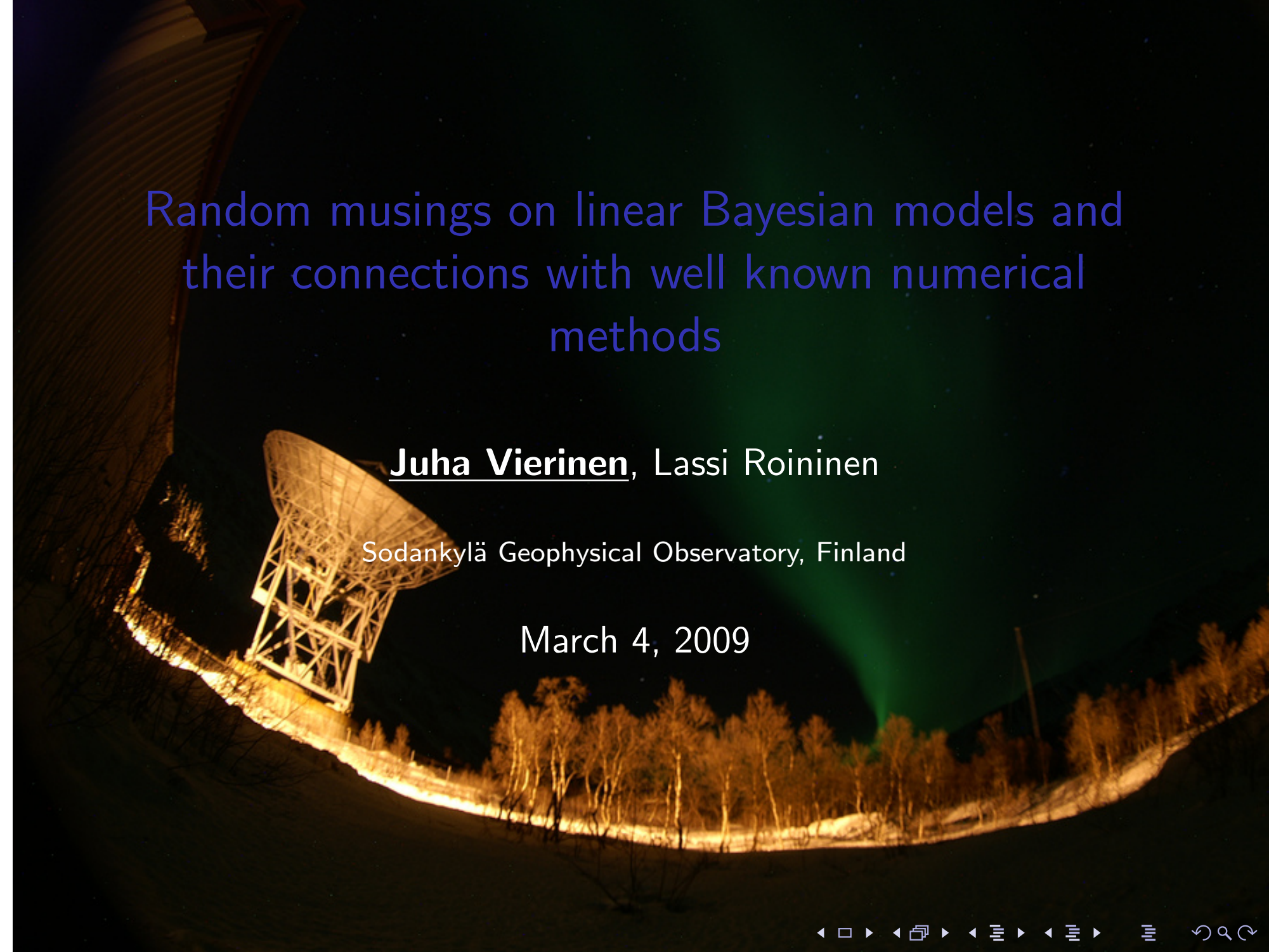


Random musings on linear Bayesian models and their connections with well known numerical methods

Juha Vierinen, Lassi Roininen

Sodankylä Geophysical Observatory, Finland

March 4, 2009

Introduction

- Tikhonov regularization, Truncated SVD, Wiener filtering and simple correlation can be understood as linear statistical models with some kind of (possibly hidden) prior assumptions.
- At least the statistical meaning of Tikhonov regularization has been reported before by several authors (e.g., Kaipio & Somersalo, and probably Lehtinen).

