Maximum a posteriori estimation of linear shape variation with application to vertebra and cartilage modeling

Crimi, Alessandro; Lilholm, Martin; Nielsen, Mads; Ghosh, Anarta; de Bruijne, Marleen; Dam, Erik B; Sporring, Jon

Published in:
IEEE Transactions on Medical Imaging

DOI:
10.1109/TMI.2011.2131150

Publication date:
2011

Document Version
Preprint (usually an early version)

Citation for published version (APA):
Maximum a Posteriori Estimation of Linear Shape Variation With Application to Vertebra and Cartilage Modeling

Alessandro Crimi*, Martin Lillholm, Mads Nielsen, Anarta Ghosh, Marleen de Bruijne, Erik B. Dam, and Jon Sporring

Abstract—The estimation of covariance matrices is a crucial step in several statistical tasks. Especially when using few samples of a high dimensional representation of shapes, the standard maximum likelihood estimation (ML) of the covariance matrix can be far from the truth, is often rank deficient, and may lead to unreliable results. In this paper, we discuss regularization by prior knowledge using maximum a posteriori (MAP) estimates. We compare ML to MAP using a number of priors and to Tikhonov regularization. We evaluate the covariance estimates on both synthetic and real data, and we analyze the estimates’ influence on a missing-data reconstruction task, where high resolution vertebra and cartilage models are reconstructed from incomplete and lower dimensional representations. Our results demonstrate that our methods outperform the traditional ML method and Tikhonov regularization.

Index Terms—Bayesian, cartilage, covariance estimation, incomplete data, maximum a posteriori (MAP), principal component analysis (PCA), reconstruction, regularization, shape model, Tikhonov regularization, vertebra.

I. INTRODUCTION

ESTIMATION of the covariance matrix is an initial and pivotal step of principal component analysis (PCA) [1], factor analysis (FA) [2], some versions of regression, and many other statistical tasks. If the sample size is small, and the number of considered variables is large, then the estimation of the covariance matrix can be poor. Regularization and specifically shrinkage have been proposed to improve estimates of covariance matrices.

Shrinkage methods aim at improving estimates by shrinking the estimation towards zero or in the general setting towards some specific value. The oldest shrinkage methods were proposed in the early 1960s [3], [4], and afterwards a series of important methods appeared [5]–[10]. These methods depend on the proper choice of a mixing parameter, and the mixing parameter is often selected to maximize the expected accuracy of the shrunken estimator by cross-validation [6], [7] or by using an analytical estimate of the shrinkage intensity [5], [8]–[10]. Regularized estimators have been shown to outperform the standard maximum likelihood (ML) estimator for small sample sizes [9]. It is most common to use a simple form of ridge regression [7], where nonzero values are added to the diagonal elements of the covariance matrix. Ridge regression was originally introduced by Tikhonov for solving an under-determined system of linear equations [4]. Friedman presented a restricted version of this regularization, linking the mixing parameter to the minimization of misclassification risk during discriminant analysis. Generally in regularization, nonzero values are added to the diagonal elements of a covariance matrix as $\Sigma_{\text{reg}} = \Sigma + \alpha^2 I$, where $\alpha^2$ is the discussed mixing parameter. It was shown that this mixing parameter is not likely to be known in advance, and finding its optimal value can be computationally demanding [6].

Probabilistic PCA considers the ML solution of a probabilistic latent variable model. This method finds the noncorrupted eigenvalues/eigenvectors, but it is not always able to regularize the covariance matrix optimally [11]. A similar method was proposed by Allassonniere et al. [12], where kernels in Hilbert space are used for estimating the covariance matrix of a generative model.

Stein estimators [5], [8]–[10] estimate the mixing parameter using just the sample covariance matrix. However, these methods rely on the estimation of a specific cost function.

An overview of methods optimized for data in nonlinear subspaces is given in [13], particularly we wish to highlight minimum noise fraction [14], minimum autocorrelation factor [15], and principal geodesic analysis (PGA) [16]. In these cases, PCA can be used as local, linear approximations.

In this paper, we investigate regularization by maximum a posteriori (MAP) estimates of covariance matrices in a PCA setting. A number of priors are discussed for this MAP method, and the usefulness of the priors is evaluated against known ground truth covariance matrices. Our underlying motivation is to improve the accuracy of medical diagnosis and prognosis of diseases, such as osteoporosis (OP) and osteoarthritis (OA). As an
example application, we demonstrate how low resolution vertebral shape outlines can be interpolated into high-resolution full contours using a shape model based on a MAP estimate of the point distribution model’s intrinsic covariance matrix.

A. Principle Component Analysis and Maximum Likelihood Covariance Estimation (ML)

In statistical shape analysis, a shape model is typically built using a set of training examples often represented as a sample matrix [17], [18]. A shape \( \mathbf{x} \in \mathbb{R}^d \) is a vector of \( \zeta \) labeled landmarks, \( \mathbf{p} \in \mathbb{R}^d \)

\[
\mathbf{x} = [\mathbf{p}_1^T \mathbf{p}_2^T \ldots \mathbf{p}_\zeta^T]^T
\]

and the collection of \( n \) shapes is collected in a sample matrix, \( \mathbf{X} \in \mathbb{R}^{d \times n} \)

\[
\mathbf{X} = [\mathbf{x}_1^{\mathbf{X}} \mathbf{x}_2^{\mathbf{X}} \ldots \mathbf{x}_n^{\mathbf{X}}].
\]

As an example, consider 30 shapes each consisting of six landmarks in 2-D X-ray images, then \( \mathbf{p}_i \in \mathbb{R}^2 \), \( \mathbf{x} \in \mathbb{R}^{12} \), and \( \mathbf{X} \in \mathbb{R}^{12 \times 30} \).

PCA [1] also known as Karhunen–Loeve transformation [19] and Hotelling transformation [20] is a widely used technique for applications such as dimensionality reduction, data compression, and data visualization [21]. PCA is an orthogonal projection of the data onto a lower dimensional linear subspace, such that the variance of the projected data is maximized sequentially along each new coordinate axis. The linear subspace is not always a precise model of the variability in the data: In the case of the shapes, the shapes may be intrinsically lie on a nonlinear submanifold. Furthermore, shape normalization using, e.g., Procrustes’ method [22], [23] for neutral position, size, and orientation will also typically project the shapes onto a nonlinear submanifold. As an example, size normalization, such as \( \mathbf{x} = \mathbf{x} / \| \mathbf{x} \|_2 \) will project the samples onto a \( q - 1 \) dimensional hypersphere, as illustrated in Fig. 1 for 2-D data. In this case, PCA may be considered an approximation of the projected variability in the tangent plane under suitable choice of mean. In part, this limitation can be addressed by exploiting the metric of the embedding manifold using techniques, such as, PGA [16]. Nevertheless, we only consider regularized linear models, since our focus is on cases, where the number of samples is small compared to the dimensionality of the embedding space, \( n \ll q \), and since we do not assume prior knowledge about any underlying shape manifolds.

