Probabilistic Graphical Models

Lecture 3: Gaussian Distributions

Volkan Cevher, Matthias Seeger
Ecole Polytechnique Fédérale de Lausanne

7/10/2011

Outline

Why Gaussians?

Gaussian (aka. normal) distribution $N(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$

$$N(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

- Marginalization, conditioning, linear transformation, posterior: All just linear algebra
- Incredible closedness properties:
  - Linear transformations
  - Conditioning
  - Marginalization

  Belief propagation needs such closedness
- Why all that?
  - Gaussians are limit distributions (central limit theorems)
  - Gaussians are maximum entropy distributions: No structure beyond mean, covariance

Gaussians are Limit Distributions

Central Limit Theorem

$\boldsymbol{x}_1 \sim P(\boldsymbol{x}_1)$, mean $\boldsymbol{\mu}$, covariance $\boldsymbol{\Sigma}$.

Imagine independent, indentically distributed (i.i.d.) replicas $\boldsymbol{x}_2, \boldsymbol{x}_3, \dots$

$$\bar{\boldsymbol{x}}^{(n)} := \sqrt{n} \underbrace{\left(n^{-1} \sum_{i=1}^{n} \boldsymbol{x}_i - \boldsymbol{\mu} \right)}_{\to 0 \text{ a.s.}} \quad \Rightarrow \; P(\bar{\boldsymbol{x}}^{(n)}) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\Sigma})$$

Gaussians are Limit Distributions

Central Limit Theorem

$\boldsymbol{x}_1 \sim P(\boldsymbol{x}_1)$, mean $\boldsymbol{\mu}$, covariance $\boldsymbol{\Sigma}$.

Imagine independent, indentically distributed (i.i.d.) replicas $\boldsymbol{x}_2, \boldsymbol{x}_3, \ldots$

$$\bar{\boldsymbol{x}}^{(n)} := \sqrt{n} \underbrace{\left(n^{-1} \sum_{i=1}^{n} \boldsymbol{x}_i - \boldsymbol{\mu} \right)}_{\to 0 \text{ a.s.}} \quad \Rightarrow \quad P(\bar{\boldsymbol{x}}^{(n)}) \xrightarrow{D} N(\boldsymbol{0}, \boldsymbol{\Sigma})$$

What does that mean?

- Averaging of i.i.d. variables: Mean, covariance retained
- Everything else smoothed away (by symmetry)
  $\Rightarrow$ What remains: Gaussian

Gaussians are Limit Distributions

Central Limit Theorem

$\boldsymbol{x}_1 \sim P(\boldsymbol{x}_1)$, mean $\boldsymbol{\mu}$, covariance $\boldsymbol{\Sigma}$.

Imagine independent, indentically distributed (i.i.d.) replicas $\boldsymbol{x}_2, \boldsymbol{x}_3, \ldots$

$$\bar{\boldsymbol{x}}^{(n)} := \sqrt{n}\underbrace{\left(n^{-1}\sum_{i=1}^{n}\boldsymbol{x}_i - \boldsymbol{\mu}\right)}_{\to 0 \text{ a.s.}} \quad \Rightarrow \quad P(\bar{\boldsymbol{x}}^{(n)}) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\Sigma})$$

What does that mean?

- Averaging of i.i.d. variables: Mean, covariance retained
- Everything else smoothed away (by symmetry)
  $\Rightarrow$ What remains: Gaussian

For the meticulous:

If $\boldsymbol{x}_1$ has no covariance, there are other stable distributions

Gaussians are Limit Distributions

Central Limit Theorem

$\boldsymbol{x}_1 \sim P(\boldsymbol{x}_1)$, mean $\boldsymbol{\mu}$, covariance $\boldsymbol{\Sigma}$.

