focus on differentiating the minutes labeled as anomalies from those labeled as background data. We study each type of anomaly/attack separately. The anomaly scores are calculated for the whole test set at once. Then, classification metrics are calculated for each anomaly type against background data only (binary classification)

The evaluation of the variance captured by the different principal components of PCA on the calibration set (see Figure 3), shows that the principal components 1-5 are the most relevant and capture a higher percentage of variance than the rest. The model with 5 P explains 33.65% of the total variance.

B. Experiments

This section focuses on describing three different experiments: *i)* Anomaly detection using the MSNM and PPCA models with a common threshold to show that detection results from both models coincide. *ii)* Anomaly detection using only either the regularization or the reconstruction error term separately, to analyze their performance for different types of anomalies. *iii)* Finally, we assess the correctness of the weighting values for both regularization and reconstruction error terms, for MSNM and PPCA models.

1) MSNM and PPCA equivalence: To experimentally demonstrate the MSNM and PPCA equivalency shown in Section V, we perform anomaly detection on the datasets presented in

²Available at: <https://github.com/josecamachop/FCParser>

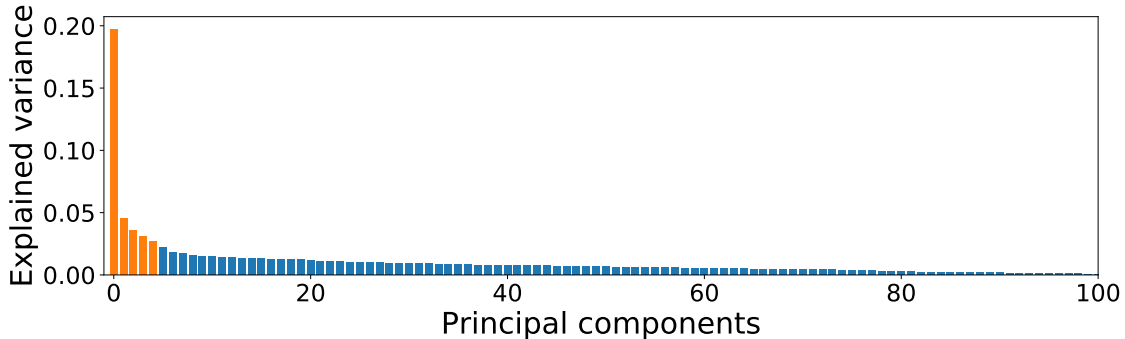


Figure 3. Explained variance per component in the UGR'16 dataset.

Table II
PERFORMANCE METRICS FOR THE SYNTHETIC DATASET USING DIFFERENT THRESHOLD VALUES SELECTED ACCORDING TO DIFFERENT CONFIDENCE LEVELS ON THE CALIBRATION SET.

Confidence level	95%	96%	97%	98%	99%
Threshold χ^2	5.99	6.43	7.01	7.82	9.21
Accuracy	0.9925	0.9933	0.9925	0.9866	0.9808
False Alarm Ratio	0.002	0.001	0.001	0.001	0

Section VI-A using both MSNM and PPCA models. The anomaly score is calculated using Eq. (26) with the condition in Eq. (28), *i.e.*, $\alpha = \sigma_{\text{ML}}^2$ and $\delta = 0$ or $\delta = \sigma_{\text{ML}}^2$ for the MSNM and PPCA models, respectively. When a threshold value is needed, it should be chosen in accordance with the values in the calibration set, according to our confidence in the absence of malicious traffic in the training data. In our case, the threshold is the 99-percentile of the anomaly scores of the calibration set for each model. The threshold can be modified if the confidence in the calibration data is reduced.

Note also that once σ_{ML}^2 and \mathbf{W}_{ML} have been calculated, Eq. (14) can be used to set a non-experimental value. Since $p(x|\mathbf{W}_{\text{ML}}, \sigma_{\text{ML}}^2)$ is a normal distribution we can select a threshold δ such as

$$\int_{\mathbf{x}^T \mathbf{C}^{-1} \mathbf{x} \leq \delta} p(x|\mathbf{W}_{\text{ML}}, \sigma_{\text{ML}}^2) dx \geq \alpha \quad (29)$$

for a confidence level α . We can take into account that $\mathbf{x}^T \mathbf{C}^{-1} \mathbf{x} \sim \chi^2(M)$ and select for this distribution a threshold δ with a given confidence level α . This theoretical approach is less used in practice.

In Figure 4 we can see how the detection area changes according to the confidence level α . Table II includes the threshold value, accuracy and False Alarm Ratio using the theoretical bound $\mathbf{x}^T \mathbf{C}^{-1} \mathbf{x} \leq \delta$.

Figure 5 shows the parallelism between MSNM and PPCA models in the synthetic test set. The same data points are identified as anomalies by both models, obtaining identical ROC curves and AUC values. The accuracy (0.9858) and the False Alarm Ratio (0.012%) were also the same values for both models. The equivalence of both models is also tested on the UGR'16 dataset. Although the anomaly score is calculated for all testing data at once, we calculate the ROC curve

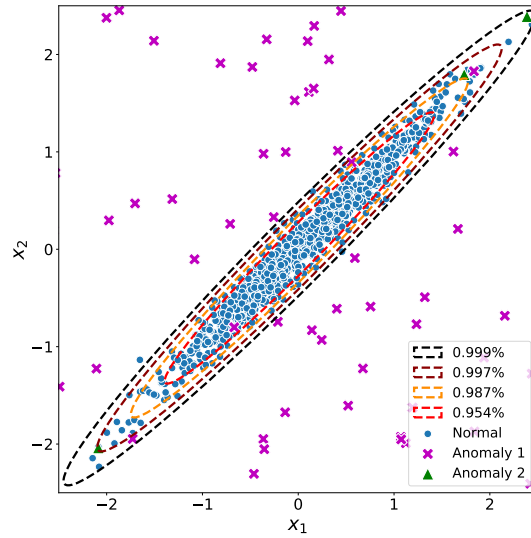


Figure 4. Detail of the synthetic test set shown in Figure 2. The detection area obtained with the threshold for different confidence levels α on the calibration data is shown in dashed lines .

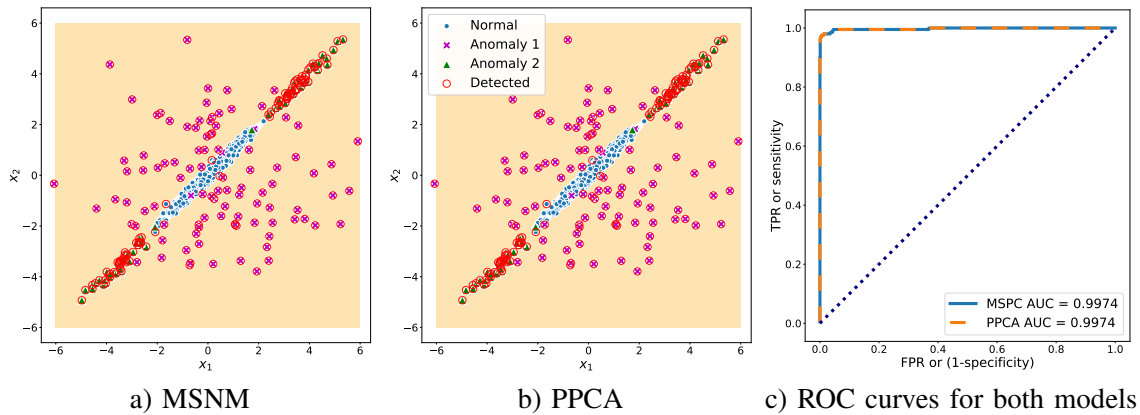


Figure 5. Anomaly detection on the synthetic dataset for both *a)* MSNM and *b)* PPCA models. Blue circles identify normal data, fuchsia xs and green triangles identify different types of anomalies. Red circumferences mark the points detected as anomalies. The orange shadows cover the detection area in each case. *c)* ROC curves for both models.

and AUC for each attack type against the background traffic. Table III includes identical results for both models when detecting different kind of attacks, using several numbers of principal components, $P = [1, 2, \dots, 5]$. For further details, Figure 6 shows the ROC curves obtained for both models in different attacks using $P = 3$.