PCA models the cloud of (normalized) shapes \( \{\mathbf{x}_i\}_{i=1}^{\ldots n} \) as a Gaussian distribution with mean \( \mu = E[\mathbf{x}] \) and covariance \( \Sigma = E[(\mathbf{x} - \mu)(\mathbf{x} - \mu)^T] \). In this paper, we assume that \( \Sigma \) has full rank except for shape normalization, e.g., when normalizing for translation, scaling and rotation, then \( \Sigma \) has rank \( q - l - 1 \). The covariance matrix \( \Sigma \) is symmetric and positive semi-definite, which is why we may write its eigenvalue decomposition as

\[
\Sigma = \mathbf{V} \Lambda \mathbf{V}^T
\]

where \( \mathbf{V} = [\mathbf{v}_1 \ldots \mathbf{v}_l] \) is the column matrix of eigenvectors, and \( \Lambda = \text{diag} ([\lambda_1 \ldots \lambda_l]) \) is the diagonal matrix of corresponding eigenvalues. Eigenvectors and values are assumed sorted according to \( |\lambda_i| \geq |\lambda_j| \) for \( i < j \). Transforming the shapes onto a new coordinate system \( \Xi = [\xi_1 \xi_2 \ldots \xi_n] \) as

\[
\Xi = \sqrt{\Lambda^{-1}} \mathbf{V}^T (\mathbf{X} - \mu_{1:n}^T)
\]

where \( \mu_{1:n} = [\mu_1 \ldots \mu_n] \) is a column vector of \( n \) ones, such that the outer product \( \mu_{1:n} \mu_{1:n}^T \) is a matrix of \( n \) \( \mu_i \), and where the inverse of zero eigenvalues are set to zero. We may consider the shape variations as ordered by variance in an uncorrelated coordinate system, and we may write an approximation of a shape \( \mathbf{x} \) by the \( l \) absolutely largest eigenvalues as

\[
\hat{\mathbf{x}} = \mu + \mathbf{V} \boldsymbol{\xi} = \mu + \sqrt{\Lambda_1 \ldots \Lambda_l, 0, \ldots, 0} \boldsymbol{\xi}
\]

where \( \boldsymbol{\xi} \) is column vector in \( \Xi \) corresponding to the column vector in \( \mathbf{X} \). The quadratic error of the approximation will be

\[
||\mathbf{x} - \hat{\mathbf{x}}||_2^2 = \sum_{i=l+1}^{n} \lambda_i.
\]

An often used estimate of the covariance matrix is the ML estimator [24]: The shapes \( \mathbf{x} \) are assumed to be independent and to have identical Gaussian density function, yielding the following joint distribution:

\[
p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n|\Sigma, \mu) = 
\frac{1}{(2\pi)^{n/2} |\Sigma|^{n/2}} \times \exp \left( -\frac{1}{2} \sum_{i=1}^{n} (\mathbf{x}_i - \mu)^T \Sigma^{-1} (\mathbf{x}_i - \mu) \right)
\]

where \( |\Sigma| \) is the determinant of \( \Sigma \). This distribution is often called the likelihood of \( \{\mathbf{x}_i\}_{i=1}^{\ldots n} \), and the point of ML for
varying \( \mu \) and \( \Sigma \) is derived for completeness in Appendix A and is found to be

\[
\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{7a}
\]

\[
\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})(x_i - \hat{\mu})^T. \tag{7b}
\]

The covariance estimate is slightly biased, but for large \( n \) the bias is negligible. The speed of the convergence of the estimator is inversely proportional to the number of shapes in a data set. However, faster convergence may be achieved, when prior knowledge is available.

### B. Tikhonov Regularization of Covariance Estimates (ML-T)

Regularization is one method for introducing prior knowledge and is often used as a tool for improving the estimate, when the number of samples is small compared to the dimensionality of the embedding space, \( n \ll q \). It is most common to use a simple form of ridge regression or Tikhonov regularization [4], [7], where nonzero values are added to the diagonal elements of the covariance matrix

\[
\Sigma_{\text{reg}} = \Sigma + \alpha^2 I \tag{8}
\]

where \( \alpha \in \mathbb{R} \). We refer to this method as ML-T. By diagonalizing \( \Sigma \), it is seen that ML-T only affects the eigenvalues of \( \Sigma_{\text{reg}} \) adding \( \alpha^2 \) to the eigenvalues of \( \Sigma \) regardless of their value. Adding a constant to zero eigenvalues is useful for regularizing an inversion but may not be for analyzing the subspace spanned by the data. It is our experience, that for \( n \ll q \), the low absolute eigenvalues require relatively more regularization than the high absolute eigenvalues. Varying \( \alpha \) as function of absolute eigenvalue is not easily done using ML-T. In this paper, we discuss Bayes estimation of the covariance matrix using the MAP method. We discuss two priors closely tied to the Wishart distribution, which allows for varying degree of regularization depending on absolute eigenvalue. Finally, we discuss a more general Gaussian prior, which allow for prior knowledge on off diagonal interactions.

### II. MAXIMUM A POSTERIORI COVARIANCE ESTIMATION

If we consider \( \Sigma \) a random variable in the space of symmetric, positive semi definite matrices, then Bayes theorem states that

\[
p(\Sigma | S) = \frac{p(S | \Sigma)p(\Sigma)}{p(S)} \tag{9}
\]

where \( p(\Sigma | S) \) is the posterior, \( p(S | \Sigma) \) the likelihood, \( p(\Sigma) \) the prior, and \( p(S) \) the evidence. We may postulate prior knowledge as prior densities and calculate the MAP estimate. As a limiting case, MAP converges to ML as the prior converges to the improper uniform distribution.

For all Bayes estimators of covariance matrices presented in this paper, we consider the same likelihood distribution \( p(S | \Sigma, n) \) of the sample covariance matrix, \( S \in \mathbb{R}^{n \times q} \)

\[
S = (X - \mu_s^T)(X - \mu_s^T)^T \tag{10}
\]

A Gaussian distribution with mean \( \mu \) and covariance \( \Sigma \) on \( \mathbf{x} \) induces a Wishart distribution on the likelihood of \( S \) [24, Ch. 7]

\[
p(S | \Sigma, n) = \frac{|S|^{(n-q-1)/2} \exp \left( -\frac{1}{2} \text{tr} \left( \Sigma^{-1} S \right) \right)}{2^{nq/2} \pi^{m/2} \Gamma_q \left( \frac{n}{2} \right)} \tag{11}
\]

where \( \Gamma_q \) is the multivariate Gamma function

\[
\Gamma_q \left( \frac{n}{2} \right) = \pi^{q(q-1)/4} \prod_{i=1}^{q} \Gamma \left( \frac{n}{2} - \frac{i-1}{2} \right). \tag{12}
\]

The zero mean distribution that best fits the samples \( S \) is found as the point of maximum log likelihood for varying \( \Sigma \) as

\[
\hat{\Sigma} = \frac{1}{n} S \tag{13}
\]

identical to (7b). A common prior for covariance matrices is the inverted Wishart to be discussed in the following.

