Estimating Covariance Matrices

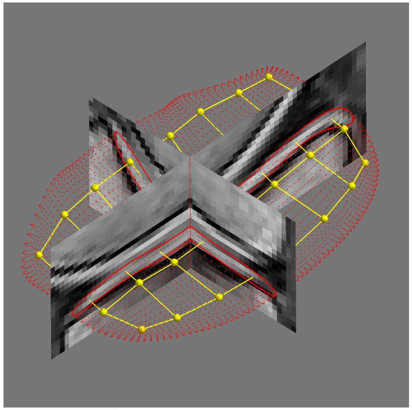
WARNING: EQUATIONS INSIDE!

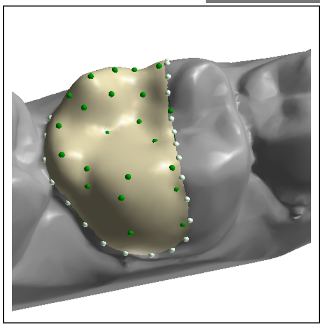
Jon Sporring

Department of Computer Science, University of Copenhagen
Universitetsparken 1, DK-2100 Copenhagen, Denmark

Ven — 2009

Motivation: Statistical shape modelling





Standard approach: Principal Component Analysis

Given a set of shapes, $\boldsymbol{x}_i \in \mathbb{R}^n$ and a mean shape $\boldsymbol{\mu}$, construct

$$\boldsymbol{X} = [\boldsymbol{x}_1 - \boldsymbol{\mu}, \boldsymbol{x}_2 - \boldsymbol{\mu}, \ldots, \boldsymbol{x}_m - \boldsymbol{\mu}] \tag{1}$$

and calculate maximum likelihood estimate of covariance matrix:

$$\boldsymbol{C} = \frac{1}{m}\boldsymbol{X}\boldsymbol{X}^T \tag{2}$$

Calculate the eigenvalue decomposition $\boldsymbol{C} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T$ with $\boldsymbol{v}_i = \boldsymbol{V}_{*i}$ and $\lambda_i = \boldsymbol{\Lambda}_{ii}$ being the i'th eigenvector and -value, then the linear space of shape variation:

$$\boldsymbol{x}(\boldsymbol{b}) = \boldsymbol{\mu} + \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{b} \tag{3}$$

Problem:

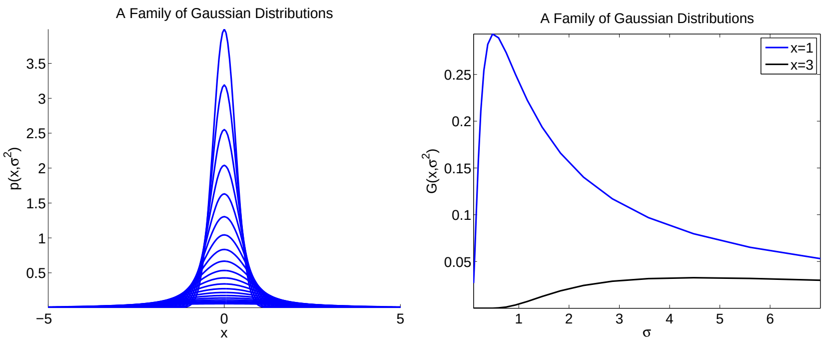
1. Undersampling, $n > m$
2. Linearity

Basics: Maximum Likelihood to Estimate Variance

Consider one random variable: Y, $E(Y) = \mu$, $X = Y - \mu$. Assume,

$$X \sim \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right) = G(x|0,\sigma^2) \tag{4}$$

Estimate σ^2 by maximizing $G(x|0,\sigma^2)$:



$$0 = \frac{\partial \log G(x|0,v)}{\partial v} = -\frac{1}{2v} + \frac{x^2}{2v^2} \quad \Rightarrow \quad \tilde{v} = x^2 \quad \Rightarrow \quad \int_{-\infty}^{\infty} x^2\, G(x|0,\sigma^2)\, dx = \sigma^2 \tag{5}$$