Linear statistical model

Let us consider the following linear measurement model $m = Ax + \xi$, and assume that $x \sim N(0, \Sigma_p)$ and $\xi \sim N(0, \Sigma)$. Using the Bayes theorem, we get the following probability distribution of x:

$$p(x|m) = \frac{p(m|x)p(x)}{p(m)} \propto p(m|x)p(x), \tag{1}$$

or

$$p(x|m) \propto \exp\left(-\overline{(m - Ax)}^T \Sigma^{-1}(m - Ax) - \overline{x}^T \Sigma_p^{-1} x\right). \tag{2}$$

The maximum a posteriori estimate, or the peak of this distribution can be obtained as:

$$x_{\mathrm{MAP}} = \left(\overline{A}^T \Sigma^{-1} A + \Sigma_p^{-1}\right)^{-1} \overline{A}^T \Sigma^{-1} m. \tag{3}$$

Wiener filtering

Now, if we assume A is infinitely extended and that it describes a convolution $m_t = \sum_{s=-\infty}^{\infty} a_{t-s} x_t$.

$$A = \begin{bmatrix} a_0 & a_{-1} & a_{-2} & \dots & \dots & 0 \\ a_1 & a_0 & a_{-1} & \ddots & & \vdots \\ a_2 & a_1 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & a_{-1} & a_{-2} \\ \vdots & & \ddots & a_1 & a_0 & a_{-1} \\ 0 & \dots & \dots & a_2 & a_1 & a_0 \end{bmatrix} \tag{4}$$

Assuming also that Σ_p and Σ are sufficiently band-dominated Toeplitz matrices with rows σ_t and $\sigma_{p,t}$, we can analyze the problem in frequency domain (everything diagonalizes) and obtain the MAP estimate as follows:

Wiener filtering: Convolution example 1

In Frequency domain:

$$\hat{x}_{\mathrm{MAP}}(f) = \left(\overline{\hat{a}(f)}\hat{\sigma}(f)^{-1}\hat{a}(f) + \hat{\sigma}_p(f)^{-1}\right)^{-1} \overline{\hat{a}(f)}\hat{\sigma}(f)^{-1}\hat{m}(f). \quad (5)$$

This simplifies to:

$$\hat{x}_{\mathrm{MAP}}(f) = \frac{\overline{\hat{a}}(f)\hat{m}(f)}{|\hat{a}(f)|^2 + \frac{\hat{\sigma}(f)}{\hat{\sigma}_p(f)}}. \quad (6)$$

This is also known as the *Wiener filter* solution. It generalizes to multiple dimensions fairly easily (one can express a ND convolution theory as a 1D convolution by organizing parameters suitably).

Wiener filtering: Image deconvolution using FFT

- Solution with FFT is fast (can solve 10^7 parameter problems in a matter of seconds).
- But assumes stationary convolution kernel and noise, not always very practical.

With terabytes of measurements, FFT makes this kind of analysis possible

Wiener filtering: Convolution example 2

Now, if we assume that $\Sigma = \mathrm{diag}(\sigma^2, ..., \sigma^2)$ and $\Sigma_p = \mathrm{diag}(\sigma_p^2, ..., \sigma_p^2)$, our measurement equation simplifies into the following form:

$$\hat{x}_{\mathrm{MAP}}(f) = \frac{\overline{\hat{a}}(f)\hat{m}(f)}{|\hat{a}(f)|^2 + \frac{\sigma^2}{\sigma_p^2}}, \tag{7}$$

and if we inspect two limiting cases (1) $\sigma^2/\sigma_p^2 \to 0$ and (2) $\sigma^2/\sigma_p^2 \gg \sup|\hat{a}(f)|^2$, we see that we get two different types of familiar filters:

1. $\lambda_t = \mathcal{F}_D^{-1}\left\{\frac{1}{\hat{a}(f)}\right\}_t$. This is the so called inverse filter, or the sidelobe-free decoding filter.
2. $h_t \approx \frac{\sigma^2}{\sigma_p^2}\overline{a_t}$, this is the so called matched filter. However, one should be careful to note that traditionally the matched filter doesn't include the σ^2/σ_p^2 term, and that when $\sigma^2/\sigma_p^2 = 0$ then our estimate also vanishes to zero.