### A. Inverted Wishart Prior (MAP-IW)

A prior giving a simple MAP estimate is the inverted Wishart distribution [24, Ch. 7]

\[
p(\Sigma | \Psi, m) = \frac{|\Psi|^{m/2} \exp \left( -\frac{1}{2} \text{tr} \left( (\Psi \Sigma)^{-1} \right) \right)}{2^{m/2} |\Sigma|^{(m+q+1)/2} \Gamma_q \left( \frac{m}{2} \right)} \tag{14}
\]

where \( \Psi \) and \( m \geq q \) are parameters of the distribution. The inverted Wishart density originates as the distribution of \( S^{-1} \), when \( S \) is Gaussian. Since the evidence is independent on \( \Sigma \), we find that the point of MAP is

\[
\hat{\Sigma} = \arg\max_{\Sigma} p(S | \Sigma, n)p(\Sigma | \Psi, m) \tag{15a}
\]

\[
= \arg\max_{\Sigma} c \left| \Psi \right|^{m/2} \exp \left( -\frac{1}{2} \text{tr} \left( (\Psi \Sigma)^{-1} \right) \right) \frac{|\Sigma|^{(n+m+q+1)/2}}{|\Sigma|^{(n+m+q+1)/2}} \tag{15b}
\]

using

\[
c = \frac{|S|^{(n-q-1)/2}}{2^{(n+m)q/2} \Gamma_q \left( \frac{n}{2} \right) \Gamma_q \left( \frac{m}{2} \right)} \tag{16}
\]

\[
\hat{\Sigma} = \frac{1}{n + m + q + 1} (S + \Psi). \tag{17}
\]

Regarding the structure of (17), we find that it can be similar to Tikhonov regularization (8), when \( \Psi = \alpha^2 I \), \( \alpha^2 \approx 1/n \), and \( n \gg m + q + 1 \). In any case, the parameter \( \Psi \) should be set by prior knowledge, e.g., as a measure of the uncertainty in landmark detection by a human operator, or using a geometric distance measure such as (32) also investigated in [25].

We wish to use the inverted Wishart to model independent landmark noise, therefore the structure of the true covariance matrix is assumed to be block diagonal (one block per landmark), and hence \( \Psi \) must also be block diagonal. Further, we observe that for typical ML estimates of \( \Sigma \) for \( n \ll q \), the close to zero eigenvalues are particularly poorly estimated. We speculate that this could be caused by overfitting of the first eigenvectors to the data leaving the directions of lesser variance to be underestimated: The first eigenvalue and corresponding eigenvector is
typically estimated through the direction of largest variance in a least squares sense. With only few available samples, this direction of largest variance and the corresponding eigenvalue will, on average, be overestimated. This is straightforward to validate empirically through repeated ML estimation on small samples from a known covariance structure; if the samples afford a direction with variance larger than the known covariance this will always be chosen by the ML estimate. This will, on average, bias the average size of the estimate of the initial eigenvalue. Thus and in contrast to ML-T, we propose to boost small eigenvalues more than large eigenvalues of $\Sigma_i$ by setting $\Psi$ to be block-wise proportional to the per landmark inverse ML estimate of its covariance matrix. As an example, for 2-D landmarks distributed as $\Sigma_i$, then $\Psi = \text{diag}(\Sigma_i^{-1}, \ldots, \Sigma_i^{-1})$

$$\hat{\Sigma} = \frac{1}{n + m + q + 1} \left( S + \alpha^2 \text{diag} \left( S_1^{-1}, \ldots, S_k^{-1} \right) \right) \quad (18)$$

where $\alpha \in \mathbb{R}$ is a free mixing parameter. We call this the MAP-IW. This is not a proper prior, since it depends on the measurements. Furthermore, it is noted that when $\alpha \neq 0$, both ML-T and MAP-IW enforces full rank. We defer discussion of its usefulness to the experimental section.

### B. Uncommitted Inverted Wishart Prior (MAP-UIW)

The MAP-IW (18) uses an improper prior that depends on an estimate of the covariance of each landmark, which in total may be considered an approximation of $\Sigma$. We propose to include $\Psi = \alpha^2 \Sigma^{-1}$ directly in the estimation, in which case the maximization results in a system of quadratic equations

$$\hat{\Sigma} = \frac{1}{n + 2m + q + 1} \left( S\Sigma + 2\alpha^2 I \right) \quad (19)$$

as derived in Appendix C. This is denoted the MAP-UIW method. As explained above, the MAP-UIW enforces full rank of the estimate. We solve this iteratively using the fix point iteration

$$\hat{\Sigma}_{t+1} = \hat{\Sigma}_t - \delta \left( \hat{\Sigma}_t - \frac{1}{n + 2m + q + 1} \left( S\hat{\Sigma}_t + 2\alpha^2 I \right) \right) \quad (20)$$

where $\delta$ is a sufficiently small constant to avoid divergence. Symmetry of the numerical solution may be guaranteed by taking the average with the transpose, or by only performing calculations for the lower left triangular matrix including the diagonal. We have experimented with both solutions and found no significant differences. For simplicity, we prefer the average. Further, we have experimented with a number of different initial points and the method appears to be robust, however, in this work we use ML-T as the initial point $\hat{\Sigma}_0$.

An alternative solution to (19) is obtained by a different factorization of the (48)

$$\hat{\Sigma} = \frac{1}{n + 2m + q + 1} \left( \sqrt{S}\Sigma \sqrt{S} + 2\alpha^2 I \right) \quad (21)$$

this solution theoretically solves the symmetrization issue and employed as a fix point iteration converges with fewer steps than (20). However, there are some numerical drawbacks. First, the need of a SVD decomposition during the square root computation slows down the algorithm. Second, since we perform iterations using a fixed step size, and the space of positive definite matrices (pd) is not a linear vector space, even this elegant solution can in practice step outsize pd-space. Therefore, in our experiments we prefer to use (20).