Many samples independently and identically distributed: $y_i, i = 1 \dots m$, $x_i = y_i - \mu$:

$$\prod_{i=1}^{m} G(x_i|0,\sigma^2) = \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^m} \exp\left(-\frac{\sum_{i=1}^{m} x_i^2}{2\sigma^2}\right) \quad \Rightarrow \quad \sigma^2 = \frac{1}{m}\sum_{i=1}^{m} x_i^2 \tag{6}$$

High Dimensional Correlated Gaussian ⇒ Hard Work

Many random and correlated variables: $X_i, i = 1 \ldots n$, $E(X_i) = \mu_i$, $E((X_i - \mu_i)(X_j - \mu_j)) = \boldsymbol{C}$, many samples $\boldsymbol{x}_j, j = 1 \ldots m$

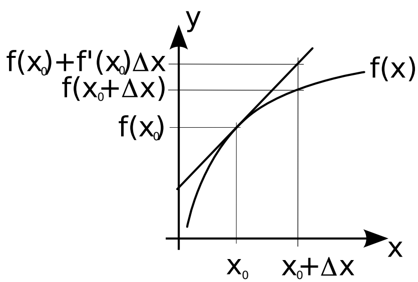
$$\frac{1}{\left(\left(\sqrt{2\pi}\right)^n \sqrt{|\boldsymbol{C}|}\right)^m} \exp\left(-\frac{1}{2}\sum_{i=1}^{m}(\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{C}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})\right) = G(\boldsymbol{x}_1, ..., \boldsymbol{x}_m|\boldsymbol{\mu}, \boldsymbol{C}) \tag{7}$$

Consider Taylor series of $f : \mathbb{R} \to \mathbb{R}$:

$$f(x + \Delta x) = f(x) + f'(x)\Delta x + \mathcal{O}(\Delta x^2) \tag{8a}$$

$$\Rightarrow \quad \Delta f(x) = f(x + \Delta x) - f(x) = f'(x)\Delta x + \mathcal{O}(\Delta x^2) \tag{8b}$$

$$\Rightarrow \quad df = f'(x)\, dx \tag{8c}$$



But What About Vector and Matrix Equations?

Taylor series of $\boldsymbol{f} : \mathbb{R}^m \to \mathbb{R}^n$:

$$f_i(\boldsymbol{x} + \Delta\boldsymbol{x}) = f_i(\boldsymbol{x}) + Df_i(\boldsymbol{x})\Delta\boldsymbol{x} + \mathcal{O}(\|\Delta\boldsymbol{x}\|^2) \quad \Rightarrow \quad d\boldsymbol{f} = D\boldsymbol{f}(\boldsymbol{x})\,d\boldsymbol{x} \tag{9}$$

where $\{D\boldsymbol{f}(\boldsymbol{x})\}_{ij} = \{\partial f_i(\boldsymbol{x})/\partial x_j\}$. Useful relations (const. $\boldsymbol{A}$):

$$\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x} \quad \Rightarrow \quad d\boldsymbol{f} = \boldsymbol{A}\,d\boldsymbol{x} \tag{10a}$$

$$\boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} \quad \Rightarrow \quad d\boldsymbol{g} = d\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} + \boldsymbol{x}^T\boldsymbol{A}\,d\boldsymbol{x} = d\boldsymbol{x}^T(\boldsymbol{A} + \boldsymbol{A}^T)\boldsymbol{x} = \boldsymbol{x}^T(\boldsymbol{A}^T + \boldsymbol{A})\,d\boldsymbol{x} \tag{10b}$$

Taylor series of $\boldsymbol{F} : \mathbb{R}^{m\times n} \to \mathbb{R}^{p\times q} \Leftrightarrow$ Taylor series of $\mathrm{vec}(\boldsymbol{F}) = [\boldsymbol{F}_{*1}^T|\ldots|\boldsymbol{F}_{*q}^T]^T \in \mathbb{R}^{pq}$ in $\mathrm{vec}(\boldsymbol{X}) \in \mathbb{R}^{mn}$,

$$\boldsymbol{f}(\mathrm{vec}(\boldsymbol{X})) = \mathrm{vec}(\boldsymbol{F}(\boldsymbol{X})) \tag{11a}$$

$$D\boldsymbol{F}(\boldsymbol{X}) = D\boldsymbol{f}(\mathrm{vec}(\boldsymbol{X})) \tag{11b}$$