Wiener filtering: Convolution example 2

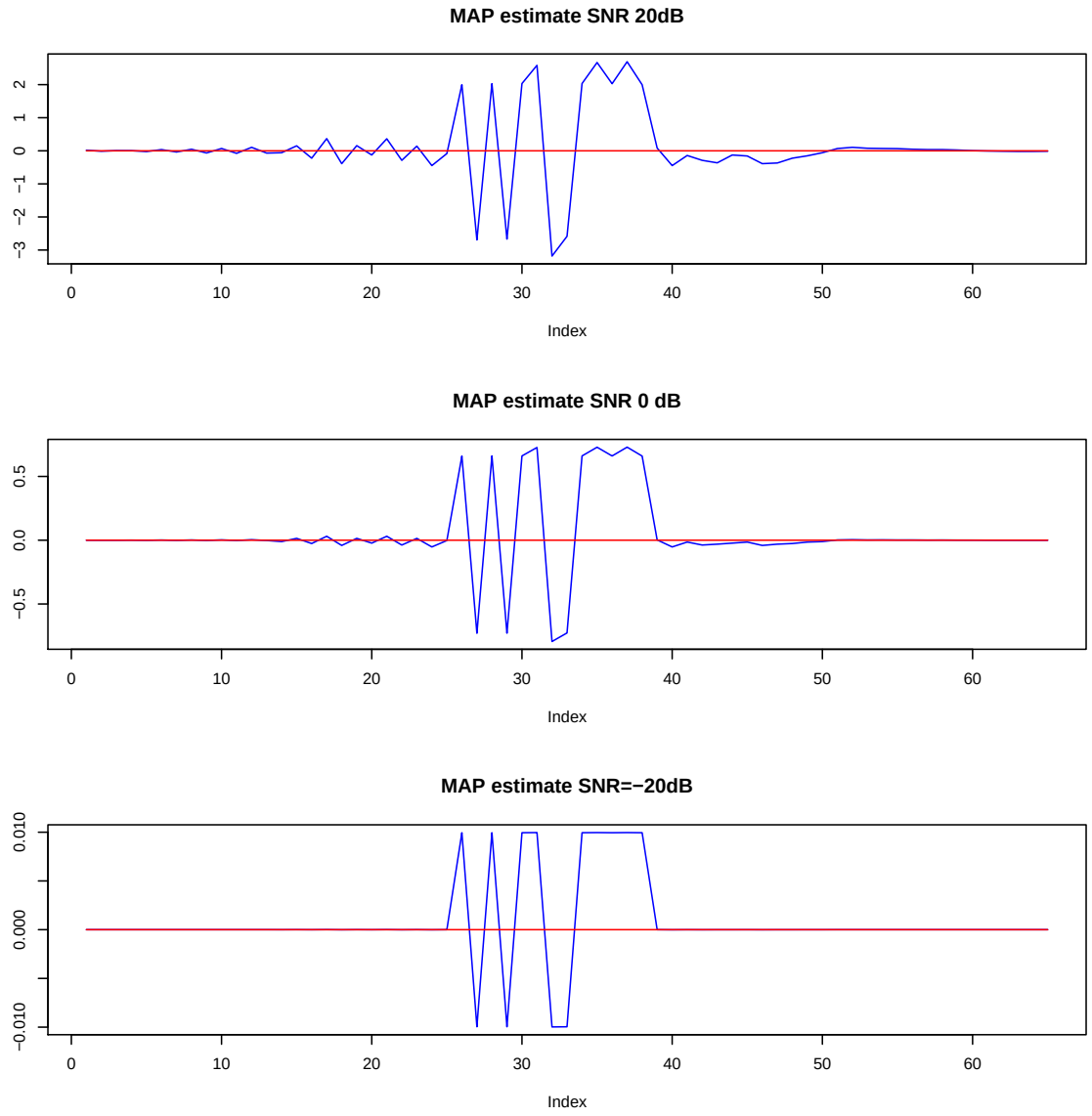


Figure: A low noise $\sigma_p/\sigma = 100$, medium noise $\sigma_p/\sigma = 1$ and high noise case $\sigma_p/\sigma = 0.01$. The transmission code is a 13 bit Barker code.

