Lecture 9: Noncausal Prediction in Gaussian Models

Chapter 4.1–4.2

Charles A. Bouman

Purdue University

Monday, September 14, 2026

Outline

1. **Data with no natural order**
   Roads, images, and data that do not lie on a grid at all.
2. **Noncausal Gaussian prediction**
   Predict from everything else; the prediction error and its variance.
3. **What changes without causality**
   Errors independent of predictors, but not of each other.
4. **From prediction to the precision matrix**
   Conditional densities, B, and the relation $B = \Gamma^{-1}(I - G)$.

Reading: Chapter 4.1–4.2, *Foundations of Computational Imaging*.

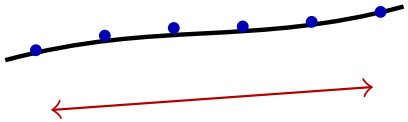
Today's Big Ideas

- With no causal order, predict each value from all the others.
- The noncausal error is still independent of its predictors.
- Errors are no longer independent of each other: whitening fails.
- The model becomes the precision matrix, which is the inverse of the covariance. It is built directly from the prediction coefficients and the prediction variances.

The zeros of the precision matrix say which values are not used to predict each other. Those zeros become the Gaussian Markov random field of the next lecture.

The trouble with causal ordering

One disadvantage of AR processes is that their construction depends on a causal ordering of points. That ordering is natural for real-time audio. For much else it is artificial.



temperatures along a road: which end is "the past"?

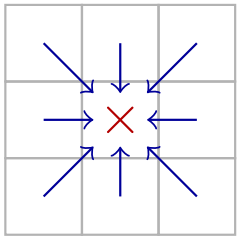
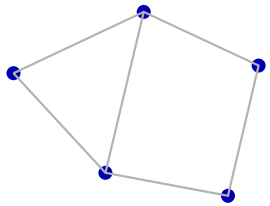


image pixels: neighbors on every side



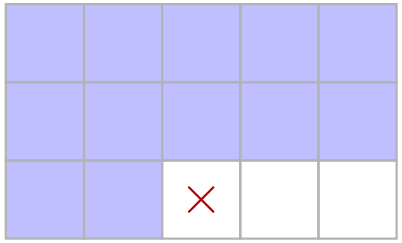
people in an organization: no grid, no order at all

The objective of this chapter

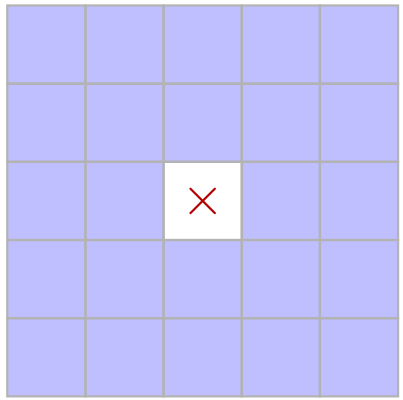
Remove causality from the modeling of images and other data. For images, this will help us avoid the raster-aligned **artifacts** that 1-D processing of ordered pixels is likely to produce.

The disadvantage of causal ordering in images

Causal prediction eliminates circular dependencies, and it gave us every result in Chapter 3. Consider what it does to an image.



causal image model



what image geometry suggests

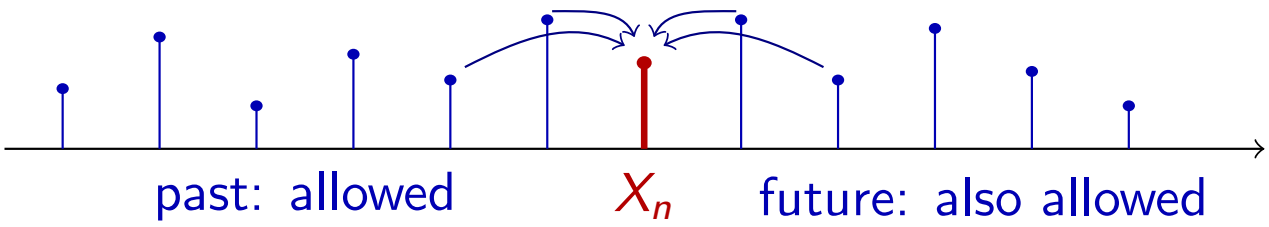
The image has no preferred direction, but the causal model does.