### C. Gaussian Prior (MAP-G)

An alternative prior is the Gaussian distribution in covariance space

$$p(\Sigma|B, s) = \frac{1}{(2\pi s^2)^{(q+1)/2}} \exp \left( -\frac{||\Sigma - B||_2^2}{2s^2} \right) \quad (22)$$

where $B \in \mathbb{R}^{q \times q}$ and $s \in \mathbb{R}$ represent the mean and variance of the covariance matrix distribution. The variance $s$ controls the trade-off between the prior and the ML estimate. A large value of $s$ would make the prior uniform and would yield the ML estimate. On the other hand, a very small $s$ would yield a delta spike around the mean value and thus bias the estimate heavily towards the mean of the Gaussian. Ideally, the Gaussian prior should be imposed on the space of symmetric, positive definite matrices, e.g., as approximated by the exponential of the norm of the differences of logarithms of matrices [26], but this makes the equation prohibitively complicated, hence we consider (22) a Euclidean approximation. The MAP of (22) is derived in Appendix D and is found to be a system of third degree polynomials

$$\hat{\Sigma} = \frac{1}{n} \left( S - s^2 \hat{\Sigma}(\hat{\Sigma} - B)^T \hat{\Sigma} \right) \quad (23)$$

We seek its solution by the following iterative scheme:

$$\hat{\Sigma}_{t+1} = \hat{\Sigma}_t + \delta \left( \hat{\Sigma}_t - \frac{1}{n} \left( S - s^2 \hat{\Sigma}_t (\hat{\Sigma}_t - B)^T \hat{\Sigma}_t \right) \right) \quad (24)$$

which we have found to converge for $t \rightarrow \infty$, when using iterative line search in $\delta$, enforcing symmetry by averaging, and using the ML-T as initial point $\hat{\Sigma}_0$. Matrix $B$ is a free positive semi-definite matrix, which is the embodiment of prior knowledge. Unfortunately, covariance matrices depend on the coordinate system, however, the matrix $B$ may be specified in a coordinate system of choice and rotated into the specific coordinate system by a suitable rotation matrix $R$ as $B = RB^T R^T$.

In this section we have considered three different priors to be imposed on our covariance matrix estimates. It is important to bear in mind that the ML-T assumes independence between all elements in the shape, and MAP-IW assume independence between landmarks. Hence both are priors of spatial noise. The MAP-UIW does not have assumption of independence, and MAP-G may be used to directly steer the estimate towards preferred shape variations. Hence both are general priors of shape variations. The classical view of the Inverse Wishart prior has a similar number of parameters as our Gaussian prior. There is, however, a clear difference in how these parameters are used: For the Inverse Wishart prior given $\Psi$, the covariance matrix of maximum probability is the solution to $\text{tr}(\Psi \Sigma^{-1}) = 0$, i.e., the $j$th column in $\Sigma^{-1}$ must be perpendicular to the $j$th row in $\Psi$. Hence, if we wish to align the Inverse Wishart prior to favor a particular covariance matrix $\Sigma$, then we should select a $\Psi$. 

from the \((q-1)(q-2) \ldots 1 = (q-1)!\) dimensional manifold fulfilling the above equation, none of which intuitively relate to \(\Sigma\). Conversely for the Gaussian prior, maximum probability is where \(|\Sigma - B|_F^2 = 0\), and in order to favor a particular covariance matrix \(\Sigma\), we simply and intuitively should choose \(B = \Sigma\). Furthermore, the additive structure of the Inverse Wishart prior estimators is directly comparable to Tikhonov regularization. In light of the difficulties of large dimensional manifolds described above, we argue that the constrained as well as the uncommitted Wishart priors are the practical extensions of Tikhonov regularization.

III. EXPERIMENTAL SETUP

We evaluate the performance of the previously described methods for estimating covariance matrices: ML, ML regularized by Tikhonov regularization (ML-T), MAP with an inverted Wishart prior (MAP-IW), MAP with an uncommitted inverted Wishart prior (MAP-UW), and finally MAP with a Gaussian prior (MAP-G). The evaluation is focused on two main tasks. Firstly, we compare the distance from a known ground-truth covariance matrix to the estimated covariance matrices for synthetic data. Secondly, we evaluate a reconstruction task from low to high-resolution shapes. The details of our validation protocol are described in the following subsections. Section III-A describes the metric used for comparing matrices, Section III-B discusses the reconstruction of high dimensional shape representations, while Section III-C presents the parameter optimization scheme used in the two validation tasks. Section III-D describes the mean value \(B\) of the Gaussian prior, which we use for our experiments. The results of this validation protocol and the experimental setups are described in Sections IV-A and IV-B, respectively, for synthetic data and for vertebrae and cartilage shape reconstruction.

A. Matrix Comparison

The first part of the validation protocol is a comparison between the ground-truth covariance matrix and different covariance matrix estimates. A matrix norm is required for this purpose. A suitable norm is the Frobenius or two-norm

\[
d(S_1, S_2) = \|S_1 - S_2\|_F
\]  

(25)

where \(S_1\) can be the ground truth matrix and \(S_2\) a matrix estimated with one of the methods described in this paper. Another useful norm is the log-Euclidean distance between two matrices [26], which approximates the distance between the matrices along the geodesics in the space of symmetric, positive definite matrices, \(d(S_1, S_2) = \|[\log(S_1) - \log(S_2)]\|_F\), where \(\log(S) = V \log(\Lambda) V^T\), when \(S = V \Lambda V^T\). With our data, we found no qualitative differences between these norms, and in Section IV-A we present our results using the Frobenius norm.

B. Increasing Shape Resolution

The direct matrix comparison experiments are not well suited for real data, where ground truth covariance matrices are not available. Therefore, we focus our attention on the example application of increasing shape resolution.

Accurate manual annotation of, e.g., full vertebrae boundaries is tedious and time consuming. Likewise for studies of cartilage shape, more detailed models may lead to more accurate results. These methodologies are described more in detail in Section IV-B. Increasing the shape resolution from a low dimensional manual annotation can be a viable alternative.

Letting \(y\) be an incomplete or lower dimensional shape vector of dimensionality \(k < q\), we would like to find the corresponding higher resolution shape \(x \in \mathbb{R}^q\) such that

\[
y = Lx
\]  

(26)

where \(L : \mathbb{R}^q \rightarrow \mathbb{R}^k\) is a linear mapping. For our experiments the matrix \(L\) is a projection matrix. The lower dimensional shape \(y\) is aligned to \(L\mu\) by the standard Procrustes method. Since the system is under-determined, the solution is not uniquely defined. Following Blanz and Vetter [27], we estimate \(x\) by minimizing the functional

\[
E(x) = \|Lx - y\|_2^2.
\]

(27)

Since \(x\) belongs to the shape model (5), the aforementioned functional transforms to

\[
E(\xi) = \|L\mu + \xi - y\|_2^2
\]

(28)

where \(\xi\) is the parameter vector of the shape model (5), and \(Q = LV_k \Lambda_v\), where \(V_k\) is the eigenvectors and \(\Lambda_v = \text{diag}(\lambda_i)\) the eigenvalues corresponding to the \(t\) principal eigenmodes of the covariance matrix. This covariance matrix is estimated using one of the methods described in the previous sections. \(Q\) is a \(k \times q\) matrix of rank at most \(k\). The corresponding reconstruction is [27]

\[
x = \mu + V_k \Lambda_v H W^+ U^T y
\]

(29)

where \(U, H, W\) are defined by the singular value decomposition of \(Q, Q = U W H^T\), and the diagonal matrix is inverted as

\[
W^+ = \begin{cases} \frac{1}{w_i^2} & \text{if } w_i > 0 \\ 0 & \text{else} \end{cases}
\]

(30)

and \(w_i\) being the \(i\)th singular value of \(W\). By construction, (29) is invariant to scaling of the underlying regularized covariance matrix with nonzero real constants.