Wiener filtering: Conclusions

- An easy way to understand Wiener filtering is through linear statistical inversion
- FFT is useful, although only when everything is stationary. This is not always the case
- There is finally a statistical justification for using matched filtering for distributed targets! Although it is only applicable for very weak targets

Truncated SVD

Bog standard linear model $\mathrm{E}\epsilon\bar{\epsilon}^T = \Sigma = \mathrm{diag}\left(\sigma^2, ..., \sigma^2\right)$:

$$m = Ax + \epsilon \tag{8}$$

MAP solution with prior Σ_p^{-1}

$$x_{\mathrm{MAP}} = \left(\overline{A}^T \Sigma^{-1} A + \Sigma_p^{-1}\right)^{-1} \overline{A}^T \Sigma^{-1} m \tag{9}$$

Singular value decomposition:

$$A = UD\overline{V}^T, \tag{10}$$

where $D = \mathrm{diag}\left(d_1, ..., d_n\right)$. Writing the prior using an eigenvalue decomposition for symmetric matrices
$\Sigma_p^{-1} = V\mathrm{diag}\left(s_1, ..., s_n\right)\overline{V}^T = V\Lambda\overline{V}^T$, we get

$$x_{\mathrm{MAP}} = V(DD + \Lambda)^{-1} D\overline{U}^T m = VD^{\dagger}\overline{U}^T m, \tag{11}$$

where $D^{\dagger} = \mathrm{diag}\left(\frac{d_1}{d_1^2 + s_1}, ..., \frac{d_n}{d_n^2 + s_n}\right)$.

TSVD

It is now easy to see what is the statistical interpretation for Truncated SVD. We assume an *a priori* covariance matrix:

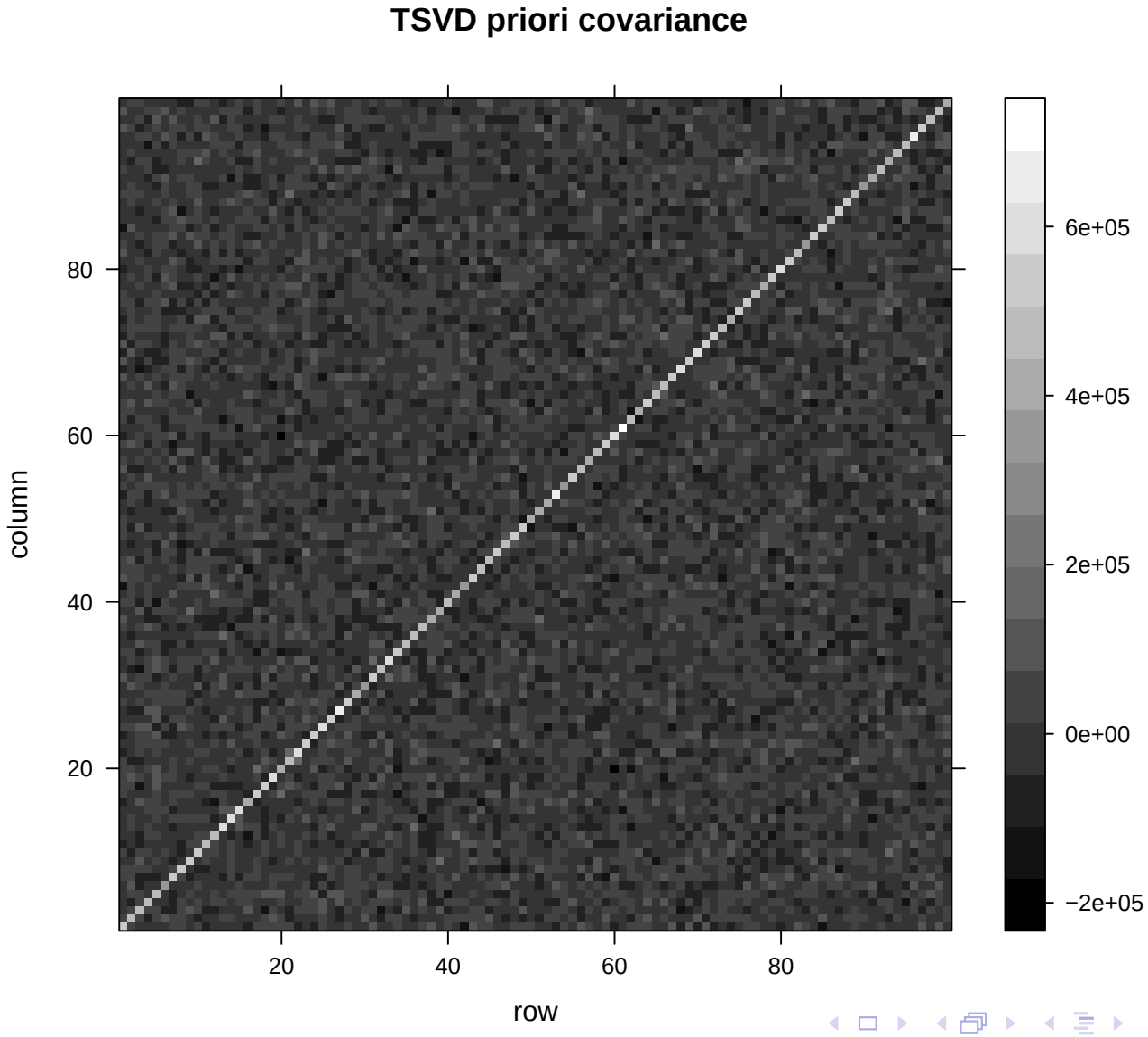
$$\Sigma_p = V \mathrm{diag}\left(s_1^{-1}, ..., s_n^{-1}\right) \overline{V}^T, \tag{12}$$