2) *Using the regularization and reconstruction error terms to detect different types of anomalies:* As previously indicated, both terms in Eq. (26) are able to identify different behaviour in the data. To quickly gain an overview of how both terms work, Figure 7 depicts the anomalies in the synthetic dataset that are detected using a single PPCA/MSNM term only. Note that when using a single term, the threshold is the 99th-percentile of the calibration set for that term. When we only use the reconstruction error term —Eq. (25)—, we can see in Figure 7(a) how we are unable

Table III
AUC VALUES FOR THE ATTACKS ON UGR'16 USING MSNM AND PPCA

Attack type	1 P		2 P		3 P		4 P		5 P	
	MSNM	PPCA	MSNM	PPCA	MSNM	PPCA	MSNM	PPCA	MSNM	PPCA
DoS	0.9118	0.9118	0.9089	0.9089	0.9089	0.9089	0.9097	0.9097	0.9091	0.9091
Scan44	0.9903	0.9903	0.9902	0.9902	0.9896	0.9896	0.9882	0.9882	0.9880	0.9880
Scan11	0.9384	0.9384	0.9390	0.9390	0.9412	0.9412	0.9318	0.9318	0.9303	0.9303
Nerisbotnet	0.8204	0.8204	0.8211	0.8211	0.8198	0.8198	0.8201	0.8201	0.8203	0.8203
UDPscan	0.7826	0.7826	0.7844	0.7844	0.7707	0.7707	0.7707	0.7707	0.7727	0.7727
SSHscan	0.5593	0.5593	0.5624	0.5624	0.5569	0.5569	0.5588	0.5588	0.5614	0.5614
Spam	0.4669	0.4669	0.4610	0.4610	0.4588	0.4588	0.4585	0.4585	0.4512	0.4512

Table IV
AUC VALUES FOR THE ATTACKS ON UGR'16 USING DIVERGENCE (PRIOR) AND ERROR (OBSERV.) TERMS SEPARATELY.

Attack type	1 P		2 P		3 P		4 P		5 P	
	Prior	Observ.	Prior	Observ.	Prior	Observ.	Prior	Observ.	Prior	Observ.
DoS	0.5242	0.9118	0.9191	0.9085	0.9178	0.9085	0.9050	0.9095	0.9152	0.9087
Scan44	0.8747	0.9903	0.8483	0.9902	0.9968	0.9895	0.9960	0.9878	0.9957	0.9875
Scan11	0.4966	0.9385	0.3771	0.9391	0.7714	0.9416	0.9691	0.9304	0.9691	0.9285
Nerisbotnet	0.4358	0.8207	0.3686	0.8214	0.4741	0.8199	0.4480	0.8203	0.4926	0.8204
UDPscan	0.6618	0.7817	0.7084	0.7838	0.8886	0.7682	0.8659	0.7674	0.8372	0.7698
Mean	0.5986	0.8886	0.6443	0.8886	0.8097	0.8855	0.8368	0.8831	0.8419	0.8830

to identify anomalies that have been generated following the linear model. Note that this case is equivalent to using a simple PCA analysis. Using the regularization term only —Eq. (24)—, see Figure 7(b), results in a similar problem. This term is calculated using only the latent $\hat{z}(\delta)$, therefore all data points whose latent value is within the distribution of the normal data remains undetected. To effectively detect all anomalies, it is clearly necessary to use both terms.

When it comes to complex networking data, such as those in UGR'16, it is difficult to predict which anomalies will be correctly detected by the two terms in the PPCA/MSNM model. Table IV shows that different attacks are captured differently by both terms. Some attacks (DoS, Scan44) are captured correctly by both terms. Neris Botnet traffic can be identified using the reconstruction error term, but cannot usually be identified with the regularization term.

The results for Scan11 and UDP Scan attacks show that the ability of the model to identify anomalies using the error reconstruction or regularization terms depends on the number of principal components used. We have included Fig 8 as an example of where the discriminative power of the regularization term increases with P , even outperforming the error reconstruction term performance. Note that the error reconstruction term is not severely affected by the number of P because the explained variance of the model is low, and therefore the error reconstruction for most anomalies remains high. As the number of P increases, the latent space is able to capture more features of the data, also allowing anomalies in the latent domain to better identified.

3) *Weighting reconstruction error and regularization terms:* The previous experiments have shown that both terms in Eq. (26) are relevant and should be used. Furthermore, a single term is not usually able to correctly identify all types of anomalies. The forthcoming question is how to combine them to obtain a single anomaly score.

On the one hand, the work in [15] suggested the use of the significance levels UCL_Q and