C. Parameter Optimization

Some of the covariance matrix estimators described in Section II require optimization of parameters such as \(\alpha, m,\) and \(s\), and we use cross-validation [28], [29] as follows. For given data, we 20 times randomly select a nonoverlapping train and test set of a specified size. For each pair, we examine a range of parameter values, and for each parameter value, we estimate the generalization error using leave-one-out on the training set. The optimal parameter value is chosen to minimize the generalization error on the training set. The error on the corresponding test set is then calculated using the optimal parameter value. Finally, we report the average performance over all test sets. Fig. 2 shows the 20 different training error
curves as a function of the mixing parameter $\alpha$ for MAP/UIW as well as their detected minima. The same approach is used for optimizing the parameters of the other methods. For the reconstruction evaluation, we note that the parameters of some priors are multiplicative constants and do not require optimization as described previously.

### D. Mean Value of the Gaussian Prior

For the Gaussian prior described in Section II, a crucial aspect is the design of the mean matrix $\mathbf{B}$ of the prior distribution of $\mathbf{X}$ in (22) and Section II-C. The design of this matrix must reflect some prior knowledge of the data it represents. For example, neighboring points often covary for typical shapes. Therefore, we find it useful to define $\mathbf{B}$ as a function of the distances between points similarly to [25]

$$A_{ij} = \text{dist}(p_i, p_j)$$

(31)

where $p_i$ is the $i$th point of the mean of the aligned training examples. Subsequently, we reweight the distances with an exponential decrease to obtain

$$\mathbf{B} = \exp\left(-\frac{1}{\beta} \mathbf{A} \otimes \mathbf{I}_l\right)$$

(32)

where $\exp()$ is taken element-wise, $\otimes$ is the Kronecker product, and $\mathbf{I}_l$ is the $l \times l$ identity matrix, and $\beta$ is a scale parameter. In this way, the resulting matrix $\mathbf{B}$ is of size $l \times q$, and $\mathbf{A}$ is of size $q/l \times q/l$. This definition of $\mathbf{B}$ focuses the regularized estimates to a region of covariance space, where shape point covariances are related by proximity.

In [25] geodesic distance were used, but in this work we prefer Euclidean. In general, whether one chooses the Euclidean or the geodesic distance is a modeling choice. If, for example, the shape change is due to friction, then it is fair to assume that shape changes correlate on the surface with nearby points. On the other hand, shape changes for soft tissue may show correlations between surface points that are better modeled by their direct Euclidean distance than their geodesic surface distance.

Regardless, considering only points on the surface of an object and for small values of $\beta$ in comparison to the distance between sample points, the Euclidean and the geodesic distances yield similar values of $\mathbf{B}$ in (32). Further, when objects are flat sheet as is the case for the tibial knee cartilage sheets described in Section IV-B, the Euclidean distance approximates the geodesic distance well. Finally, the Euclidean distance is the easiest to calculate. We choose to model only correlation between nearby points and therefore small values of $\beta$ and Euclidean distance.

In this section, we have given an example of the mean value matrix $\mathbf{B}$ based on the idea that proximity implies covariance. Many other, and perhaps better, choices are possible, but exhaustive investigation is beyond the scope of this paper.

### IV. Experiments

We now describe the results obtained using the validation protocol and setup described in Section III. First, we report the matrix comparisons between ground truth and covariance estimates for the synthetic data. Secondly, we discuss the shape reconstruction for vertebrae and cartilage shapes.

### A. Synthetic Shapes, Setup, and Data

In this section, we describe the data setup of the experiments with synthetic data and the results of performing a matrix comparison as described in Section III-A. The synthetic data set is designed to explore two types of variability: linear shape change and independent landmark random variations. First, we design a basis shape resembling a vertebra. Second, the linear shape variation is created by covarying two points along the vertical direction by a normally distributed displacement, which is a typical shape variation seen in real data. Third, normally distributed noise is added to each landmark to simulate annotation noise. We impose three different covariance matrices

$$C_1 = \gamma \begin{bmatrix} \kappa_1 & -\kappa_2 \\ -\kappa_2 & \kappa_2 \end{bmatrix}$$

(33a)

$$C_2 = \gamma \begin{bmatrix} \kappa_1 & \kappa_2 \\ \kappa_2 & \kappa_2 \end{bmatrix}$$

(33b)

$$C_3 = \gamma \begin{bmatrix} \kappa_3 & 0 \\ 0 & \kappa_2 \end{bmatrix}$$

(33c)

Fig. 3(a) illustrates the linear shape variation and Fig. 3(b) examples of noise. For the noise, we have used $\gamma = 2$, $\kappa_1 = 0.3$, and $\kappa_2 = 0.1$, and $C_1$ for the top right and bottom left corners, $C_2$ for the remaining corners, and $C_3$ for the mid sections. Examples of resulting shapes are shown in Section IV. We estimate the ground truth covariance matrix by using ML on a very large and independent data set of 10,000 samples. Furthermore, we generate 5000 samples, from which the 25 crossfolds are generated as described in Section III-C.

The experiments on the synthetic data previously described lead to the results depicted in Figs. 4 and 5. Here, (25) is used to compare the ground truth and the estimated matrices on a small sample size using the different methods (ML, ML-T, MAP-G, MAP-IW, and MAP/UIW) for two noise levels. For each method, the variation for cross-validation is illustrated using standard error of the mean (SEM) error bars. Furthermore, significance of differences between the best method
Fig. 3. (a) Mean shape of the synthetic data and an example of linear shape variations obtained from a Gaussian distribution, where the two central points co-vary vertically. (b) Example of noise variation, where the ellipses denote the standard deviation. Samples of the resulting shapes are shown in Figs. 4 and 5.

and Tikhonov regularization is illustrated by stars, where ** denotes a t-test p-value below 0.05, *** below 0.01, and **** below 0.001. We observe that all the suggested methods outperform ML, and further that the MAP-IW and MAP-UIW priors are suitable choices when the shape variation is dominated by noise (large $\gamma$ from (33), illustrated in Fig. 5), while the MAP-G prior is a better choice when there is a dominant shape variation (illustrated in Fig. 4). In addition, Figs. 4 and 5 show that with more available data less regularization is required; and the differences between the methods are less pronounced. It is also important to bear in mind that except for MAP-G, all the methods share the same eigenvectors and differ only in the eigenvalues. This is likely reflected by the clustering of graphs not including MAP-G.