Imagine independent, indentically distributed (i.i.d.) replicas $\boldsymbol{x}_2, \boldsymbol{x}_3, \dots$

$$\bar{\boldsymbol{x}}^{(n)} := \sqrt{n}\underbrace{\left(n^{-1}\sum_{i=1}^{n}\boldsymbol{x}_i - \boldsymbol{\mu}\right)}_{\to 0 \text{ a.s.}} \quad \Rightarrow \quad P(\bar{\boldsymbol{x}}^{(n)}) \stackrel{D}{\to} N(\boldsymbol{0}, \boldsymbol{\Sigma})$$

Implications for Statistics:

- Most models with finite number of parameters: Maximum likelihood estimator asymptotically normal

Gaussians are Limit Distributions

Central Limit Theorem

$\boldsymbol{x}_1 \sim P(\boldsymbol{x}_1)$, mean $\boldsymbol{\mu}$, covariance $\boldsymbol{\Sigma}$.

Imagine independent, indentically distributed (i.i.d.) replicas $\boldsymbol{x}_2, \boldsymbol{x}_3, \dots$

$$\bar{\boldsymbol{x}}^{(n)} := \sqrt{n} \underbrace{\left(n^{-1} \sum_{i=1}^{n} \boldsymbol{x}_i - \boldsymbol{\mu} \right)}_{\to 0 \text{ a.s.}} \quad \Rightarrow \quad P(\bar{\boldsymbol{x}}^{(n)}) \stackrel{D}{\to} N(\boldsymbol{0}, \boldsymbol{\Sigma})$$

Implications for closedness:

- Linear transformations, marginalization:
  Limit distributions have to be closed

Gaussians are Maximum Entropy Distributions

How much uncertainty / little structure is in a distribution?

Differential Entropy

$$\mathrm{H}[P] = \mathrm{E}_P[-\log P(\boldsymbol{x})] = \int P(\boldsymbol{x})(-\log P(\boldsymbol{x}))\, d\boldsymbol{x}$$

Gaussians are Maximum Entropy Distributions

How much uncertainty / little structure is in a distribution?
Differential Entropy

$$\mathrm{H}[P] = \mathrm{E}_P[-\log P(\boldsymbol{x})] = \int P(\boldsymbol{x})(-\log P(\boldsymbol{x}))\, d\boldsymbol{x}$$

Information theory (Shannon)

Immensely useful, basis of probabilistic machine learning.
Part II: Scratch surface. But dig for yourself:

- Cover, Thomas: Elements of Information Theory (1991)

One of my top five all times favourite textbooks. Read it!

Gaussians are Maximum Entropy Distributions

How much uncertainty / little structure is in a distribution?
Differential Entropy

$$\mathrm{H}[P] = \mathrm{E}_P[-\log P(\boldsymbol{x})] = \int P(\boldsymbol{x})(-\log P(\boldsymbol{x}))\, d\boldsymbol{x}$$

- Given mean $\boldsymbol{\mu}$, covariance $\boldsymbol{\Sigma}$: Maximum entropy distribution?

$$N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathrm{argmax}_P \left\{ \mathrm{H}[P] \;\middle|\; \mathrm{E}_P[\boldsymbol{x}] = \boldsymbol{\mu},\, \mathrm{Cov}_P[\boldsymbol{x}] = \boldsymbol{\Sigma} \right\}$$

Gaussians are Maximum Entropy Distributions

How much uncertainty / little structure is in a distribution?
Differential Entropy

$$\mathrm{H}[P] = \mathrm{E}_P[-\log P(\boldsymbol{x})] = \int P(\boldsymbol{x})(-\log P(\boldsymbol{x}))\, d\boldsymbol{x}$$

- Given mean $\boldsymbol{\mu}$, covariance Σ: Maximum entropy distribution?

$$N(\boldsymbol{\mu}, \Sigma) = \mathrm{argmax}_P \left\{ \mathrm{H}[P] \;\middle|\; \mathrm{E}_P[\boldsymbol{x}] = \boldsymbol{\mu},\, \mathrm{Cov}_P[\boldsymbol{x}] = \Sigma \right\}$$

- What does that mean?
  - Gaussian “nothing but mean and covariance”.
    Any other structure? It’s not a Gaussian
  - Would expect nice closedness properties for MaxEnt distributions
  - Upper bound on entropy:

$$\mathrm{H}[P] \le \mathrm{H}\left[N(\mathbf{0}, \mathrm{Cov}_P[\boldsymbol{x}])\right] = \frac{1}{2} \log |2\pi\, e\, \mathrm{Cov}_P[\boldsymbol{x}]|$$

Too Simple for Real World?