with

$$s_i \rightarrow \begin{cases} 0 & \text{when} \quad d_i > c \\ \infty & \text{otherwise} \end{cases} \tag{13}$$

TSVD: Example

A is a 100×100 random matrix. The following is a prior covariance Σ_p that corresponds to truncating 50 smallest singular values:

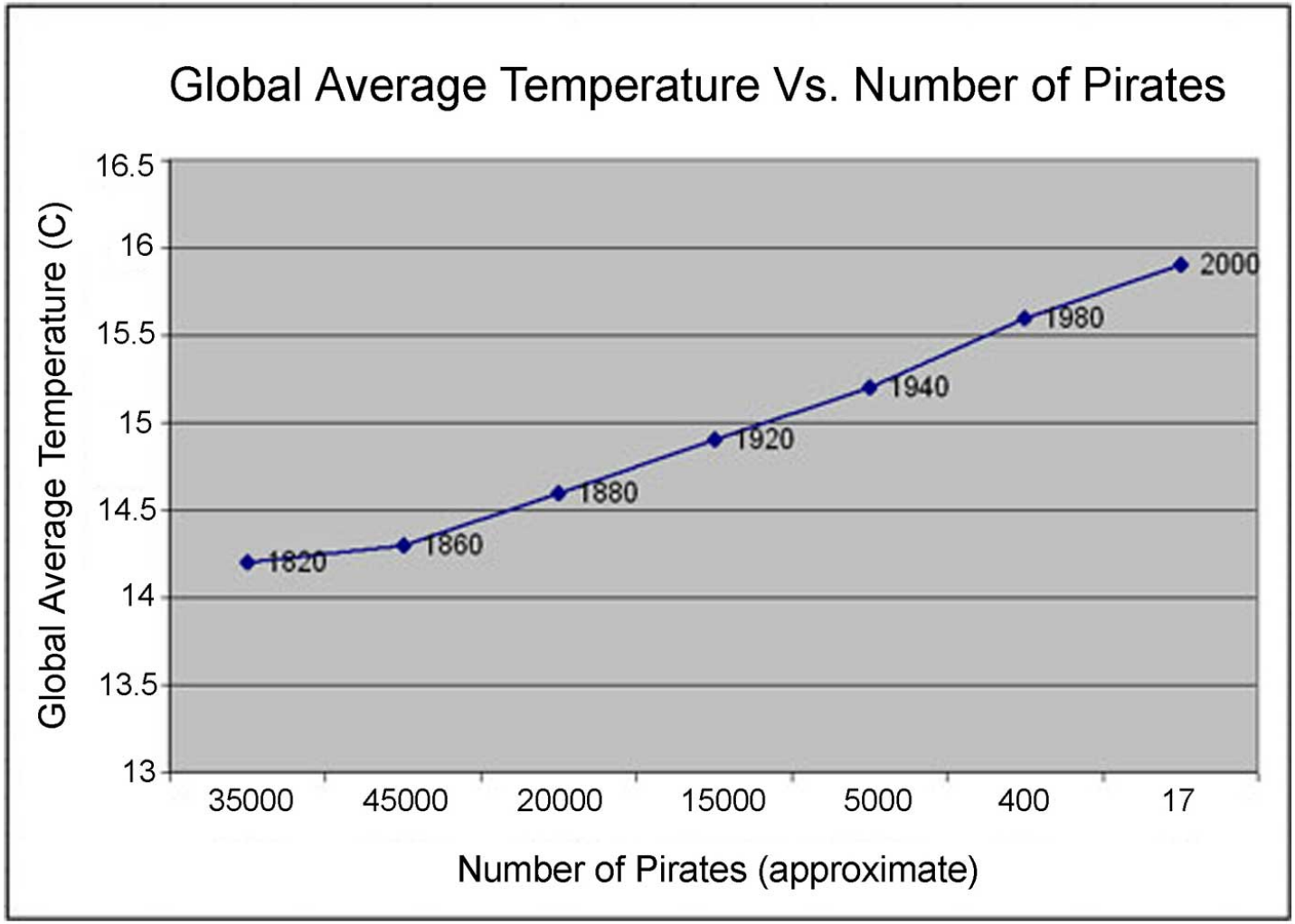


TSVD: Conclusions

- A better alternative to TSVD is to simply add some large enough values to the unstable singular values. Nearly anything should be better than setting them to infinity.

Correlation

STOP GLOBAL WARMING: BECOME A PIRATE



WWW.VENGANZA.ORG

Correlation

Again, let's start with a plain vanilla linear model:

$$m = Ax + \epsilon \tag{14}$$

Assuming no prior and that ϵ is zero mean Gaussian and $\mathrm{E}\epsilon\bar{\epsilon}^T = \Sigma = \mathrm{diag}\left(\sigma^2, ..., \sigma^2\right)$, we get a MAP estimate:

$$x_{\mathrm{MAP}} = \left(\overline{A}^T A\right)^{-1} \overline{A}^T m, \tag{15}$$

now what if $A^T A = \alpha I$? Then we get a very simple solution:

$$x_{\mathrm{MAP}} = \alpha^{-1} \overline{A}^T m, \tag{16}$$

which is basically correlating the measurement with the theory. Examples: Perfect radar transmission codes or a Fourier transform model.