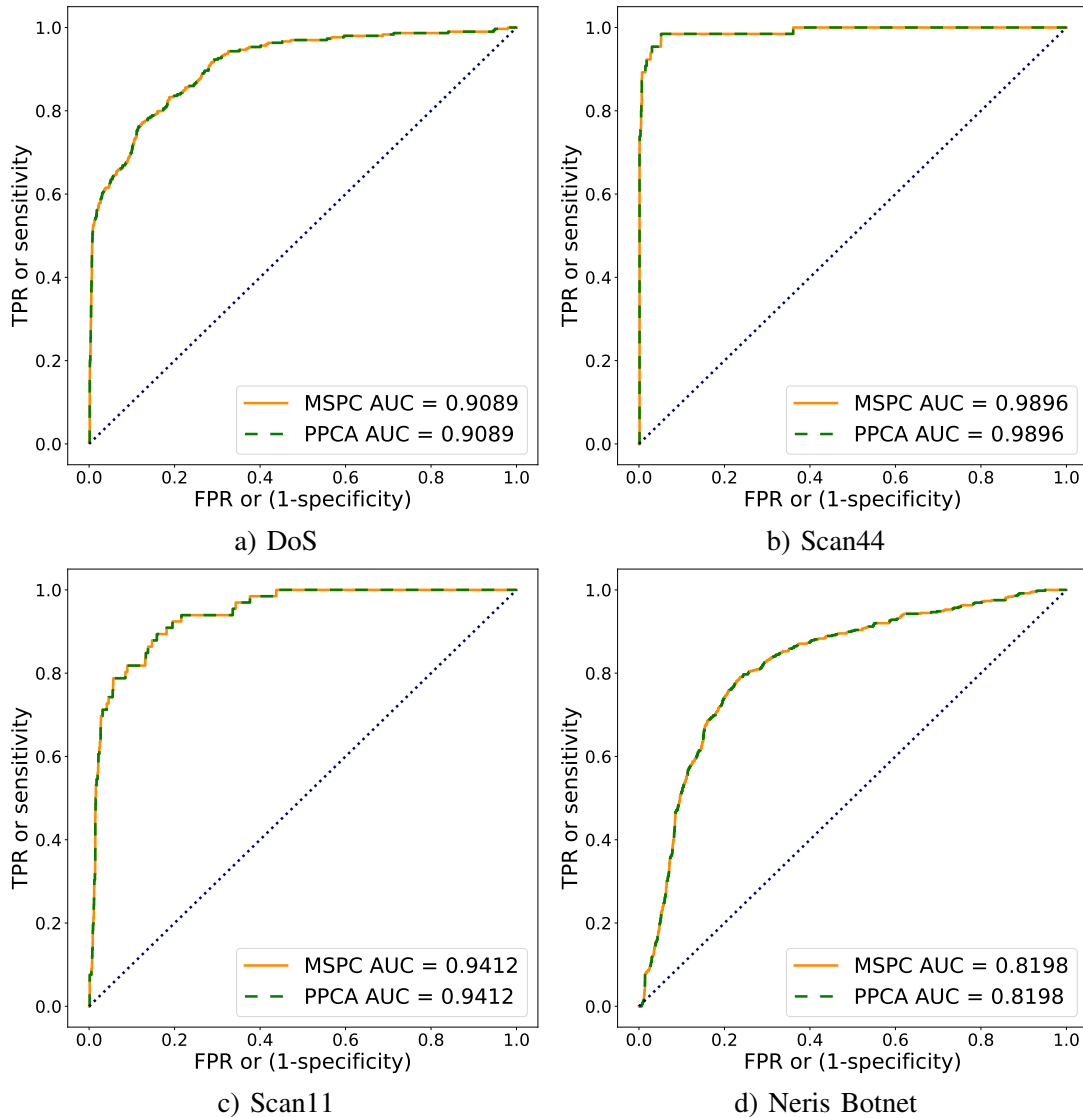


Figure 6. ROC curves for different attacks on UGR'16 using MSNM and PPCA with $P=3$.

UCL_D and a weighting parameter calculated as P/M (see Section III). On the other hand, PPCA provides a probabilistic interpretation for the weighting parameter, combining both terms according to their contribution to the marginal likelihood—see Eq. (17)—. In this section, we compare the MSNM approach in Eq. (4) with the PPCA approach in Eq. (17).

In the synthetic dataset, both strategies obtain a similar result in AUC: 0.9974 for PPCA and 0.9973 for MSNM. Note that in both cases, the combination of both terms obtains a better AUC than the values reported in the previous section for each term separately.

The difference between the results from PPCA and MSNM is more clearly shown in Table VII, where the UGR'16 dataset is analyzed. Even when the MSNM approach results in slightly better AUC values for some combinations of attacks and number of P , the PPCA AUC score

Table V
ACCURACY VALUES FOR THE ATTACKS ON UGR'16 USING DIVERGENCE (PRIOR) AND ERROR (OBSERV.) TERMS SEPARATELY.

Attack type	1 P		2 P		3 P		4 P		5 P	
	Prior	Observ.	Prior	Observ.	Prior	Observ.	Prior	Observ.	Prior	Observ.
Dos	0.9237	0.8946	0.9655	0.8939	0.9273	0.8932	0.8939	0.8943	0.9287	0.8922
scan44	0.9507	0.9011	0.9713	0.9008	0.9320	0.9004	0.8979	0.9015	0.9347	0.8991
scan11	0.9460	0.8999	0.9667	0.8995	0.9300	0.8992	0.8974	0.9001	0.9337	0.8977
nerisbotnet	0.9022	0.8801	0.9223	0.8798	0.8850	0.8795	0.8541	0.8806	0.8880	0.8788
anomaly-udpscan	0.9520	0.9001	0.9728	0.8998	0.9312	0.8995	0.8969	0.9005	0.9340	0.8981
mean	0.9349	0.8952	0.9597	0.8948	0.9211	0.8943	0.8880	0.8954	0.9238	0.8932

Table VI
MEAN FALSE ALARM RATIO ON UGR'16 USING DIVERGENCE (PRIOR) AND ERROR (OBSERV.) TERMS SEPARATELY.

1 P		2 P		3 P		4 P		5 P	
Prior	Observ.	Prior	Observ.	Prior	Observ.	Prior	Observ.	Prior	Observ.
0.0474	0.0995	0.0267	0.0998	0.0684	0.1002	0.1027	0.0992	0.0656	0.1016

Table VII
AUC VALUES FOR THE ATTACKS ON UGR'16 USING DIFFERENT NUMBER OF PCs AND MODELS: MSNM —Eq. (4)— AND PCCA —Eq. (17)—.

Attack type	1 P		2 P		3 P		4 P		5 P	
	MSNM	PCCA	MSNM	PCCA	MSNM	PCCA	MSNM	PCCA	MSNM	PCCA
DoS	0.9118	0.9118	0.9085	0.9089	0.9085	0.9089	0.9095	0.9097	0.9087	0.9091
Scan44	0.9903	0.9903	0.9902	0.9902	0.9895	0.9896	0.9878	0.9882	0.9875	0.9880
Scan11	0.9385	0.9384	0.9391	0.9390	0.9415	0.9412	0.9305	0.9318	0.9286	0.9303
Nerisbotnet	0.8207	0.8204	0.8214	0.8211	0.8199	0.8198	0.8203	0.8201	0.8204	0.8203
UDPscan	0.7817	0.7826	0.7838	0.7844	0.7683	0.7707	0.7674	0.7707	0.7700	0.7727
Mean	0.8886	0.8887	0.8886	0.8887	0.8855	0.8860	0.8831	0.8841	0.8830	0.8841

Table VIII
ACCURACY VALUES FOR THE ATTACKS ON UGR'16 USING DIFFERENT NUMBER OF PCs AND MODELS: MSNM —Eq. (4)— AND PCCA —Eq. (17)—.

	MSNM	PCCA	MSNM	PCCA	MSNM	PCCA	MSNM	PCCA	MSNM	PCCA
Dos	0.8946	0.8947	0.8939	0.8946	0.8932	0.9655	0.8944	0.8947	0.8922	0.8932
scan44	0.9011	0.9012	0.9008	0.9011	0.9004	0.9713	0.9015	0.9017	0.8991	0.9004
scan11	0.8999	0.9000	0.8995	0.8999	0.8992	0.9667	0.9001	0.9003	0.8977	0.8992
nerisbotnet	0.8801	0.8802	0.8798	0.8801	0.8794	0.9223	0.8806	0.8807	0.8788	0.8795
anomaly-udpscan	0.9001	0.9003	0.8998	0.9001	0.8995	0.9728	0.9005	0.9007	0.8981	0.8995
mean	0.8952	0.8953	0.8948	0.8952	0.8943	0.9597	0.8954	0.8956	0.8932	0.8943

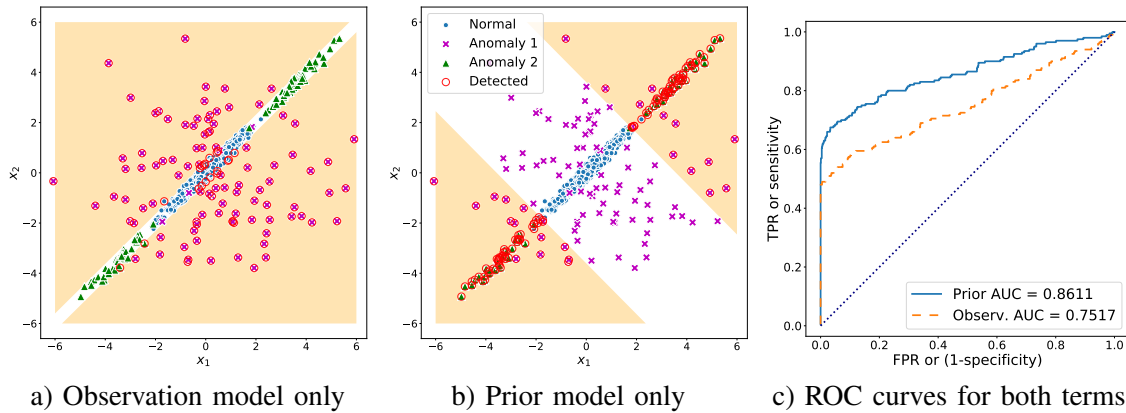


Figure 7. Detected anomalies in the synthetic dataset using only a single term. Blue circles identify normal data, fuchsia xs and green triangles identify different types of anomalies. Red circumferences mark the points detected as anomalies. The orange shadows cover the detection area in each case.