I want to model / learn structure.
Why should I care for an unstructured distribution?

Too Simple for Real World?

I want to model / learn structure.
Why should I care for an unstructured distribution?

Gaussians are elementary building blocks

- Gaussian + Structure (latent variables) $\rightarrow$ Wealth of models
  $\Rightarrow$ We'll see a few in what follows
- Many distributions are Gaussian scale mixtures [part II]
- Gaussian (implicitly) behind much of classical estimation methodology
- Carrier distribution for approximate inference [part II]

Too Simple for Real World?

I want to model / learn structure.
Why should I care for an unstructured distribution?

Gaussians are elementary building blocks

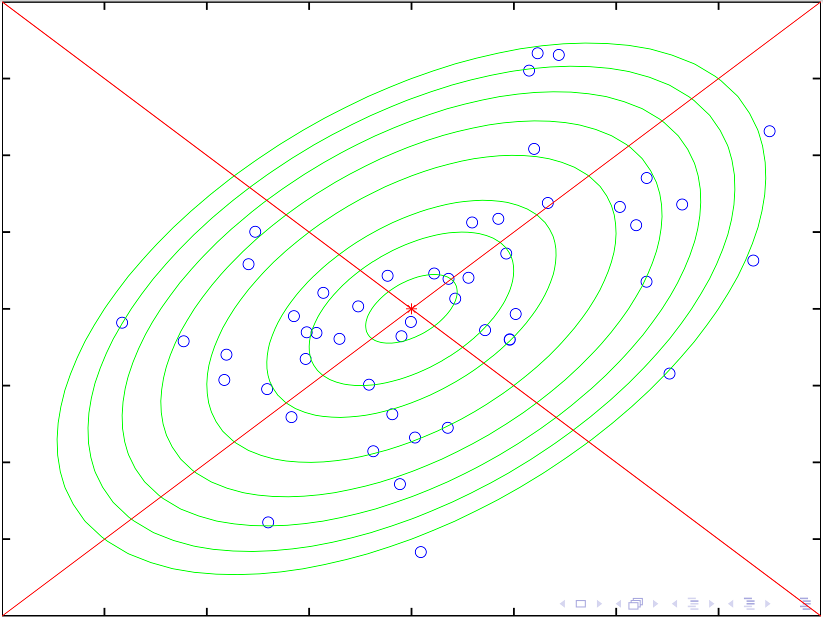
- Gaussian + Structure (latent variables) $\rightarrow$ Wealth of models
  $\Rightarrow$ We'll see a few in what follows
- Many distributions are Gaussian scale mixtures [part II]
- Gaussian (implicitly) behind much of classical estimation methodology
- Carrier distribution for approximate inference [part II]

Maximum entropy for general variables / moments?
$\Rightarrow$ Exponential families
Not in this lecture, but dig for yourself:

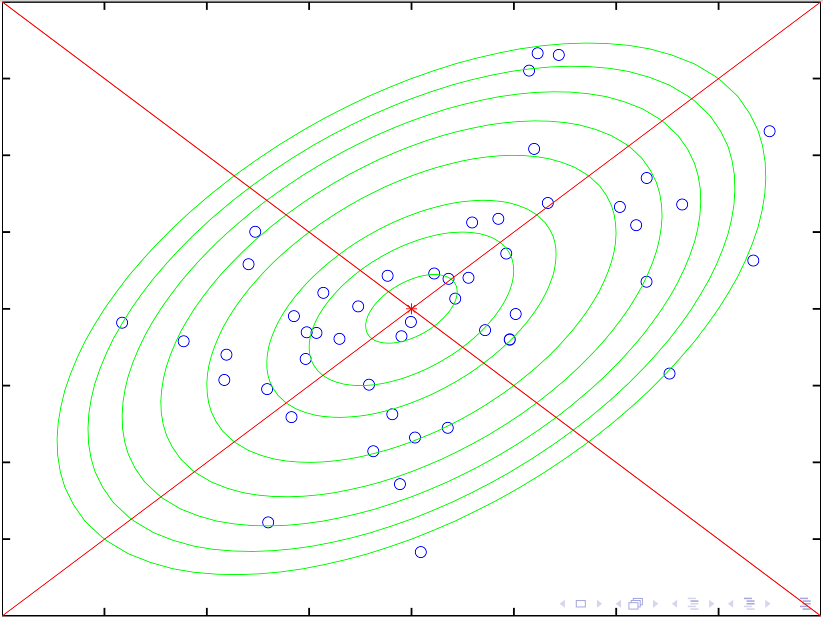
- M. Seeger: PhD thesis, Appendix A.4.1

http://people.mmci.uni-saarland.de/~mseeger/papers/thesis-appa.pdf

Gaussian Contours: Ellipsoids



Linear Transformations



Linear Transformations

- For any random variable $\boldsymbol{x}$ with covariance [expectation linear!]

$$
\begin{aligned}
\mathrm{E}[\boldsymbol{A}\boldsymbol{x}] &= \boldsymbol{A}\mathrm{E}[\boldsymbol{x}] \\
\mathrm{Cov}[\boldsymbol{A}\boldsymbol{x}] &= \mathrm{E}[\boldsymbol{A}\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{A}^T] - \mathrm{E}[\boldsymbol{A}\boldsymbol{x}]\mathrm{E}[\boldsymbol{A}\boldsymbol{x}]^T \\
&= \boldsymbol{A}\left(\mathrm{E}[\boldsymbol{x}\boldsymbol{x}^T] - \mathrm{E}[\boldsymbol{x}]\mathrm{E}[\boldsymbol{x}]^T\right)\boldsymbol{A}^T = \boldsymbol{A}\mathrm{Cov}[\boldsymbol{x}]\boldsymbol{A}^T
\end{aligned}
$$

Linear Transformations

- For any random variable $\boldsymbol{x}$ with covariance [expectation linear!]

$$\begin{aligned} \mathrm{E}[\boldsymbol{A}\boldsymbol{x}] &= \boldsymbol{A}\mathrm{E}[\boldsymbol{x}] \\ \mathrm{Cov}[\boldsymbol{A}\boldsymbol{x}] &= \mathrm{E}[\boldsymbol{A}\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{A}^T] - \mathrm{E}[\boldsymbol{A}\boldsymbol{x}]\mathrm{E}[\boldsymbol{A}\boldsymbol{x}]^T \\ &= \boldsymbol{A}\left(\mathrm{E}[\boldsymbol{x}\boldsymbol{x}^T] - \mathrm{E}[\boldsymbol{x}]\mathrm{E}[\boldsymbol{x}]^T\right)\boldsymbol{A}^T = \boldsymbol{A}\mathrm{Cov}[\boldsymbol{x}]\boldsymbol{A}^T \end{aligned}$$

- Gaussian is just mean and covariance

$$\boldsymbol{x} \sim N(\mu, \Sigma) \quad \Rightarrow \; \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b} \sim N(\boldsymbol{A}\mu + \boldsymbol{b}, \boldsymbol{A}\Sigma\boldsymbol{A}^T)$$

Linear Transformations

- For any random variable $\boldsymbol{x}$ with covariance [expectation linear!]

$$
\begin{aligned}
\mathrm{E}[\boldsymbol{A}\boldsymbol{x}] &= \boldsymbol{A}\mathrm{E}[\boldsymbol{x}] \\
\mathrm{Cov}[\boldsymbol{A}\boldsymbol{x}] &= \mathrm{E}[\boldsymbol{A}\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{A}^T] - \mathrm{E}[\boldsymbol{A}\boldsymbol{x}]\mathrm{E}[\boldsymbol{A}\boldsymbol{x}]^T \\
&= \boldsymbol{A}\left(\mathrm{E}[\boldsymbol{x}\boldsymbol{x}^T] - \mathrm{E}[\boldsymbol{x}]\mathrm{E}[\boldsymbol{x}]^T\right)\boldsymbol{A}^T = \boldsymbol{A}\mathrm{Cov}[\boldsymbol{x}]\boldsymbol{A}^T
\end{aligned}
$$