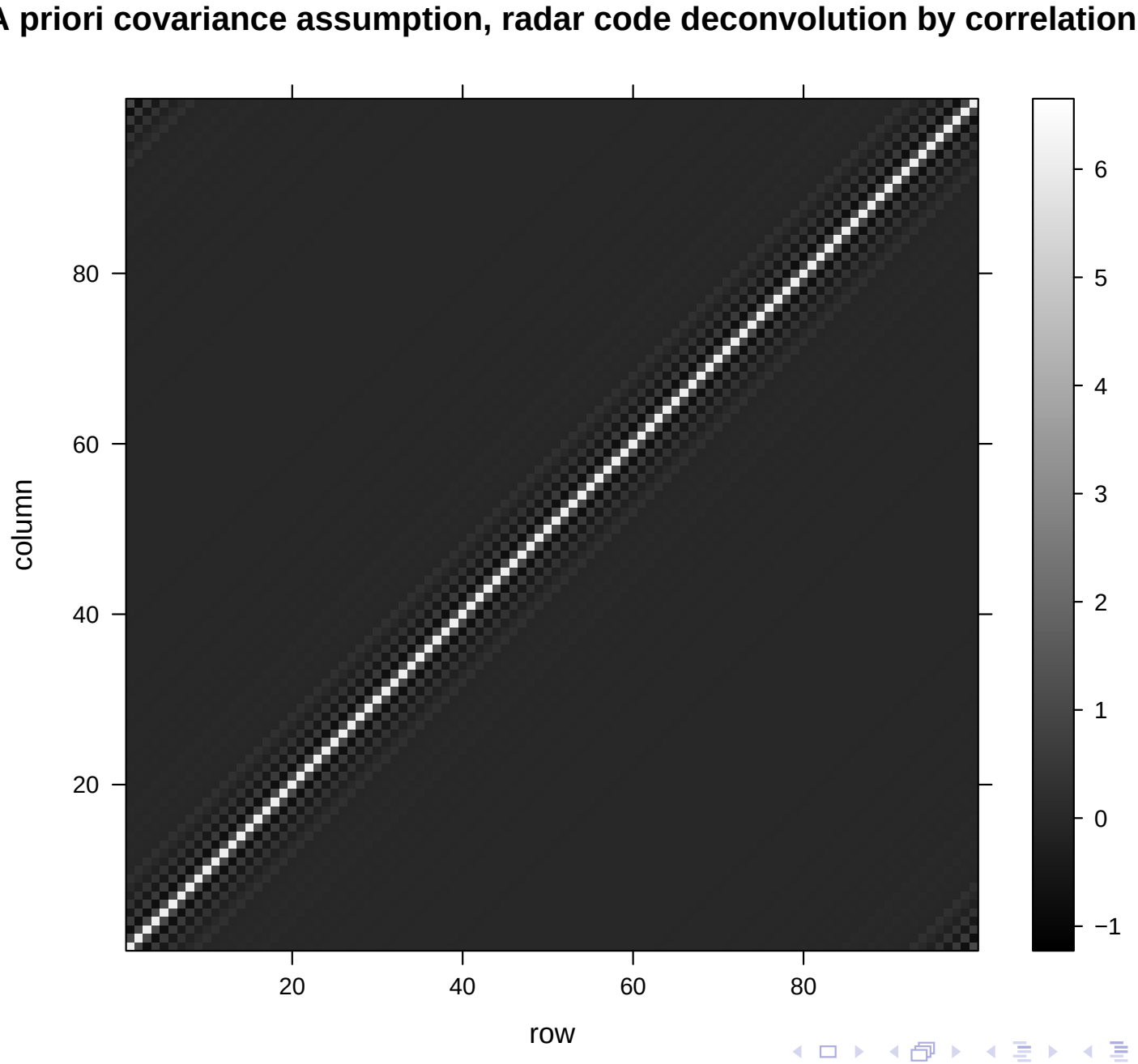
Correlation, hidden a priori assumption

Now what if $\overline{A}^T A \neq I$, but one still insists on correlating. Does this have a statistical meaning? It turns out yes. Again, we use singular value decomposition $A = UD\overline{V}^T$, where $D = \mathrm{diag}\,(d_1, ..., d_n)$. We then assume that there is some "magic" prior Σ_p that turns our maximum a posteriori solution into a correlation. Writing the prior using an eigenvalue decomposition for symmetric matrices $\Sigma_p^{-1} = V\mathrm{diag}\,(s_1, ..., s_n)\,\overline{V}^T = V\Lambda\overline{V}^T$, we get

$$x_{\mathrm{MAP}} = V(DD + \Lambda)^{-1}\overline{V}^T VD\overline{U}^T m. \tag{17}$$

Now, if $DD + \Lambda = \alpha I$, then in fact $x_{\mathrm{MAP}} = \alpha^{-1}\overline{A}^T m$, which means correlating and scaling the data.

Correlation, hidden a priori assumption



Correlation: Conclusions

- However, when $\overline{A}^T A = \alpha I$, correlation is actually a maximum likelihood solution. In the case of, e.g., Fourier transform (data is modeled as a sum of sinusoids) or perfect radar transmission codes, this is the case.
- When $\overline{A}^T A \approx \alpha I$, the solution might in some cases be a good first order approximation of the maximum likelihood solution, e.g., for *random radar code groups* or the *Lomb-Scargle periodogram* (non-uniform timestep discrete Fourier transform).

Conclusions

- Many traditional tools, such as Wiener filtering, TSVD and correlation can be understood through the framework of linear statistical inversion in a more general form. This helps to understand what the numerical method actually does and possibly to propose enhancements.

Conclusions

THERE ARE LIES,
DAMN LIES,
STATISTICS,
AND THEN THERE IS

LOGIC