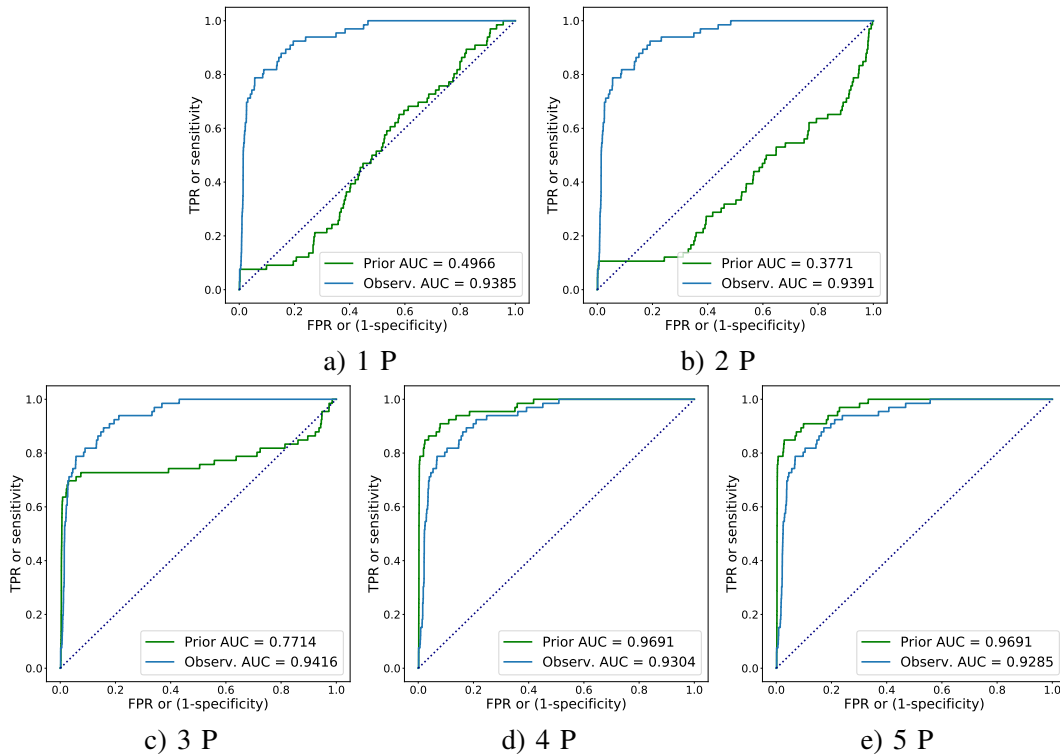


Figure 8. Evolution of the ROC curves of both divergence and error terms on the Scan11 attack for different latent space sizes.

yields a better mean value in each case. While the differences are subtle, they indicate that PPCA is better when it comes to combining information from both regularization and reconstruction error terms. Looking at Table VII, we can observe that PPCA obtains better values than MSNM in those attacks where the regularization term is more informative. In addition, note that while a single term might be better at recognising a given attack type, the mean AUC values are always

Table IX
 MEAN FALSE ALARM RATIO ON UGR'16 USING DIFFERENT NUMBER OF PCS AND MODELS: MSNM —EQ. (4)— AND
 PPCA —EQ. (17)—.

1 P		2 P		3 P		4 P		5 P	
MSNM	PPCA	MSNM	PPCA	MSNM	PPCA	MSNM	PPCA	MSNM	PPCA
0.0995	0.0994	0.0998	0.1000	0.1002	0.1000	0.0992	0.0989	0.1016	0.1018

higher than those obtained using only a single term.

VII. CONCLUSIONS

In this paper we have analyzed the use of the generative model known as probabilistic PCA (PPCA) for detection of anomalies in network security. Specifically, we have provided a detailed mathematical model that connects MSNM, a well-known framework for network anomaly detection, and PPCA. The generative PPCA model provides a probabilistic point of view to explain the MSNM framework and the meaning of its principal elements: The use of Q and D statistics, which are derived as an error reconstruction term and a regularization term, respectively, in the PPCA formulation.

Understanding the role of both terms in the anomaly detection process is a key step towards the correct use of generative models in this security research field. Specifically, a direct application is that of correctly using other generative models like VAEs and GANs. In a review of research works that use these models, we note that while the error reconstruction term is widely used due to its high anomaly detection capability, the regularization term is often forgotten or discarded. We have theoretically and experimentally assessed that both terms are relevant and capture complementary information. This implies that they should be used together for a robust anomaly detection.

In addition, the PPCA generative framework provides a combination of both terms that considers their contribution to the marginal distribution $p(\mathbf{x})$ with a probabilistic interpretation, thus offering a non-empirical solution for the weighting parameter required by MSNM.

Although the linear PPCA generative model is easy to understand and helps obtain the above conclusions, its simplicity limits its generalization to non-linear data. Non-linearity is often present in real traffic data and would be difficult to capture and detect. Also, PPCA inherits some of the disadvantages from PCA, which is quite common when working with latent space models. Even with a linear model, the latent combination of the original features is not easy to interpret. So, as shown in Figure 8, the choice of the latent space size P is critical for the detection of some attacks. While the use of linear detection models is useful in the later diagnosis of network incidents, this paper intends to establish a first step for more complex generative approaches to network anomaly detection. Further research on the combination of both error reconstruction and regularization terms for those models is also needed. Finally, the choice of the threshold is a well-known problem for anomaly detection. The calibration set provides a benchmark to choose the decision boundary, but it should be determined according to the confidence in the calibration data. The Gaussian form of $p(x|\mathbf{W}_{ML}, \sigma_{ML}^2)$ allows us to choose the confidence-based threshold. Although the threshold might also be experimentally determined with the use of testing sets, this approach relies on known attacks, and therefore does not provide information about the optimal threshold for new or unknown attacks. Finally, system requirements usually determine the threshold to use in industry applications. A high cost of undetected false negatives

might induce the use of a lower threshold that will produce a higher number of false positives. When combining both decision terms, the use of a combined or separated threshold for each term is also an open field for future research.

ACKNOWLEDGMENT

This work was sponsored in part by the Agencia Estatal de Investigación under project PID2019-105142RB-C22/AEI/10.13039/501100011033, Spanish MINECO (Ministerio de Economía y Competitividad) project TIN2017-83494-R and FEDER/Junta de Andalucía-Consejería de Transformación Económica, Industria y Universidades/ project A-TIC-215-UGR18. The work by Fernando Pérez-Bueno was sponsored by Ministerio de Economía, Industria y Competitividad under FPI contract BES-2017-081584.

The authors would like to thank Daniel Cortés Troya for his collaboration in the early stages of this work.