- Gaussian is just mean and covariance

$$
\boldsymbol{x} \sim N(\mu, \Sigma) \quad \Rightarrow \; \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b} \sim N(\boldsymbol{A}\mu + \boldsymbol{b}, \boldsymbol{A}\Sigma\boldsymbol{A}^T)
$$

[Missing here: Formal proof that $P(\boldsymbol{y})$ is Gaussian. $\Rightarrow$ Ask me offline]

Related Points

- Checking the normalization factor:
  $\boldsymbol{x} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$. $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$ eigendecomposition
  ($\boldsymbol{U}$ orthonormal (like rotation), $\boldsymbol{\Lambda}$ diagonal).
  $\boldsymbol{y} = \boldsymbol{U}^T\boldsymbol{x}$ (rotate eigenvectors $\to$ axes) $\Rightarrow d\boldsymbol{y} = d\boldsymbol{x}$

$$\int e^{-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}}\, d\boldsymbol{x} = \int e^{-\frac{1}{2}\boldsymbol{y}^T\boldsymbol{U}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{U}\boldsymbol{y}}\, d\boldsymbol{y} = \prod_i \int e^{-\frac{1}{2}y_i^2/\lambda_i}\, dy_i$$

$$= \prod_i (2\pi\lambda_i)^{1/2} = |2\pi\boldsymbol{\Sigma}|^{1/2} \text{ [determinant} = \prod \text{eigenvalues]}$$

  Recall: $\boldsymbol{U}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{U} = \boldsymbol{\Lambda}^{-1}$.

Related Points

- Checking the normalization factor:
  $\boldsymbol{x} \sim N(\mathbf{0}, \Sigma)$. $\Sigma = \boldsymbol{U}\Lambda\boldsymbol{U}^T$ eigendecomposition
  ($\boldsymbol{U}$ orthonormal (like rotation), Λ diagonal).
  $\boldsymbol{y} = \boldsymbol{U}^T\boldsymbol{x}$ (rotate eigenvectors $\rightarrow$ axes) $\Rightarrow d\boldsymbol{y} = d\boldsymbol{x}$

  $$\int e^{-\frac{1}{2}\boldsymbol{x}^T\Sigma^{-1}\boldsymbol{x}}\, d\boldsymbol{x} = \int e^{-\frac{1}{2}\boldsymbol{y}^T\boldsymbol{U}^T\Sigma^{-1}\boldsymbol{U}\boldsymbol{y}}\, d\boldsymbol{y} = \prod_i \int e^{-\frac{1}{2}y_i^2/\lambda_i}\, dy_i$$
  $$= \prod_i (2\pi\lambda_i)^{1/2} = |2\pi\Sigma|^{1/2} \text{ [determinant} = \prod \text{eigenvalues]}$$

  Recall: $\boldsymbol{U}^T\Sigma^{-1}\boldsymbol{U} = \Lambda^{-1}$.
- For Gaussian:

  Σ diagonal $\Rightarrow P(\boldsymbol{x}) = \prod_i P(x_i)$
  Uncorrelated components $\Rightarrow$ Independent components

Marginal Distribution

- $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \Sigma)$.
  $I \subset \{1, \dots, n\}$. $\boldsymbol{x}_I := (x_i)_{i \in I}$.
  Prize question: What is $P(\boldsymbol{x}_I)$?