Useful relations ($\boldsymbol{X}$ invertible):

$$d\mathrm{tr}\,(\boldsymbol{X}) = \mathrm{tr}\,(d\boldsymbol{X}) \tag{12a}$$

$$d\,|\boldsymbol{X}| = |\boldsymbol{X}|\,\mathrm{tr}\left(\boldsymbol{X}^{-1}\,d\boldsymbol{X}\right) \tag{12b}$$

$$d\,\|\boldsymbol{X}\|^2 = d\mathrm{tr}\left(\boldsymbol{X}^T\boldsymbol{X}\right) = \mathrm{tr}\left(d\boldsymbol{X}^T\,\boldsymbol{X} + \boldsymbol{X}^T\,d\boldsymbol{X}\right) = 2\mathrm{tr}\left(\boldsymbol{X}^T\,d\boldsymbol{X}\right) \tag{12c}$$

$$\boldsymbol{I} = \boldsymbol{X}\boldsymbol{X}^{-1} \quad \Rightarrow \quad \boldsymbol{0} = d\boldsymbol{X}\,\boldsymbol{X}^{-1} + \boldsymbol{X}\,d\boldsymbol{X}^{-1} \quad \Rightarrow \quad d\boldsymbol{X}^{-1} = -\boldsymbol{X}^{-1}\,d\boldsymbol{X}\,\boldsymbol{X}^{-1} \tag{12d}$$

Back to Multivariate Gaussian

Trick: $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m] - \boldsymbol{\mu}\mathbf{1}_m^T$, $\sum_{i=1}^m (\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{C}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}) = \operatorname{tr}\left(\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{C}^{-1}\right)$,

$$G(\boldsymbol{x}_1, ..., \boldsymbol{x}_m | \boldsymbol{\mu}, \boldsymbol{C}) = \frac{1}{\left(\left(\sqrt{2\pi}\right)^n \sqrt{|\boldsymbol{C}|}\right)^m} \exp\left(-\frac{1}{2}\operatorname{tr}\left(\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{C}^{-1}\right)\right) \tag{13a}$$

$$\Rightarrow \log G(\boldsymbol{x}_1, ..., \boldsymbol{x}_m | \boldsymbol{\mu}, \boldsymbol{C}) = -\frac{m}{2}\left(n \log 2\pi + \log|\boldsymbol{C}|\right) - \frac{1}{2}\operatorname{tr}\left(\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{C}^{-1}\right) = L \tag{13b}$$

$$\Rightarrow dL = -\frac{m}{2} d\log|\boldsymbol{C}| - \frac{1}{2} d\operatorname{tr}\left(\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{C}^{-1}\right) \tag{13c}$$

$$d\log|\boldsymbol{C}| = \frac{1}{|\boldsymbol{C}|} d|\boldsymbol{C}| = \frac{1}{|\boldsymbol{C}|}|\boldsymbol{C}| \operatorname{tr}\left(\boldsymbol{C}^{-1}\, d\boldsymbol{C}\right) = \operatorname{tr}\left(\boldsymbol{C}^{-1}\, d\boldsymbol{C}\right) \tag{14a}$$

$$d\operatorname{tr}\left(\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{C}^{-1}\right) = \operatorname{tr}\left(d\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{C}^{-1} + \boldsymbol{X}\, d\boldsymbol{X}^T\boldsymbol{C}^{-1} + \boldsymbol{X}\boldsymbol{X}^T\, d\boldsymbol{C}^{-1}\right) \tag{14b}$$

$$= \operatorname{tr}\left(\boldsymbol{X}^T\boldsymbol{C}^{-1}\, d\boldsymbol{X} + (\boldsymbol{C}^{-1})^T\, d\boldsymbol{X}\, \boldsymbol{X}^T - \boldsymbol{X}\boldsymbol{X}^T\boldsymbol{C}^{-1}\, d\boldsymbol{C}\, \boldsymbol{C}^{-1}\right) \tag{14c}$$