APPENDIX

A. Laplace approximation

It follows that:

$$p(\mathbf{x}) = \int p(\mathbf{z})p(\mathbf{x}|\mathbf{z})d\mathbf{z} = \int \exp[\ln p(\mathbf{z})p(\mathbf{x}|\mathbf{z})]d\mathbf{z} \quad (30)$$

where $f(\mathbf{z}) = \ln(p(\mathbf{z})p(\mathbf{x}|\mathbf{z}))$ is a quadratic function that can be expanded around the maximum a posteriori (MAP)

$$\hat{\mathbf{z}} = \arg \max_{\mathbf{z}} f(\mathbf{z}) = \arg \max_{\mathbf{z}} \ln(p(\mathbf{z})p(\mathbf{x}|\mathbf{z})) \quad (31)$$

to obtain

$$\begin{aligned} p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) &= \exp[\ln p(\mathbf{z})p(\mathbf{x}|\mathbf{z})] \\ &\propto \exp\left[f(\hat{\mathbf{z}}) - \frac{1}{2}(\mathbf{z} - \hat{\mathbf{z}})^T \left(\mathbf{I} + \frac{1}{\sigma^2} \mathbf{W}^T \mathbf{W}\right) (\mathbf{z} - \hat{\mathbf{z}})\right], \end{aligned} \quad (32)$$

which produces, from Eq. (30), $\ln p(\mathbf{x}) = f(\hat{\mathbf{z}}) + \text{const}$, and so

$$\ln p(\mathbf{x}|\mathbf{W}, \sigma^2) = -\frac{1}{2}(\hat{\mathbf{z}}^T \hat{\mathbf{z}} + \frac{1}{\sigma^2} \|\mathbf{x} - \mathbf{W}\hat{\mathbf{z}}\|^2) + \text{const}. \quad (33)$$

B. Analysis of $f_\alpha(\delta)$

Let us define

$$f_\alpha(\delta) = \frac{1}{2}(\hat{\mathbf{z}}^T(\delta)\hat{\mathbf{z}}(\delta) + \frac{1}{\alpha} \|\mathbf{x} - \mathbf{W}(\delta)\hat{\mathbf{z}}(\delta)\|^2), \quad (34)$$

and study its properties.

We observe that this function depends on \mathbf{x} but we do not make this dependency explicit in order to simplify the notation. Note also that we will end up studying the basic properties of the associated quadratic form.

Theorem 1. *Let us consider $f_\alpha(\delta)$ defined in Eq. (34) with $\alpha < \lambda_P$ and $\delta \in [0, \lambda_P)$. Then*

- $f_\alpha(\delta)$ is a convex function on δ with minimum at $\alpha/2$,
- and $f_\alpha(0) = f_\alpha(\alpha)$.

Proof. Therefore,

$$\begin{aligned}
f_\alpha(\delta) &= \frac{1}{2}(\mathbf{x}^T \mathbf{U}(\mathbf{L} - \delta \mathbf{I})\mathbf{L}^{-2} \mathbf{U}^T \mathbf{x} + \frac{1}{\alpha} \|\mathbf{x} - \mathbf{U}(\mathbf{L} - \delta \mathbf{I})\mathbf{L}^{-1} \mathbf{U}^T \mathbf{x}\|^2) \\
&= \frac{1}{2}(\mathbf{x}^T \mathbf{U}(\mathbf{L} - \delta \mathbf{I})\mathbf{L}^{-2} \mathbf{U}^T \mathbf{x} \\
&\quad + \frac{1}{\alpha}(\mathbf{x}^T \mathbf{x} + \mathbf{x}^T \mathbf{U}(\mathbf{L} - \delta \mathbf{I})\mathbf{L}^{-2} \mathbf{U}^T \mathbf{x} - 2\mathbf{x}^T \mathbf{U}(\mathbf{L} - \delta \mathbf{I})\mathbf{L}^{-1} \mathbf{U}^T \mathbf{x}))
\end{aligned}$$

and

$$\begin{aligned}
f'_\alpha(\delta) &= \frac{1}{2}(-\mathbf{x}^T \mathbf{U} \mathbf{L}^{-2} \mathbf{U}^T \mathbf{x} \\
&\quad + \frac{1}{\alpha}(-2\mathbf{x}^T \mathbf{U}(\mathbf{L} - \delta \mathbf{I})\mathbf{L}^{-2} \mathbf{U}^T \mathbf{x} + 2\mathbf{x}^T \mathbf{U} \mathbf{L}^{-1} \mathbf{U}^T \mathbf{x})).
\end{aligned} \tag{35}$$

Furthermore, we have the following identity for the sum of the matrices involved in $f'_\alpha(\delta)$,

$$-\mathbf{L}^{-2} + \frac{2}{\alpha}(-(\mathbf{L} - \delta \mathbf{I})\mathbf{L}^{-2} + \mathbf{L}^{-1}) = \mathbf{L}^{-2}[-\mathbf{I} + \frac{2}{\alpha} \delta \mathbf{I}] \tag{36}$$

from which we can see that the sum is the zero matrix iff $\delta = \alpha/2$.

We also note that

$$f''_\alpha(\delta) = \frac{1}{2\alpha} \mathbf{x}^T \mathbf{U} \mathbf{L}^{-2} \mathbf{U}^T \mathbf{x} \geq 0 \tag{37}$$

So, $f_\alpha(\delta)$ is a convex quadratic function on δ whose minimum is achieved at $\delta = \alpha/2$.

Furthermore, using the Taylor expansion around the minimum it follows that

$$f_\alpha(\delta) = f_\alpha\left(\frac{\alpha}{2}\right) + \frac{1}{2}\left(\delta - \frac{\alpha}{2}\right)^2 f''_\alpha\left(\frac{\alpha}{2}\right) \tag{38}$$

and so $f_\alpha(0) = f_\alpha(\alpha)$.

In summary, $f''_\alpha(\delta)$ is convex, its minimum value is achieved at $\delta = \alpha/2$, and $f_\alpha(0) = f_\alpha(\alpha)$. \square

REFERENCES

- [1] R. Contu, D. Kish, C. Canales, S. Deshpande, E. Kim, and D. Gardner, "Forecast analysis: Information security, worldwide, 2q18 update," url<https://www.gartner.com/en/documents/3889055>, 2018.
- [2] New York Times, "Blackout Hits Iran Nuclear Site in What Appears to Be Israeli Sabotage," <https://www.nytimes.com/2021/04/11/world/middleeast/iran-nuclear-natanz.html>, 2021.
- [3] A. B. Nassif, M. A. Talib, Q. Nasir, and F. M. Dakalbab, "Machine learning for anomaly detection: A systematic review," *IEEE Access*, vol. 9, pp. 78 658–78 700, 2021.
- [4] D. Kwon, H. Kim, J. Kim, S. C. Suh, I. Kim, and K. J. Kim, "A survey of deep learning-based network anomaly detection," *Cluster Computing*, vol. 22, no. 1, pp. 949–961, 2019.
- [5] R. Chalapathy and S. Chawla, "Deep learning for anomaly detection: A survey," 2019.
- [6] L. Ruff, J. R. Kauffmann, R. A. Vandermeulen, G. Montavon, W. Samek, M. Kloft, T. G. Dietterich, and K.-R. Müller, "A unifying review of deep and shallow anomaly detection," *Proceedings of the IEEE*, 2021.
- [7] A. Lakhina, M. Crovella, and C. Diot, "Diagnosing network-wide traffic anomalies," *ACM SIGCOMM computer communication review*, vol. 34, no. 4, pp. 219–230, 2004.
- [8] —, "Characterization of network-wide anomalies in traffic flows," in *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, 2004, pp. 201–206.
- [9] G. Maciá-Fernández, J. Camacho, P. García-Teodoro, and R. A. Rodríguez-Gómez, "Hierarchical PCA-based multivariate statistical network monitoring for anomaly detection," in *2016 IEEE international workshop on information forensics and security (WIFS)*. IEEE, 2016, pp. 1–6.