For some types of data, causal ordering is quite natural. For images, which are the primary subject of this book, it is not. Imposing a causal order can often produce artifacts in the results of processing. That is the reason for the noncausal models of this chapter.

The noncausal predictor

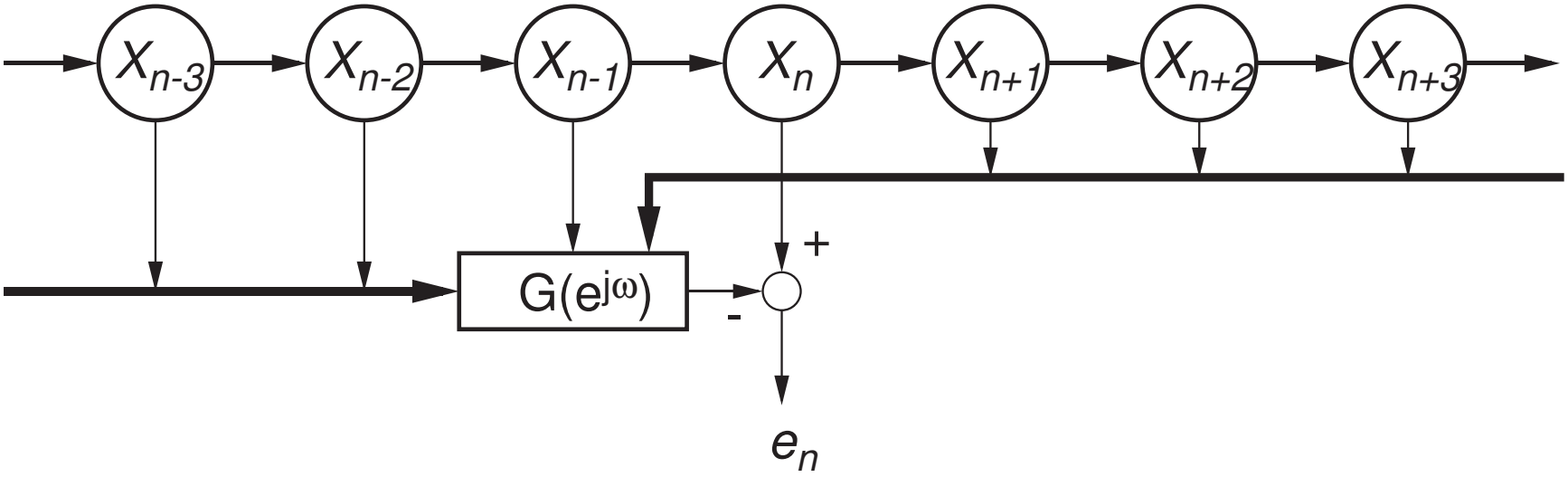
We start with 1-D signals. Let $X_1, \cdots, X_N$ again be a zero-mean discrete-time Gaussian random process. Rather than predicting from the past alone, we predict X_n from everything except X_n,

$$\hat{X}_n = \mathbb{E}\left[X_n | X_i \text{ for } i \neq n\right] \ .$$



- Lecture 6 conditioned on the past; now on everything else.

Diagram of the noncausal predictor



- The predictor uses a window of $n \pm P$ around the predicted point.
- Causal used the left side of the window; noncausal uses both sides.
- Errors $\mathcal{E}_n$ are not white, even when the predictor is optimal.