$$= \operatorname{tr}\left(\boldsymbol{X}^T\boldsymbol{C}^{-1}\, d\boldsymbol{X} + \boldsymbol{X}^T(\boldsymbol{C}^{-1})^T\, d\boldsymbol{X} - \boldsymbol{C}^{-1}\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{C}^{-1}\, d\boldsymbol{C}\right) \tag{14d}$$

$$= \operatorname{tr}\left(2\boldsymbol{X}^T\boldsymbol{C}^{-1}\, d\boldsymbol{X} - \boldsymbol{C}^{-1}\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{C}^{-1}\, d\boldsymbol{C}\right) \tag{14e}$$

Differential may be Treated in Parts

Mean part:

$$d\boldsymbol{X} = -d\boldsymbol{\mu}\,\mathbf{1}_m^T \tag{15a}$$

$$\Rightarrow 0 = dL_{\boldsymbol{\mu}} = -\mathrm{tr}\left(\boldsymbol{X}^T\boldsymbol{C}^{-1}\,d\boldsymbol{X}\right) = \mathrm{tr}\left(\boldsymbol{X}^T\boldsymbol{C}^{-1}\,d\boldsymbol{\mu}\,\mathbf{1}_m^T\right) = \mathrm{tr}\left(\mathbf{1}_m^T([\boldsymbol{x}_1,\ldots,\boldsymbol{x}_m]^T - \mathbf{1}_m\boldsymbol{\mu}^T)\boldsymbol{C}^{-1}\,d\boldsymbol{\mu}\right) \tag{15b}$$

$$\Rightarrow 0 = \mathbf{1}_m^T([\boldsymbol{x}_1,\ldots,\boldsymbol{x}_m]^T - \mathbf{1}_m\boldsymbol{\mu}^T) = \mathbf{1}_m^T[\boldsymbol{x}_1,\ldots,\boldsymbol{x}_m]^T - m\boldsymbol{\mu}^T \tag{15c}$$

$$\Rightarrow \boldsymbol{\mu} = \frac{1}{m}\sum_{i=1}^{m}\boldsymbol{x}_i \tag{15d}$$

Covariance part:

$$0 = dL_{\boldsymbol{C}} = -\frac{m}{2}\mathrm{tr}\left(\boldsymbol{C}^{-1}\,d\boldsymbol{C}\right) + \frac{1}{2}\mathrm{tr}\left(\boldsymbol{C}^{-1}\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{C}^{-1}\,d\boldsymbol{C}\right) = \mathrm{tr}\left(-m\boldsymbol{C}^{-1}\,d\boldsymbol{C} + \boldsymbol{C}^{-1}\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{C}^{-1}\,d\boldsymbol{C}\right) \tag{16a}$$

$$\Rightarrow 0 = \boldsymbol{C}^{-1}\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{C}^{-1} - m\boldsymbol{C}^{-1} \tag{16b}$$

$$\Rightarrow m\boldsymbol{C}^{-1} = \boldsymbol{C}^{-1}\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{C}^{-1} \tag{16c}$$

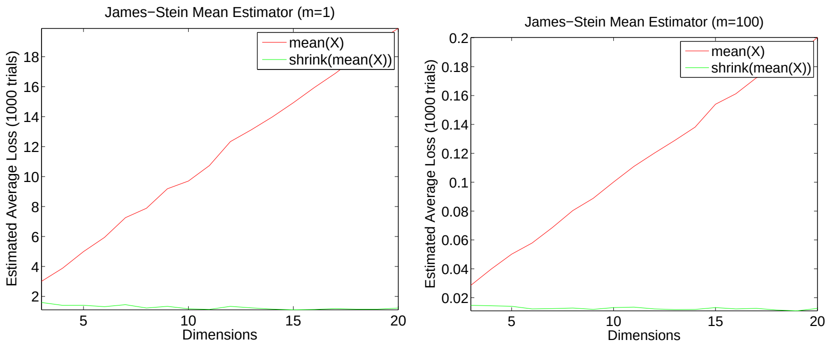
$$\Rightarrow \boldsymbol{C} = \frac{1}{m}\boldsymbol{X}\boldsymbol{X}^T \tag{16d}$$

$$\tag{16e}$$

Log-Likelihood Mean Estimator does not Min. Quadratic Loss!