- [10] H. Hotelling, "Multivariate quality control," *Techniques of Statistical Analysis*, 1947. [Online]. Available: <https://ci.nii.ac.jp/naid/10021322508/en/>
- [11] J. Camacho, J. M. García-Jiménez, N. M. Fuentes-García, and G. Maciá-Fernández, "Multivariate big data analysis for intrusion detection: 5 steps from the haystack to the needle," *Computers & Security*, vol. 87, p. 101603, 2019.
- [12] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 3, pp. 611–622, 1999.
- [13] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2014.
- [14] H. Xu, W. Chen, N. Zhao, Z. Li, J. Bu, Z. Li, Y. Liu, Y. Zhao, D. Pei, Y. Feng, J. Chen, Z. Wang, and H. Qiao, "Unsupervised Anomaly Detection via Variational Auto-Encoder for seasonal KPIs in Web Applications," in *Proc. of the 2018 World Wide Web Conference*, 2018, pp. 187–196.
- [15] J. Camacho, P. García-Teodoro, and G. Maciá-Fernández, "Traffic monitoring and diagnosis with multivariate statistical network monitoring: a case study," in *2017 IEEE security and privacy workshops (SPW)*. IEEE, 2017, pp. 241–246.
- [16] A. Soule, A. Lakhina, N. Taft, K. Papagiannaki, K. Salamantian, M. C. Antonio Nucci, and C. Diot, "Traffic matrices: balancing measurements, inference and modeling," vol. 33, no. 1. USENIX Association, Berkeley, CA, USA, 2005, pp. 331–344.
- [17] Y. Zhang, Z. Ge, A. Greenberg, and M. Roughan, "Network anomography," in *Proceedings of the 5th ACM SIGCOMM conference on Internet Measurement (IMC'05)*. USENIX Association, Berkeley, CA, USA, 2005, pp. 317–330.
- [18] L. Huang, X. Nguyen, M. Garofalakis, M. Jordan, A. Joseph, and N. Taft, "In-network pcas and anomaly detection," in *Proceedings of Neural Information Processing Systems (NIPS)*. NIPS Foundation, 2006.
- [19] H. Ringberg, A. Soule, J. Rexford, and C. Diot, "Sensitivity of PCA for traffic anomaly detection," *SIGMETRICS. Perform Eval. Rev.*, vol. 35, no. 1, pp. 109–120, 2007.
- [20] B. Rubinstein, B. Nelson, L. Huang, A. Joseph, S. hon Lau, R. Satish, N. Taft, , and J. Tygar, "Antidote: understanding and defending against poisoning of anomaly detectors," 2009.
- [21] Z. Wang, K. Hu, K. Xu, Y. Baolin, and X. Dong, "Structural analysis of network traffic matrix via relaxed principal component pursuit," *Computer Networks*, vol. 56, pp. 2049–2067, 2012.
- [22] C. Pascoal, M. De Oliveira, R. Valadas, P. Filzmoser, P. Salvador, and A. Pacheco, "Robust feature selection and robust pca for internet traffic anomaly detection," 2012, pp. 1755–1763.
- [23] D. Brauckhoff, K. Salamantian, and M. May, "Applying PCA for Traffic Anomaly Detection: Problems and Solutions," in *Proceedings of IEEE INFOCOM*, 2009, pp. 2886–2870.
- [24] T. Kourti and J. F. MacGregor, "Multivariate SPC methods for process and product monitoring," *Journal of Quality Technology*, vol. 28, no. 4, 1996.
- [25] A. Ferrer, "Latent Structures-Based Multivariate Statistical Process Control: A Paradigm Shift," *Quality Engineering*, vol. 26, pp. 72–91, Jan. 2014, publisher: Taylor & Francis _eprint: <https://doi.org/10.1080/08982112.2013.846093>.
- [26] J. M. González-Martínez, A. Ferrer, and J. A. Westerhuis, "Real-time synchronization of batch trajectories for on-line multivariate statistical process control using Dynamic Time Warping," *Chemometrics and Intelligent Laboratory Systems*, vol. 105, pp. 195–206, 2011.
- [27] S. Soufiane, R. Magán-Carrión, I. Medina-Bulo, and H. Bouden, "Preserving authentication and availability security services through multivariate statistical network monitoring," *Journal of Information Security and Applications*, vol. 58, p. 102785, 2021.
- [28] M. López-Martin, B. Carro, A. Sanchez-Esguevillas, and J. Lloret, "Conditional variational autoencoder for prediction and feature recovery applied to intrusion detection in iot," *Sensors*, vol. 17, p. 1967, 2017.
- [29] J. An and S. Cho, "Variational autoencoder based anomaly detection using reconstruction probability," *Technical Report*, pp. 1–18, 2015.
- [30] S. Zavrak and M. Iskefiyelo, "Anomaly-based intrusion detection from network flow features," *IEEE Access*, vol. 8, pp. 108 346–108 358, 2020.
- [31] M. Sölch, J. Bayer, M. Ludersdorfer, and P. van der Smagt, "Variational inference for on-line anomaly detection in high-dimensional time series," in *International Conference on Machine Learning anomaly detection workshop*. IEEE, 2016.
- [32] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006.
- [33] G. Maciá-Fernández, J. Camacho, R. Magán-Carrión, P. García-Teodoro, and R. Therón, "UGR'16: A new dataset for the evaluation of cyclostationarity-based network IDSs," *Computers & Security*, vol. 73, pp. 411–424, 2018. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0167404817302353>
- [34] M. Ring, S. Wunderlich, D. Scheuring, D. Landes, and A. Hotho, "A survey of network-based intrusion detection data sets," *Computers & Security*, vol. 86, pp. 147–167, 2019. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S016740481930118X>
- [35] M. Catillo, A. Pecchia, M. Rak, and U. Villano, "Demystifying the role of public intrusion datasets: A replication study of dos network traffic data," *Computers & Security*, vol. 108, p. 102341, 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0167404821001656>
- [36] S. García, M. Grill, J. Stiborek, and A. Zunino, "An empirical comparison of botnet detection methods," *Computers & Security*, vol. 45, pp. 100–123, 2014. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0167404814000923>

CHAPTER 7

Other works (published, submitted, in preparation)

This chapter includes three additional works developed during this Ph.D. in which the candidate had a relevant role in their elaboration. Since they are not part of the compendium of publications presented to obtain the Ph.D. degree, they will only be mentioned along with their relevant contributions. We also include here the 3 Minute Thesis (3MT) competition. This Ph.D. thesis won the first prize at the University of Granada in 2021.

7.1 The Devil Is in the Details: Whole Slide Image Acquisition and Processing for Artifact Detection, Color Variation, and Data Augmentation. A Review.

7.1.1 JCR Publication Details

Authors: Neel Kanwal*, Fernando Pérez-Bueno*, Arne Schmidt, Kjersti Engan, Rafael Molina (* Indicates equal contribution)

Title: The devil is in the details: Whole Slide Image acquisition and processing for artifact detection, color variation, and data augmentation. A review.

Reference: IEEE Access, 2022, 10, 58821-58844

Status: Published

DOI: 10.1109/ACCESS.2022.3176091

Quality indices:

- Impact Factor (JCR 2021): 3.476
 - Rank 105/276 (Q2) in Engineering, Electrical and Electronic
 - Rank: 79/164 (Q2) in Computer Science, Information Systems
- Journal Citation Indicator (JCR 2021): 0,93
 - Rank 104/344 (Q2) in Engineering, Electrical and Electronic
 - Rank: 75/246 (Q2) in Computer Science, Information Systems

7.1.2 Abstract

Whole Slide Images (WSI) are widely used in histopathology for research and the diagnosis of different types of cancer. The preparation and digitization of histological tissues leads to the introduction of artifacts and variations that need to be addressed before the tissues are analyzed. WSI preprocessing can significantly improve the performance of computational pathology systems and is often used to facilitate human or machine analysis. Color processing techniques are usually the main concern, while other areas are frequently ignored. In this paper, we present a detailed study of the state-of-the-art in three different areas of WSI preprocessing: Artifact detection, color variation, and the emerging field of pathology-specific data augmentation. We include a summary of evaluation techniques along with a discussion of possible limitations and future research directions for new methods.