The error is linear, and $g_{n,n} = 0$

As with the causal predictor, the noncausal predictor is a linear function of the data when the random process is zero mean and Gaussian, by the property of Lecture 3. So the **noncausal prediction error** is

$$\mathcal{E}_n = X_n - \sum_{i=1}^{N} g_{n,i} X_i \ , \qquad \text{with } g_{n,n} = 0 \text{ for all } n \ .$$

G: zero diagonal

$$\begin{bmatrix} 0 & g & g & g & g \\ g & 0 & g & g & g \\ g & g & 0 & g & g \\ g & g & g & 0 & g \\ g & g & g & g & 0 \end{bmatrix}$$

the target cannot be one of its own predictors

The condition that $g_{n,n} = 0$ is very important. Otherwise the value X_n would be used to predict X_n perfectly. In the causal case the zero diagonal followed from H being lower triangular. Here it has to be imposed on G directly.

The noncausal prediction variance

The **noncausal prediction variance** is defined as

$$\sigma_n^2 \stackrel{\triangle}{=} \mathbb{E}\left[\mathcal{E}_n^2 | X_i \text{ for } i \neq n\right] = \mathbb{E}\left[\mathcal{E}_n^2\right] \ ,$$

where the second equality holds for the Gaussian case we consider.

- Linearity (Lecture 3): this variance does not depend on X_i, $i \neq n$.
- It may still depend on the time index n.

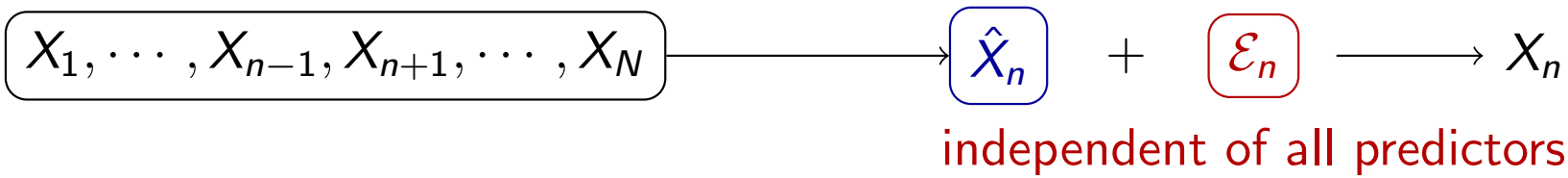
Property: independence from past and future

The noncausal prediction errors are independent of the values used in the prediction.

Independence of noncausal Gaussian prediction errors from past and future

The MMSE noncausal prediction errors for a zero-mean Gaussian random process are independent of the past and future of the random process:

$$\mathcal{E}_n \perp\!\!\!\perp \{X_i\}_{i \neq n} \ .$$



- Analogue of Lecture 6: the error is independent of every other value.

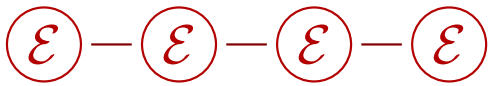
The prediction errors are no longer independent

One very important property is lost when we use noncausal prediction.

In general, the noncausal prediction errors are no longer uncorrelated and no longer independent of each other.



causal: independent



noncausal: correlated

Removing the causal ordering removes the whitening property.

- Independent prediction errors were essential to calculating the density $p(x)$.
- How will we calculate $p(x)$ for non-causal prediction?

The precision matrix

If the errors no longer factor, we start from the density itself. Since X is a zero-mean Gaussian random vector, it must have a density function of the form

$$p(x) = \frac{1}{(2\pi)^{N/2}} |B|^{1/2} \exp\left\{ -\frac{1}{2} x^t B x \right\} ,$$

where $B = R^{-1}$ is the inverse of the covariance matrix for X.

- B is the precision matrix; as it grows, the variance of X shrinks.
- R gives pairwise covariation; B gives how each value is predicted.

Two descriptions of the same Gaussian vector

Since X is zero-mean Gaussian, the conditional distribution of X_n given all the remaining X_i is

$$p(x_n|x_i \text{ for } i \neq n) = \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left\{ -\frac{1}{2\sigma_n^2} \left(x_n - \sum_{i=1}^{N} g_{n,i} x_i \right)^2 \right\} .$$

Global description
$p(x) \propto e^{-\frac{1}{2}x^t B x}$

?