Assume $\boldsymbol{X} \in \mathbb{R}^n$, $E(\boldsymbol{X}) = \boldsymbol{\mu}$, and $\tilde{\boldsymbol{\mu}} = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{x}_i$

$$L_q(\boldsymbol{\mu}, \tilde{\boldsymbol{\mu}}) = \|\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}\|^2 \quad (17)$$



Consider $m = 1$, $\boldsymbol{x}_1 = \boldsymbol{\mu} + \boldsymbol{\epsilon}$, $\tilde{\boldsymbol{\mu}} = \boldsymbol{x}_1$:

$$L_q = \sum_{i=1}^{m} (\mu_i - \boldsymbol{x}_1)^2 = \sum_{i=1}^{m} \boldsymbol{\epsilon}_i^2 \simeq m \quad (18)$$

Hence, shrink for $n > 2$ (Stein 1956, James-Stein 1961):

$$\hat{\boldsymbol{\mu}}(\tilde{\boldsymbol{\mu}}, \boldsymbol{w}) = \tilde{\boldsymbol{\mu}} - \frac{n-2}{m(\tilde{\boldsymbol{\mu}} - \boldsymbol{w})^T C^{-1} (\tilde{\boldsymbol{\mu}} - \boldsymbol{w})} (\tilde{\boldsymbol{\mu}} - \boldsymbol{w}) \quad (19)$$

Log-Likelihood Covariance Estimator does not Min. Loss!

Consider:

$$L_q(\boldsymbol{C}, \tilde{\boldsymbol{C}}) = \operatorname{tr}\left(\left(\tilde{\boldsymbol{C}}\boldsymbol{C}^{-1} - \boldsymbol{I}\right)^2\right) \tag{20}$$

The optimal estimator invariant to $\boldsymbol{C} \rightarrow \boldsymbol{H}\boldsymbol{C}\boldsymbol{H}^T$, $\tilde{\boldsymbol{C}} - \boldsymbol{H}\tilde{\boldsymbol{C}}\boldsymbol{H}^T$, where $\boldsymbol{H}$ lower triangular is,

$$\hat{\boldsymbol{C}}(m\tilde{\boldsymbol{C}}) = \boldsymbol{T}\boldsymbol{D}\boldsymbol{T}^T \tag{21}$$

where $\boldsymbol{T}$ lower triangular such that $m\tilde{\boldsymbol{C}} = \boldsymbol{T}\boldsymbol{T}^{\boldsymbol{T}}$, and $\boldsymbol{D} = \operatorname{diag}\left(\boldsymbol{F}^{-1}\boldsymbol{f}\right)$, where

$$F_{ii} = (n + m - 2i + 1)(n + m - 2i + 3) \tag{22a}$$

$$F_{ij} = (n + m - 2j + 1) \tag{22b}$$

$$f_i = n + m + 2i + 1 \tag{22c}$$

Undersampling ⇒ Maximum A Posteriori

Small set of samples requires added knowledge, i.e. Bayes theorem

$$P(\boldsymbol{\mu}, \boldsymbol{C} | \boldsymbol{x}_1, ..., \boldsymbol{x}_m) = \frac{P(\boldsymbol{x}_1, ..., \boldsymbol{x}_m | \boldsymbol{\mu}, \boldsymbol{C}) P(\boldsymbol{\mu}, \boldsymbol{C})}{P(\boldsymbol{x}_1, ..., \boldsymbol{x}_m)}, \tag{23}$$