7.1.3 Main Contributions

- We present a complete review on WSI preprocessing techniques connecting the WSI acquisition procedure to the causes for WSI variations and the crucial preprocessing steps.
- Three areas of interest for WSI preprocessing are included: artifact detection, color variation, and data augmentation.
- We discuss the current challenges and future directions for WSI preprocessing.

7.2 Deep Variational Bayesian Stain Separation of Histopathological Images Using Blind Color Deconvolution

7.2.1 JCR Publication Details

Authors: Shuowen Yang, Fernando Pérez-Bueno, Hanlin Qin, Rafael Molina, Aggelos K.Katsaggelos

Title: Deep Variational Bayesian Stain Separation of Histopathological Images Using Blind Color Deconvolution.

Status: Submitted

Quality indices:

- Impact Factor (JCR 2021): 11.041
 - Rank 12/144 (D1) in Computer Science, Artificial Intelligence
 - Rank 12/276 (D1) in Engineering, Electrical and Electronic
- Journal Citation Indicator (JCR 2021): 2.16
 - Rank 13/189 (D1) in Computer Science, Artificial Intelligence
 - Rank 17/344 (D1) in Engineering, Electrical and Electronic

7.2.2 Abstract

Histological images are often tainted with two or more stains to reveal their underlying structures and conditions. Blind Color Deconvolution (BCD) techniques separate colors (stains) and structural information (concentrations). This is a process useful for the processing, data augmentation, and classification of such images.

Classical BCD methods are often computationally expensive models in two different senses. Firstly, for a given image, the estimation of the corresponding colors and concentrations is time consuming and, secondly, the whole estimation process has to be carried out on each image independently (non-amortized). Deep neural networks learn a mapping from input to output (or probability distributions over the output) whose estimation may be time consuming but once it has been learned it can be used in a fast, amortized manner on unseen inputs.

Due to the lack of large databases of ground truth color and concentrations, deep learning methods have hardly been applied to BCD. In this work, we propose a deep variational Bayesian BCD neural network (BCDnet) for stain separation and concentration estimation. Under this framework, we tackle the lack of ground

truth by using Bayesian modeling and inference. A prior distribution on the stain colors, which does not require the knowledge of the true colors, and the use of maximum likelihood to estimate the concentrations are proposed. BCDnet is trained by maximizing the evidence lower bound of the observed images. Fidelity to the observed images (in a transformed space) and the Kullback-Leibler divergence between the estimated posterior distribution of the colors and the chosen prior are the terms that define the bound to be optimized. The model is trained, validated, and tested on two multicenter databases: Camelyon-17 and a stain separation benchmark with three different tissue types. The proposed approach performs well in comparison to classical non-amortized methods and paves the way for the use of deep learning techniques on BCD problems.

7.2.3 Main Contributions

- We introduce the first Bayesian approach to Blind Color Deconvolution using Deep Neural Networks.
- We tackle the lack of the true underlying ground truth color stains and concentrations for training by: a) the introduction of relevant priors on the color-vector matrix and concentrations and, b) the use of variational inference to approximate the posterior distribution of the color-vector matrix and concentrations given the observed optical density image.
- The proposed approach is the first amortized model presented for BCD of histological images.
- The proposed approach was evaluated on stain separation and showed competitive results with state-of-the-art BCD methods.

7.3 Robust blind color deconvolution and blood detection on H&E histological images using Bayesian K-Singular Value Decomposition

7.3.1 Publication Details

Authors: Fernando Pérez-Bueno, Kjersti Engan, Rafael Molina

Title: Robust blind color deconvolution and blood detection on H&E histological images using Bayesian K-Singular Value Decomposition.

Status: In preparation

7.3.2 Abstract

Color variation between histological images from different laboratories is a known issue that degrades the performance of Computer Aided Diagnosis (CAD) systems. These variations are caused by differences in the staining protocol (e.g., with Hematoxylin and Eosin (H&E)). Histology-specific models to solve color variation are designed taking into account the staining procedure. In particular, Blind Color Deconvolution (BCD) methods aim to identify the observed color the stains in the image and to separate the tissue structure from the color information. A common assumption is that images are stained with and only with the expected protocol (e.g., two stains for H&E). This assumption might not hold true in the presence of common artifacts such as blood in the image, were the blood cells usually obtain a third different color. The presence of blood usually hampers the ability of color standardization algorithms to correctly identify the stains in the image, producing unexpected outputs. In this work, we use the recently proposed Bayesian K-Singular Value Decomposition including a third ‘stain’ channel to detect blood in histological images and produce a robust blind color deconvolution. Our method was tested on synthetic and real images containing different amount of blood pixels.

7.3.3 Main Contributions

- We extend the BKSVD method for BCD to make it robust against artifacts in the image.
- We relate the fields of artifact detection and color variation by focusing on blood and how its presence affects BCD methods on H&E histological images.

- We propose the use of BCD for blood detection.

7.4 3 Minute Thesis (3MT) Competition

7.4.1 JCR Publication Details

Authors: Fernando Pérez-Bueno, Valery Naranjo, Rafael Molina.

Title: In Cancer Detection, the Devil Is in the Details.

Reference: 2021 Coimbra Group 3MT Competition.

Quality indices: First Prize (University of Granada, institutional finals)

7.4.2 Summary:

Developed by The University of Queensland (UQ), the 3MT competition consists of effectively explaining one's research in three minutes, in a language appropriate to a non-specialist audience. Competitors are allowed one PowerPoint slide, but no other resources. The event was streamed live on the UGR-media YouTube channel and the recording is available in [16].

The speech presented at the 3MT competition introduced the problem of color variation of histopathological images and how the models proposed in this Ph.D. thesis can improve CAD performance when using images from different hospitals. It was awarded with the first prize at the University of Granada and chosen to represent the university in the international competition held by the Coimbra Group.

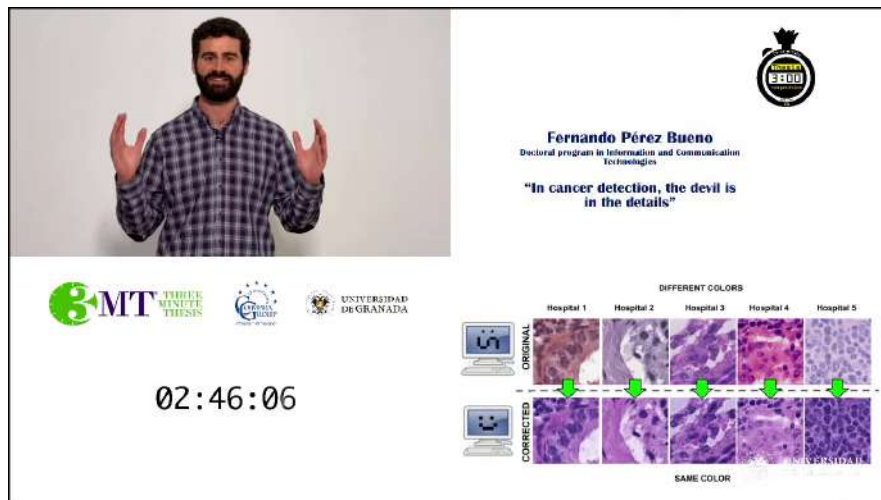


Figure 7.1: A frame from the speech presented in [16].