B. Increasing Shape Resolution, Setup and Real Data

In this section, we explain the experimental setup and results for increasing shape resolution. These experiments are performed on two different data sets: vertebrae and cartilage shapes. For the aim of detecting osteoporotic fragility fractures, the use of high-resolution (full boundary) vertebral shapes may lead to more reliable results than an analysis with low-resolution shapes [30], [31]. For studies on osteoarthritis, more detailed 3D shape models of cartilage may also lead to more accurate results.

Using the shape reconstruction model (29), a high-resolution shape was reconstructed from a low-dimensional shape. The reconstruction error for a single shape is computed for the $\zeta$ points of the boundary as the root mean square error (RMSE)

$$E_{\text{full}} = \sqrt{\frac{1}{\zeta} \sum_{i=1}^{\zeta} \| \bar{p}_{i,\text{reconstr}} - \bar{p}_{i,\text{series}} \|_2^2}.$$  (34)
The performances of ML, ML-T, MAP-IW, and MAP-UIW methods are compared using mean $E_{\text{full}}$ over all the test shapes for different number of principal eigenmodes.

1) Vertebra Shapes: For general shape analysis, a detailed outline of the shapes are typically used [as illustrated in Fig. 6(a)]. However, for clinical studies on osteoporosis, height ratios are often used for fracture quantification [32], [33] and only the six vertebra contour points needed to compute these height ratios are routinely marked by expert radiologists. These six landmark points are shown as large stars in Fig. 6(b). The aim is thus to interpolate the full contour [shown as full outline in Fig. 6(b)] from these six points. During the experiments, the shape model is built from the full outline using 52 landmarks per shape. Hence, in our experiments $q = 104$ and $k = 12$.

We study the vertebra shape variation using images of 75 healthy elderly women, who maintain skeletal integrity over a seven years observation period. The subjects had lateral X-ray of the lumbar spine taken twice, once in 1992–93 (baseline) and again in 2000–01 (follow-up). The used images were digitized using a Vidar DosimetryPro Advantage scanner at 45 μm (570 dpi) providing an image resolution of 9651 × 4008 pixels with 12-bit gray scale. The four lumbar vertebrae L1-L4 were annotated using an annotation tool developed in Matlab. Using both the baseline and follow-up images, we obtain 600 (75 patients ×2 visits × 4 vertebrae) vertebra shape annotations. To avoid biased results, we only use shapes from the follow-up set. Furthermore, we only use 1 of 4 possible vertebrae per patient selected at random from L1 to L4. Therefore, the total data set of 600 shapes is reduced to 75 shapes. From this we either train on 10 or 20 shapes, and always test on 35 shapes. Fig. 7 shows the mean reconstruction errors for vertebrae shapes obtained using (34) for two different training sample sizes as a function of

---

Fig. 5. Frobenius norm of the difference between the ground truth covariance matrix and the estimation with ML, ML-T, MAP-IW, MAP-UIW, and MAP-G priors. All the methods are compared for different number of samples, all the parameters are determined by cross-validation as described in Section III-C. Subfigure (a) shows the data obtained using (33) with $\gamma = 2$, (b) shows the difference as described in Section IV-A, while (c) shows a zoom of the graph in (b). For the estimation procedure we used $\theta = 0.25$, and for the cross-validation optimization we investigated the following intervals: $m \in \{1, \ldots, 5\}$, $s \in \{0.01, 0.1\}$, $\alpha_{\text{ML-T}} \in \{0, 0.1\}$ and $\alpha_{\text{IW}}, \alpha_{\text{UIW}} \in \{0, 0.5\}$. The stars indicate statistical significance of the differences between MAP-UIW and ML-T (see Fig. 4 for details).

Fig. 6. (a) An image of a vertebra with the shape annotation. (b) An original (continuous line) and reconstructed (dashed line) shape annotation. The shape’s 52 points are reconstructed using only the six points depicted as the big stars.
2) Cartilage Shapes: The cartilage data set was composed of 320 knee MRI scans from 80 subjects including both left and right knees and baseline and follow-up scans from a longitudinal 21-month study. In particular, we focus on the tibial cartilage sheets, which are approximately flat. Similarly to the vertebrae, the data set was reduced to 80 samples to ensure only one shape per patient. The MRI acquisition was done on an Esaote C-Scan low field 0.18T clinical scanner using 3D, T1-weighted sequences (flip angle 40°, TR 50 ms, TE 16 ms). The scans were sagittal with slice resolution of 0.7 mm × 0.7 mm and slice thickness of 0.8 mm. The dimensions of the scans were 256 × 256 pixels with around 110 slices. The population includes healthy and diseased knees with varying degree of osteoarthritis from both men and women at ages from 21 to 78.

We represent the knee cartilage sheets using the m-rep model [35]. Here, a shape is defined by a spatially regular lattice of medial atoms as depicted in Fig. 8(a). For each knee, we have a three-dimensional m-rep model of the medial tibial cartilage compartment resulting from a fully automatic segmentation [36]. In Fig. 8(a), a cartilage shape model is illustrated for a cropped knee MRI.

For m-reps, each atom is representation by a tuple of position, orientation, and radius parameters. A more detailed explanation of this model is given in [37]. For this paper, we only consider the position parameters to allow an analysis similar to the point distribution model used for the vertebra shapes.

We evaluate the performance of our method for interpolating medial atoms in order to produce a higher resolution medial model of the cartilage sheets. We perform this evaluation by removing atoms from the models and by measuring how well the interpolation allows for reconstruction of the original model. We train on 10 or 20 shapes, and always test on 35 shapes from the remaining shapes. We produce a low-resolution lattice by removing atoms from the cartilage model, and calculate the mean reconstruction error between the original and the reconstructed 3D shapes using (34). To represent a medial model for a cartilage sheet, ζ = 32 points are used, hence the dimensionality q = 96. During the experiments 24 points are removed leaving four points, such that the reduced dimensions is k = 12. A typical result of the reconstruction is shown in Fig. 8. The mean reconstruction error using cartilage shapes is obtained using (34) and depicted in Fig. 9. The cartilage shapes show an improvement using the Gaussian prior. Like for the vertebra, the MAP-G method is more appealing because the shapes are dominated more by global shape variations rather than noise. The observation that MAP-G achieves a given reconstruction error using fewer eigenmodes compared to the other methods indicate that MAP-G estimates a more compact representation of the data.