The point of maximum a posteriori density is practical and found by

$$0 = d \log P(\boldsymbol{\mu}, \boldsymbol{C} | \boldsymbol{x}_1, ..., \boldsymbol{x}_m) = d \log P(\boldsymbol{x}_1, ..., \boldsymbol{x}_m | \boldsymbol{\mu}, \boldsymbol{C}) + d \log P(\boldsymbol{\mu}, \boldsymbol{C}) - d \log P(\boldsymbol{x}_1, ..., \boldsymbol{x}_m), \tag{24}$$

$$d \log P(\boldsymbol{x}_1, ..., \boldsymbol{x}_m | \boldsymbol{\mu}, \boldsymbol{C}) = \frac{1}{2} \mathrm{tr} \left(\left(\boldsymbol{C}^{-1} \boldsymbol{X} \boldsymbol{X}^T \boldsymbol{C}^{-1} - m \boldsymbol{C}^{-1} \right) d\boldsymbol{C} \right) \tag{25}$$

Inverted Wishart prior gives Simple Structure

Assuming independent $P(\boldsymbol{\mu}, \boldsymbol{C}) = P(\boldsymbol{\mu})P(\boldsymbol{C})$ and consider prior $P(\boldsymbol{C})$. E.g. Inverted Wishart distribution

$$\mathcal{W}^{-1}(\boldsymbol{C}|\boldsymbol{\Psi}, \eta) = \frac{|\boldsymbol{\Psi}|^{\eta/2} \exp -\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Psi}\boldsymbol{C}^{-1}\right)}{2^{\eta n/2} |\boldsymbol{C}|^{(\eta+n+1)/2} \Gamma_n\left(\frac{\eta}{2}\right)} \tag{26a}$$

$$\Rightarrow d\log \mathcal{W}^{-1}(\boldsymbol{C}|\boldsymbol{\Psi}, \eta) = -\frac{(\eta+n+1)}{2} d\log|\boldsymbol{C}| - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Psi} d\boldsymbol{C}^{-1}\right) \tag{26b}$$

$$= -\frac{(\eta+n+1)}{2}\mathrm{tr}\left(\boldsymbol{C}^{-1} d\boldsymbol{C}\right) + \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Psi}\boldsymbol{C}^{-1}\, d\boldsymbol{C}\,\boldsymbol{C}^{-1}\right) \tag{26c}$$

$$= \mathrm{tr}\left(\left(\frac{1}{2}\boldsymbol{C}^{-1}\boldsymbol{\Psi}\boldsymbol{C}^{-1} - \frac{(\eta+n+1)}{2}\boldsymbol{C}^{-1}\right) d\boldsymbol{C}\right) \tag{26d}$$

$$d\log P(\boldsymbol{x}_1, ..., \boldsymbol{x}_m|\boldsymbol{\mu}, \boldsymbol{C}) + d\log \mathcal{W}^{-1}(\boldsymbol{C}|\boldsymbol{\Psi}, \eta) \tag{27a}$$

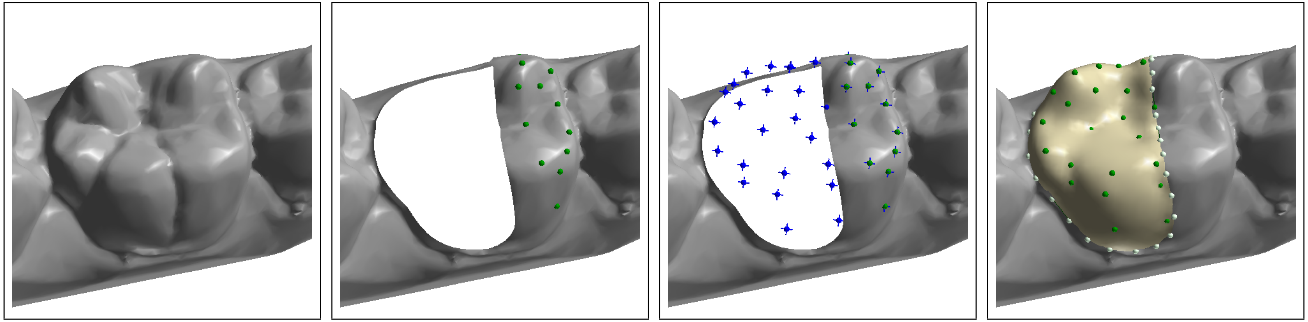
$$= \frac{1}{2}\mathrm{tr}\left(\left(\boldsymbol{C}^{-1}\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{C}^{-1} - m\boldsymbol{C}^{-1}\right) d\boldsymbol{C}\right) + \mathrm{tr}\left(\left(\frac{1}{2}\boldsymbol{C}^{-1}\boldsymbol{\Psi}\boldsymbol{C}^{-1} - \frac{(\eta+n+1)}{2}\boldsymbol{C}^{-1}\right) d\boldsymbol{C}\right) \tag{27b}$$