CHAPTER 8

Concluding Remarks

The main conclusion of this Ph.D. thesis is that Bayesian probabilistic modeling and inference can be used to improve histological images making them easier to classify and interpret using CAD systems. We explored Bayesian blind color deconvolution to separate the observed images into the latent elements that compose them (i.e. stain color and stain concentrations (Chapters 2-4)). Bayesian modeling and inference can also be applied in other areas such as pansharpening (Chapter 5) and network anomaly detection (Chapter 6). This can be specified through the following specific conclusions:

- The use of prior knowledge on the stain color and concentrations can lead to a better separation of the information on the observed image. Probabilistic models and Bayesian inference provide a consistent framework to introduce prior knowledge and to manage uncertainty in histopathological images.
- Improving the image might have a different sense depending on the task to perform. Higher fidelity to the original tissue is desired for visual analysis, while better classification features for CAD systems might be obtained from images where the noise and residual elements are removed.
- Sparsity is a desired feature for the latent separation of the stains in the image. We have explored sparsity on the stain concentration in three different approaches: by using the TV prior directly on the concentrations, by using the SG family of priors on the high-pass filtered concentrations to remark the edges, and with a two-tiered hierarchical prior (equivalent to a Laplacian prior) on the concentrations for each pixel to promote a separation where each pixel is assigned to only one stain.
- On the one hand, reference-based BCD is a robust approach that is not affected by artifacts on the image, but lacks the flexibility to adapt to color

distributions that are far from the reference. On the other hand, Dictionary Learning for BCD is able to estimate a color-vector matrix that better represents the differential staining, but is exposed to artifacts (e.g. blood or cauterized areas) on the images.

- Bayesian BCD is computationally expensive but outperforms non-probabilistic approaches for stain separation. However, due to the reduced number of stains in histological images, its application to massive WSI processing can be boosted with pixel sampling for the estimation of the BKSVD model parameters.
- BCD has a high potential for the improvement and interpretation of histological images. Stain separation can be used for color normalization and color augmentation, or directly for CAD purposes. Feeding CAD systems with the single-stained concentration images instead of the RGB observed or normalized image can improve the performance of the diagnostic. This approach mimics the analysis performed by pathologists, as they differentiate the stains on the image and not the color they present.
- Some of the lessons learned while working with histological images can be extended to other areas. Using probabilistic models to separate the information in their latent components can help to highlight information previously confused or disguised in the observed variables. Specifically:
 - The estimation of high-resolution multispectral images from a low-resolution multispectral images and a high-resolution panchromatic image can be improved by using the SG priors, separating the contribution of the panchromatic image to each channel of the HR MS image.
 - Probabilistic PCA provides a latent space that can be used for robust network anomaly detection. In addition, PPCA establishes a bridge between previous network anomaly detection models (i.e. MSNM) and recent generative approaches such as VAEs that can be used to better understand the latter.
- The works included in chapter 7 show that there is room for future research in histological image processing and probabilistic modeling. In the paper in section 7.1 reviews the state-of-the-art and remarks future directions and challenges of WSI processing. The recently submitted work in section 7.2 is, to the best of our knowledge, the first Bayesian approach to BCD using deep neural networks, and paves the way for new works using deep learning. The work in preparation in 7.3, also opens a new research line, presenting the

use of BCD for artifact detection in histological images. Finally, the 3MT award in [7.4](#), shows that the research developed in this thesis is of interest for a general audience.

Bibliography

- [1] M. G. Hanna, A. Parwani, and S. J. Sirintrapun, “Whole Slide Imaging: Technology and Applications,” *Advances in Anatomic Pathology*, vol. 27, no. 4, pp. 251–259, Jul. 2020.
- [2] World Health Organization, “Cancer fact sheets. Sep. 2021.” <https://www.who.int/news-room/fact-sheets/detail/cancer>, (accessed: January 2022).
- [3] S. Morales, K. Engan, and V. Naranjo, “Artificial intelligence in computational pathology – challenges and future directions,” *Digital Signal Processing*, p. 103196, 2021.
- [4] D. Tellez, G. Litjens, P. Bándi, W. Bulten, J.-M. Bokhorst, F. Ciompi, and J. van der Laak, “Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology,” *Medical Image Analysis*, vol. 58, p. 101544, 2019.
- [5] N. Kanwal, F. Pérez-Bueno, A. Schmidt, K. Engan, and R. Molina, “The devil is in the details: Whole slide image acquisition and processing for artifacts detection, color variation, and data augmentation. a review.” *IEEE Access*, pp. 58 821 – 58 844, 2022.
- [6] M. Salvi, U. R. Acharya, F. Molinari, and K. M. Meiburger, “The impact of pre- and post-image processing techniques on deep learning frameworks: A comprehensive review for digital pathology image analysis,” *Computers in Biology and Medicine*, vol. 128, p. 104129, 2021.
- [7] F. Pérez-Bueno, J. Serra, M. Vega, J. Mateos, R. Molina, and A. K. Katsaggelos, “Bayesian K-SVD for H&E Blind Color Deconvolution. Applications to Stain Normalization, Data Augmentation, and Cancer Classification.” *Computerized Medical Imaging and Graphics*, 2022.
- [8] N. Trahearn, D. Snead, I. Cree, and N. Rajpoot, “Multi-class stain separation using independent component analysis,” in *Medical Imaging 2015: Digital*

- Pathology*, vol. 9420. International Society for Optics and Photonics, 2015, p. 94200J.
- [9] A. E. Esteban, M. Lopez-Perez, A. Colomer, M. A. Sales, R. Molina, and V. Naranjo, “A new optical density granulometry-based descriptor for the classification of prostate histological images using shallow and deep Gaussian processes,” *Computer Methods and Programs in Biomedicine*, vol. 178, pp. 303–317, 2019.
- [10] R. Duggal, A. Gupta, R. Gupta, and P. Mallick, “Sd-layer: stain deconvolutional layer for cnns in medical microscopic imaging,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 435–443.
- [11] F. Pérez-Bueno, M. López-Pérez, M. Vega, J. Mateos, V. Naranjo, R. Molina, and A. K. Katsaggelos, “A tv-based image processing framework for blind color deconvolution and classification of histological images,” *Digital Signal Processing*, vol. 101, p. 102727, 2020.
- [12] F. Perez-Bueno, M. Vega, V. Naranjo, R. Molina, and A. K. Katsaggelos, “Super Gaussian Priors for Blind Color Deconvolution of Histological Images,” *Proceedings - International Conference on Image Processing, ICIP*, vol. 2020-October, pp. 3010–3014, 2020.
- [13] A. C. Ruifrok and D. A. Johnston, “Quantification of histochemical staining by color deconvolution,” *Analytical and quantitative cytology and histology*, vol. 23, pp. 291–299, 2001.
- [14] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006, pp. 454–455.
- [15] N. Hidalgo-Gavira, J. Mateos, M. Vega, R. Molina, and A. K. Katsaggelos, “Variational Bayesian blind color deconvolution of histopathological images,” *IEEE Transactions on Image Processing*, vol. 29, no. 1, pp. 2026–2036, 2020.
- [16] F. Pérez-Bueno. In cancer detection, the devil is in the details. Coimbra group 3 Minute Thesis (3MT) competition 2021. UGR Institutional finals. UGR Media. Youtube. Accessed: July 2022. [Online]. Available: <https://youtu.be/zsdDF53CcFI?t=2141>