V. CONCLUSION

Efficient estimation of the covariance matrix is of high importance in statistical shape analysis. Often, the number of available training examples is limited and the estimation of a covariance matrix is a challenging task. We have studied a Bayesian approach to the problem as an alternative to the well-known methods of ML and Tikhonov regularization (ML-T), and we have investigated three different priors: the Inverted Wishart (MAP-IW), the Uncommitted Inverted Wishart (MAP/UIW),
and the Gaussian prior (MAP-G). Comparing with respect to the parameters, the ML method has no parameters, the ML-T has one parameter, the MAP-UIW prior has two, while the remaining two priors have \((q + 1) \cdot q/2 + 1\) parameters in their unconstrained form. Both the MAP-IW prior and the MAP-G prior require a full matrix of parameters, but we have suggested constrained forms, that can be deduced automatically. The MAP-IW method can reflect the noise variation of independent landmark building the \(\Psi\) matrix as a block diagonal matrix as described in Section II-A. Contrarily, the MAP-G mean matrix \(B\) allows for inter-landmark interactions, like the suggested geometrical relationships in Section III-D. The Wishart priors also depend on a nuisance parameter \(m\), but we remind that for the reconstruction task the results do not seem to be very sensitive to its value. The presented methods perform as well as or better than the ML method especially in the case of small number of training examples. The choice of prior is naturally of great importance for the result. The MAP-IW and MAP-UIW priors assume statistically independent points, and therefore perform better for variations caused by uncorrelated noise. The MAP-G prior is not limited to zero off-block-diagonal elements.
and may be used to steer the estimate towards preferred shape variations. Therefore, the choice of prior is related to various factors: the number of samples available, the complexity of the data, the noise variance in the data and the intended application context in which estimated covariance matrices will be used. In general, the prior and specifically for the MAP-G the $B$ matrix has to be chosen using prior knowledge optimizing this interplay between prior, data, and application. The focus of this paper has been to introduce novel prior frameworks for covariance matrix estimation and to give simple examples of how they can be used to introduce actual prior knowledge in both artificial and real data sets. But even with the slightly naïve $B$ matrix suggested for the MAP-G prior, the improvements over standard approaches are significant for all but one of the examples given. Finally, the paper presented one case with a strong or near-optimal link between prior and data is present. This was the case for the very noisy version of the artificial data, where as the expected the Uncommitted Inverted Wishart prior excelled. In summary, with a sufficient number of samples the standard ML estimation may be the right choice, while for low sample sizes regularisation is required. In this case, we argue that our methods can be valid substitutes to Tikhonov regularisation.

**APPENDIX A**

**DERIVATION OF MAXIMUM LIKELIHOOD ESTIMATOR**

For completeness the following gives the derivation of the log-likelihood estimates of the mean vector and covariance matrix using matrix differential calculus [38]. This formulation will be used as a reference point for deriving expressions for the MAP estimates.

Using (6), for estimating $\Sigma$ and $\mu$ from a set of samples, it is possible to seek the maximum point of $p(x_1, x_2, \ldots, x_n|\Sigma, \mu)$ as

$$\frac{\partial \log p(x_1, x_2, \ldots, x_n|\Sigma, \mu)}{\partial \Sigma} = 0 \quad (35a)$$

$$\frac{\partial \log p(x_1, x_2, \ldots, x_n|\Sigma, \mu)}{\partial \mu} = 0 \quad (35b)$$

For practical reasons it is possible to rewrite the sum under the exponential function of (6) as

$$\sum_{i=1}^{n} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) = tr \left( \Sigma^{-1} XX^T \right) \quad (36)$$

and to rewrite the logarithm of (6) as

$$L(x_1, \ldots, x_n|\mu, \Sigma) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\Sigma) - \frac{1}{2} tr \left( \Sigma^{-1} XX^T \right) \quad (37)$$

The differential of $L$, considering only $\Sigma$ and $\mu$ as variables, is found to be

$$dL = -tr \left( \frac{n}{2} \Sigma^{-1} d\Sigma + \frac{1}{2} (d\Sigma^{-1}) XX^T + \frac{1}{2} \Sigma^{-1} d(XX^T) \right) \quad (38)$$

For minimizing the partial derivatives of $dL$ for $\Sigma$, it is possible to isolate the first two terms of (38) considering $d\Sigma^{-1} = -\Sigma^{-1} (d\Sigma) \Sigma^{-1}$

$$0 = -tr \left( \frac{n}{2} \Sigma^{-1} d\Sigma - \frac{1}{2} \Sigma^{-1} (d\Sigma) \Sigma^{-1} XX^T \right) \quad (39a)$$

$$0 = -tr \left( \Sigma^{-1} (d\Sigma)(nI - \Sigma^{-1} \Sigma^{-1} XX^T) \right) \quad (39b)$$

where a nontrivial solution is $0 = nI - \Sigma^{-1} \Sigma^{-1} XX^T$ that is

$$\Sigma = \frac{1}{n} XX^T \quad (40)$$

equivalent to (7b). Alternatively, starting from the Wishart likelihood

$$p(S|\Sigma, n) = \frac{|S|^{(n-q-1)/2} \exp \left( -\frac{1}{2} tr \left( \Sigma^{-1} S \right) \right)}{2^{\frac{na}{2}} (\Sigma)^{n/2} \Gamma_q \left( \frac{n}{2} \right)} \quad (41)$$

the point of ML for varying $\Sigma$ is found to be

$$0 = d \log p(S|\Sigma, n) = -\frac{1}{2} dtr \left( \Sigma^{-1} S \right) - \frac{n}{2}/log|\Sigma| \quad (42a)$$

$$= \frac{1}{2} tr \left( \Sigma^{-1} (d\Sigma) \Sigma^{-1} S - n \Sigma^{-1} d\Sigma \right) \quad (42b)$$

which implies that

$$0 = \Sigma^{-1} S - nI \quad (43)$$

or therefore $\Sigma = S/n$, also equivalent to (7b).