Local prediction description
$p(x_n|x_{i\neq n})$: mean $\sum_i g_{n,i}x_i$, variance σ_n^2

Our objective

Determine the matrix B in terms of the noncausal prediction parameters σ_n^2 and $g_{n,i}$.

Differentiating the log likelihood

Differentiate with respect to x_n, and everything unrelated to x_n disappears:

$$\begin{aligned}
\frac{d}{dx_n} \log p(x) &= \frac{d}{dx_n} \log \left\{ p(x_n | x_i \text{ for } i \neq n)\, p(x_i \text{ for } i \neq n) \right\} \\
&= \frac{d}{dx_n} \log p(x_n | x_i \text{ for } i \neq n) + \frac{d}{dx_n} \log p(x_i \text{ for } i \neq n) \\
&= \frac{d}{dx_n} \log p(x_n | x_i \text{ for } i \neq n) \ .
\end{aligned}$$

- Line 1: the joint density factors as conditional times marginal.
- Line 2: the log of a product is a sum of logs.
- Line 3: the second term has no x_n, so its derivative is zero.

Evaluate both sides

Evaluating the two sides of this equality with the two density expressions yields the relationship

$$\sum_{i=1}^{N} B_{n,i} x_i = \frac{1}{\sigma_n^2} \left(x_n - \sum_{i=1}^{N} g_{n,i} x_i \right) .$$

- Left side: differentiating $-\frac{1}{2} x^t B x$ with respect to x_n gives $-\sum_{i=1}^{N} B_{n,i} x_i$.
- Right side: differentiating $-\frac{1}{2\sigma_n^2} \left(x_n - \sum_i g_{n,i} x_i \right)^2$ with respect to x_n gives $-\frac{1}{\sigma_n^2} \left(x_n - \sum_i g_{n,i} x_i \right)$, using $g_{n,n} = 0$.
- The minus signs cancel from both sides.

This equality holds for every x. What does that force the coefficients to be?

The precision matrix and the predictor determine each other

Matching coefficients term by term, the precision matrix is

$$B_{i,j} = \frac{1}{\sigma_i^2}\left(\delta_{i-j} - g_{i,j}\right) \ , \quad \text{that is,} \quad B_{i,i} = \frac{1}{\sigma_i^2} \ , \quad B_{i,j} = -\frac{g_{i,j}}{\sigma_i^2} \ \ (i \neq j) \ .$$

Alternatively, we can invert this relationship to compute the MMSE noncausal predictor parameters given a specification of B:

$$\begin{aligned} \sigma_n^2 &= (B_{n,n})^{-1} \ , \\ g_{n,i} &= \delta_{n-i} - \sigma_n^2 B_{n,i} \ . \end{aligned}$$

- Diagonal of B: inverse variances. Off-diagonal: scaled coefficients.
- Note $g_{n,n} = \delta_0 - \sigma_n^2 B_{n,n} = 1 - 1 = 0$, as required.

Matrix form

The previous slide compresses to one matrix equation. Define

- G is the **noncausal prediction matrix**, zero diagonal.
- $\Gamma = \text{diag}\left\{\sigma_1^2, \cdots, \sigma_N^2\right\}$, the diagonal matrix of noncausal prediction variances.
- $\mathcal{E}$, the column vector of noncausal prediction errors.

Then we have that $\mathcal{E} = (I - G)X$ and

$$B = \Gamma^{-1}(I - G)$$

or, alternatively,

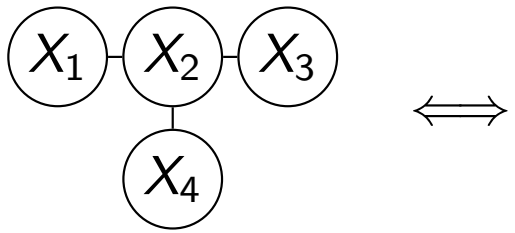
$$\begin{aligned} \Gamma &= \text{diag}(B)^{-1}\ , \\ G &= I - \Gamma B\ . \end{aligned}$$

Prediction parameters and the precision matrix

The noncausal prediction parameters (G, Γ) and the precision matrix B carry the same information. Given either one, the formulas above produce the other.