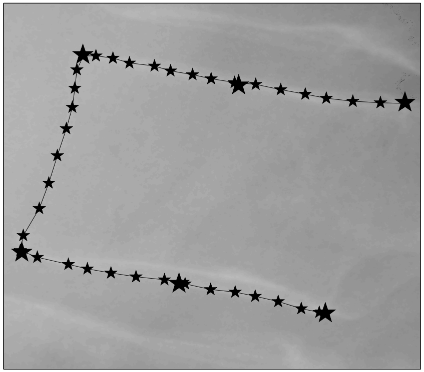
$$\Rightarrow 0 = \boldsymbol{C}^{-1}\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{C}^{-1} - m\boldsymbol{C}^{-1} + \boldsymbol{C}^{-1}\boldsymbol{\Psi}\boldsymbol{C}^{-1} - (\eta+n+1)\boldsymbol{C}^{-1} \tag{27c}$$

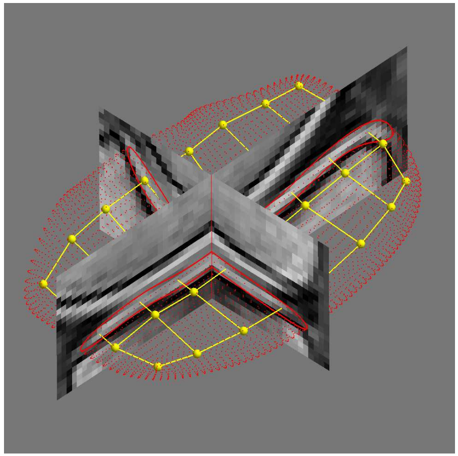
$$\Rightarrow 0 = \boldsymbol{X}\boldsymbol{X}^T - m\boldsymbol{C} + \boldsymbol{\Psi} - (\eta+n+1)\boldsymbol{C} \tag{27d}$$

$$\Rightarrow \boldsymbol{C} = \frac{1}{\eta+m+n+1}\left(\boldsymbol{X}\boldsymbol{X}^T + \boldsymbol{\Psi}\right) \tag{27e}$$

What are good values for Ψ?







What do you believe in?

$$P_{\text{Gauss}}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-x^2}{2\sigma^2}\right) \tag{28a}$$

$$P_{\text{Exp}}(x) = \frac{1}{\boldsymbol{\mu}} \exp\left(\frac{-x}{\boldsymbol{\mu}^2}\right), \quad x \geq 0 \tag{28b}$$

Gives polynomial systems of equations:

$$m\boldsymbol{C} = \boldsymbol{X}\boldsymbol{X}^T + \tag{29}$$

	P_{Gauss}	P_{Exp}
$\operatorname{tr}(\boldsymbol{C})$	$\frac{\operatorname{tr}(\boldsymbol{C})}{\sigma^2}\boldsymbol{C}^2$	$\frac{1}{\boldsymbol{\mu}^2}\boldsymbol{C}^2$
$\lvert\boldsymbol{C}\rvert$	$\frac{\lvert\boldsymbol{C}\rvert^2}{\sigma^2}\boldsymbol{C}$	$\frac{1}{\boldsymbol{\mu}^2}\lvert\boldsymbol{C}\rvert\,\boldsymbol{C}$
$\lVert\boldsymbol{C}\rVert$	$\frac{1}{\sigma^2}\boldsymbol{C}\boldsymbol{C}^T\boldsymbol{C}$	$\frac{1}{\boldsymbol{\mu}^2\lVert\boldsymbol{C}\rVert}\boldsymbol{C}\boldsymbol{C}^T\boldsymbol{C}$