**APPENDIX B**

**DERIVATION OF MAXIMUM A POSTERIORI INVERTED WISHART ESTIMATOR (MAP-IW)**

In the following, we will derive (17). Starting from (15) and only considering varying $\Sigma$, we calculate the differential log as

$$d \log p(\Sigma|\Psi, m) = \frac{-(m+q+1)}{2} d \log |\Sigma| - \frac{1}{2} tr \left( \Psi \Sigma^{-1} \right) \quad (44a)$$

$$= \frac{-(m+q+1)}{2} d \log |\Sigma| - \frac{1}{2} tr \left( \Psi \Sigma^{-1} \right) \quad (44b)$$

$$= \frac{1}{2} tr \left( \Psi \Sigma^{-1} \right) - \frac{(m+q+1)}{2} d \log |\Sigma| \quad (44c)$$

The MAP using i.i.d. Gaussian likelihood is found to be

$$0 = d \log p(x_1, \ldots, x_n|\mu, \Sigma) + d \log p(\Sigma|\Psi, m) \quad (45a)$$

$$= \frac{1}{2} tr \left( \left( \Sigma^{-1} XX^T \Sigma^{-1} - n \Sigma^{-1} \right) d\Sigma \right) + tr \left( \left( \frac{1}{2} \Sigma^{-1} \Psi \Sigma^{-1} - \frac{(m+q+1)}{2} \Sigma^{-1} \right) d\Sigma \right) \quad (45b)$$
which implies that

\[
\begin{align*}
0 &= \Sigma^{-1} XX^T \Sigma^{-1} - n \Sigma^{-1} \\
&\quad + \Sigma^{-1} \Psi \Sigma^{-1} - (m + q + 1) \Sigma^{-1} \\
&= XX^T - n \Sigma + \Psi - (m + q + 1) \Sigma
\end{align*}
\]  

(46a)

(46b)

and

\[
\Sigma = \frac{1}{m + n + q + 1} (XX^T + \Psi)
\]

(47)

which is seen to be identical to (17).

**APPENDIX C**

**DERIVATION OF MAXIMUM A POSTERIORI UNCOMMITTED INVERTED WISHART ESTIMATE (MAP-UIW)**

In the following, we will derive (19). For the moment assume that \( \Sigma \) is invertible, consider (14) and let \( \Psi = \alpha^2 \Sigma^{-1} \). We use the improper inverted Wishart prior \( p(\Sigma | \alpha^2 \Sigma^{-1}, m) \). Considering only varying \( \Sigma \), the differential log of \( p \) istmi-crimi-2131150.xml

\[
\begin{align*}
d \log p(\Sigma | \alpha^2 \Sigma^{-1}, m) &= \frac{m}{2} d \log \left[ \frac{1}{\alpha^2 \Sigma^{-1}} \right] - \frac{1}{2} \left( (m + q + 1) d \log | \Sigma | \right) \\
&= \frac{m}{2} \text{tr} \left( (\alpha \Sigma) \Sigma^{-1} \right) - \frac{1}{2} \text{tr} \left( (m + q + 1) \Sigma^{-1} \right) \\
&= \frac{m}{2} \text{tr} \left( (\alpha \Sigma) \Sigma^{-1} \right) - \frac{1}{2} \text{tr} \left( (m + q + 1) \Sigma^{-1} \right) \\
&= \frac{m}{2} \text{tr} \left( (\alpha \Sigma) (\Sigma^{-1} - (m + q + 1) I) \right) \\
&= \frac{1}{2} \text{tr} \left( (XX^T \Sigma^{-1} - n I) (\alpha \Sigma) (\Sigma^{-1} - (m + q + 1) I) \right) \\
&= \frac{1}{2} \text{tr} \left( (XX^T \Sigma^{-1} - n \Sigma + \Psi - (m + q + 1) \Sigma) \right)
\end{align*}
\]

(48a)

(48b)

(48c)

(48d)

The MAP using independent and identically distributed (i.i.d.) Gaussian likelihood is found to be

\[
\begin{align*}
0 &= d \log p(x_1, \ldots, x_n | \mu, \Sigma) + d \log p(\Sigma | \alpha^2 \Sigma^{-1}, m) \\
&= \frac{1}{2} \text{tr} \left( (XX^T \Sigma^{-1} - n \Sigma) (\alpha \Sigma) (\Sigma^{-1} - (m + q + 1) I) \right) \\
&= \frac{1}{2} \text{tr} \left( XX^T \Sigma^{-1} + \alpha^2 \Sigma^{-1} - \frac{n + 2m + q + 1}{2} I \right)
\end{align*}
\]

(49a)

(49b)

which implies that

\[
\begin{align*}
0 &= \frac{1}{2} XX^T \Sigma^{-1} + \alpha^2 \Sigma^{-1} - \frac{n + 2m + q + 1}{2} I \\
&= \frac{1}{2} XX^T \Sigma + \alpha^2 I - \frac{n + 2m + q + 1}{2} \Sigma^2.
\end{align*}
\]

(50a)

(50b)

Notice, that the above equation is also valid for positive semidefinite matrices \( \Sigma \). Hence, we conclude

\[
\Sigma^2 = \frac{1}{(n + 2m + q + 1)} (XX^T \Sigma + 2 \alpha^2 I)
\]

(51)

which is seen to be (19).

**APPENDIX D**

**DERIVATION OF MAXIMUM A POSTERIORI GAUSSIAN ESTIMATE (MAP-G)**

In the following, we will derive (II-C). Starting from (22) and only considering varying \( \Sigma \), we find the differential log as

\[
\begin{align*}
\frac{d \log p(\Sigma)}{d \Sigma} &= -\frac{1}{2 \sigma^2} \text{tr} \left( (\Sigma - B)^T (\Sigma - B) \right) \\
&= -\frac{1}{\sigma^2} \text{tr} \left( (\Sigma - B)^T d \Sigma \right)
\end{align*}
\]

(52)

(53)

The MAP using i.i.d. Gaussian likelihood is found to be

\[
\begin{align*}
0 &= d \log p(x_1, \ldots, x_n | \mu, \Sigma) + d \log p(\Sigma | B, \sigma) \\
&= \frac{1}{2} \text{tr} \left( (\Sigma^{-1} XX^T \Sigma^{-1} - n \Sigma^{-1}) (\sigma \Sigma^{-1}) \right) \\
&= -\frac{1}{\sigma^2} \text{tr} \left( (\Sigma - B)^T d \Sigma \right)
\end{align*}
\]

(54a)

(54b)

which implies that

\[
\begin{align*}
0 &= \Sigma^{-1} XX^T \Sigma^{-1} - n \Sigma^{-1} - \frac{2}{\sigma^2} (\Sigma - B)^T \Sigma \\
&= XX^T - n \Sigma - \frac{2}{\sigma^2} \Sigma (\Sigma - B)^T \Sigma \\
&= XX^T - n \Sigma - \frac{2}{\sigma^2} \Sigma (\Sigma - B)^T \Sigma
\end{align*}
\]

(55a)

(55b)

and

\[
\Sigma = \frac{1}{n} \left( XX^T - \frac{2}{\sigma^2} \Sigma (\Sigma - B)^T \Sigma \right)
\]

(56)

The equation above can be seen as the third degree matrix polynomial given in (II-C).

**ACKNOWLEDGMENT**

The authors would like to acknowledge the Center for Clinical and Basic Research, Ballerup, Denmark, for providing the annotated vertebrae radiographs and the MRI knee scans. The authors would also like to thank the anonymous reviewers for their useful comments.

**REFERENCES**