- In Lecture 6 the density came from a change of variables, and $R_X = (A^t \Lambda^{-1} A)^{-1}$. Here it comes directly from $B = \Gamma^{-1}(I - G)$. In both cases an accurate predictor is an accurate model.

The zeros of the precision matrix say which values are not used to predict each other.



$\Longleftrightarrow$

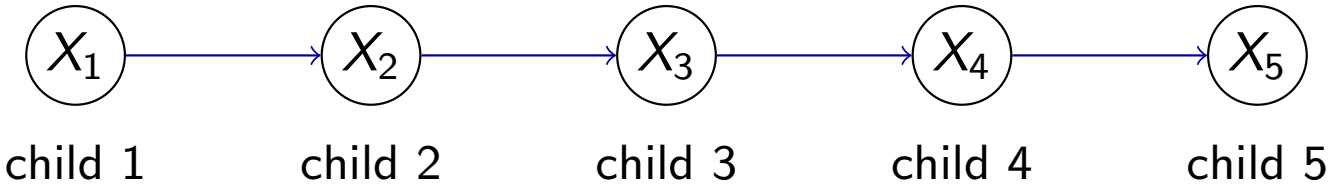
sparsity of B

$*$	$*$	0	0
$*$	$*$	$*$	$*$
0	$*$	$*$	0
0	$*$	0	$*$

- $B_{i,j} = 0$ exactly when $g_{i,j} = 0$, which happens when X_j is not used to predict X_i. That observation becomes the Gaussian Markov random field of Sections 4.3 to 4.6.

Conditional independence

In the **telephone game**, each child whispers to the next only what they just heard. This forms a homogeneous Markov chain.



Conditional independence

X and Z are *conditionally independent given* Y, written $X \perp\!\!\!\perp Z \mid Y$, when $p(x \mid y, z) = p(x \mid y)$. Once Y is known, Z says nothing more about X.

- For the chain, $X_1 \perp\!\!\!\perp X_3 \mid X_2$: given what child 2 heard, child 1 and child 3 say nothing more about each other. Yet child 1 and child 5 are still dependent, through the whole line.
- This is the zero pattern of B: for a chain B is tridiagonal, and $B_{i,j} = 0$ is exactly $X_i \perp\!\!\!\perp X_j$ given the rest.

Try it live: The Telephone Game thingy

Summary

- Noncausal prediction removes the arbitrary ordering. Each value is predicted from all the others, and the prediction error is $\mathcal{E}_n = X_n - \sum_i g_{n,i} X_i$ with $g_{n,n} = 0$.
- The prediction variance $\sigma_n^2 = \mathbb{E}\left[\mathcal{E}_n^2 | X_i \text{ for } i \neq n\right]$ does not depend on the conditioning values.
- Each error stays independent of the values used to predict it, $\mathcal{E}_n \perp\!\!\!\perp \{X_i\}_{i \neq n}$, but different errors are no longer independent of each other.
- The density is written with the precision matrix, $p(x) \propto \exp\{-\frac{1}{2} x^t B x\}$, where $B_{i,j} = \frac{1}{\sigma_i^2}(\delta_{i-j} - g_{i,j})$, or in matrix form $B = \Gamma^{-1}(I - G)$.
- The relationship runs both ways, with $\Gamma = \text{diag}(B)^{-1}$ and $G = I - \Gamma B$. Also $B_{i,j} = 0$ exactly when X_j is absent from the predictor of X_i.

Reading: Chapter 4.1–4.2. **Next lecture:** Chapter 4.3–4.4, Gaussian Markov random fields.