Marginal Distribution

- $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
  $I \subset \{1, \dots, n\}$. $\boldsymbol{x}_I := (x_i)_{i \in I}$.
  Prize question: What is $P(\boldsymbol{x}_I)$?
- Pick selection matrix $\boldsymbol{I}_{I,\cdot} \Rightarrow \boldsymbol{x}_I = \boldsymbol{I}_{I,\cdot}\boldsymbol{x}$

$$P(\boldsymbol{x}_I) = N(\boldsymbol{I}_{I,\cdot}\boldsymbol{\mu}, \boldsymbol{I}_{I,\cdot}\boldsymbol{\Sigma}\boldsymbol{I}_{\cdot,I}) = N(\boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I)$$

Marginal Distribution

- $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
  $I \subset \{1, \ldots, n\}$. $\boldsymbol{x}_I := (x_i)_{i \in I}$.
  Prize question: What is $P(\boldsymbol{x}_I)$?
- Pick selection matrix $\boldsymbol{I}_{I,\cdot} \Rightarrow \boldsymbol{x}_I = \boldsymbol{I}_{I,\cdot}\boldsymbol{x}$

$$P(\boldsymbol{x}_I) = N(\boldsymbol{I}_{I,\cdot}\boldsymbol{\mu}, \boldsymbol{I}_{I,\cdot}\boldsymbol{\Sigma}\boldsymbol{I}_{\cdot,I}) = N(\boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I)$$

- Marginalization (linear transformations):
  Very simple if you have $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$

Conditioning would be easy if . . .

- $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
  $I \subset \{1, \dots, n\}$. $R = \{1, \dots, n\} \setminus I$.
  Next prize question: What is $P(\boldsymbol{x}_I | \boldsymbol{x}_R)$?

Conditioning would be easy if . . .

- $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
  $I \subset \{1, \dots, n\}$. $R = \{1, \dots, n\} \setminus I$.
  Next prize question: What is $P(\boldsymbol{x}_I | \boldsymbol{x}_R)$?
- Not so simple. But it would be if . . .

Conditioning would be easy if . . .

- $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \Sigma)$.
  $I \subset \{1, \ldots, n\}$. $R = \{1, \ldots, n\} \setminus I$.
  Next prize question: What is $P(\boldsymbol{x}_I | \boldsymbol{x}_R)$?
- Not so simple. But it would be if . . .

$$P(\boldsymbol{x}_I|\boldsymbol{x}_R) \propto e^{-\frac{1}{2}((\boldsymbol{x}_I-\boldsymbol{\mu}_I)^T \boldsymbol{A}_I(\boldsymbol{x}_I-\boldsymbol{\mu}_I)+2(\boldsymbol{x}_R-\boldsymbol{\mu}_R)^T \boldsymbol{A}_{I,R}^T(\boldsymbol{x}_I-\boldsymbol{\mu}_I))},$$
$$\boldsymbol{A} = \Sigma^{-1}$$

Natural and Moment Parameterization

Two ways of parameterizing a Gaussian. You know:

Gaussian in moment (aka. mean) parameters μ, Σ

$$\propto e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$$

$$\propto e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$$

Natural and Moment Parameterization

Two ways of parameterizing a Gaussian. You know:

Gaussian in moment (aka. mean) parameters μ, Σ

$$\propto e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$$

$$\propto e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{\mu})}, \quad \boldsymbol{A} = \boldsymbol{\Sigma}^{-1}$$

Natural and Moment Parameterization

Two ways of parameterizing a Gaussian. You know:

Gaussian in moment (aka. mean) parameters μ, Σ

$$\propto e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$$

$$\propto e^{-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}+(\boldsymbol{A}\boldsymbol{\mu})^T\boldsymbol{x}}, \quad \boldsymbol{A} = \Sigma^{-1}$$

Natural and Moment Parameterization

Two ways of parameterizing a Gaussian. You know:

Gaussian in moment (aka. mean) parameters μ, Σ

$$\propto e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$$

$$\propto e^{-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}+\boldsymbol{r}^T\boldsymbol{x}}, \quad \boldsymbol{A} = \boldsymbol{\Sigma}^{-1},\ \boldsymbol{r} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$$

Natural and Moment Parameterization

Two ways of parameterizing a Gaussian. You know:

Gaussian in moment (aka. mean) parameters μ, Σ

$$\propto e^{-\frac{1}{2}(\boldsymbol{x}-\mu)^T \Sigma^{-1} (\boldsymbol{x}-\mu)}$$

Now you know:

Gaussian in natural (aka. canonical) parameters $\boldsymbol{r}, \boldsymbol{A}$

$$\propto e^{-\frac{1}{2}\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} + \boldsymbol{r}^T \boldsymbol{x}}, \quad \boldsymbol{A} = \Sigma^{-1}, \ \boldsymbol{r} = \Sigma^{-1} \mu$$

Natural and Moment Parameterization

Gaussian in moment (aka. mean) parameters $\boldsymbol{\mu}, \boldsymbol{\Sigma}$

$$\propto e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$$

Gaussian in natural (aka. canonical) parameters $\boldsymbol{r}, \boldsymbol{A}$

$$\propto e^{-\frac{1}{2}\boldsymbol{x}^T \boldsymbol{A}\boldsymbol{x} + \boldsymbol{r}^T\boldsymbol{x}}, \quad \boldsymbol{A} = \boldsymbol{\Sigma}^{-1}, \ \boldsymbol{r} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$$

- Why two parameterizations for the same thing?
  - Some things simple in moment parameters:
    Linear transforms, marginalization [everything “sum”]
  - Some things simple in natural parameters:
    Conditioning, density product [everything “product”]

Natural and Moment Parameterization

Gaussian in moment (aka. mean) parameters μ, Σ

$$\propto e^{-\frac{1}{2}(\boldsymbol{x}-\mu)^T\Sigma^{-1}(\boldsymbol{x}-\mu)}$$

Gaussian in natural (aka. canonical) parameters $\boldsymbol{r}, \boldsymbol{A}$

$$\propto e^{-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}+\boldsymbol{r}^T\boldsymbol{x}}, \quad \boldsymbol{A} = \Sigma^{-1}, \ \boldsymbol{r} = \Sigma^{-1}\mu$$

- Why two parameterizations for the same thing?
  - Some things simple in moment parameters:
    Linear transforms, marginalization [everything “sum”]
  - Some things simple in natural parameters:
    Conditioning, density product [everything “product”]
- For belief propagation (sum-product): Conversions all the time
- Conversion $\leftrightarrow$ Matrix inversion
  $\Rightarrow$ Makes Gaussian propagation numerically difficult

Conditional Distribution

- $P(\boldsymbol{x}_I|\boldsymbol{x}_R)$: What does it mean? Factorization
  1. Sample $\boldsymbol{x}_R \sim N(\boldsymbol{\mu}_R, \Sigma_R)$
  2. Sample $\boldsymbol{x}_I$ from Gaussian depending on $\boldsymbol{x}_R$
- $\mathrm{E}[\boldsymbol{x}] = \boldsymbol{\mu}$, $\mathrm{Cov}[\boldsymbol{x}] = \Sigma$ afterwards?
  $\Rightarrow$ Rule (2) must be $P(\boldsymbol{x}_I|\boldsymbol{x}_R)$!

For meticulous: We already know that $P(\boldsymbol{x}_I|\boldsymbol{x}_R)$ is Gaussian (by inspection)

Conditional Distribution

- $P(\boldsymbol{x}_I|\boldsymbol{x}_R)$: What does it mean? Factorization
  1. Sample $\boldsymbol{x}_R \sim N(\boldsymbol{\mu}_R, \Sigma_R)$
  2. Sample $\boldsymbol{x}_I$ from Gaussian depending on $\boldsymbol{x}_R$
- $\mathrm{E}[\boldsymbol{x}] = \boldsymbol{\mu}$, $\mathrm{Cov}[\boldsymbol{x}] = \Sigma$ afterwards?
  $\Rightarrow$ Rule (2) must be $P(\boldsymbol{x}_I|\boldsymbol{x}_R)$!

For meticulous: We already know that $P(\boldsymbol{x}_I|\boldsymbol{x}_R)$ is Gaussian (by inspection)

- Ansatz: $\boldsymbol{y} = \boldsymbol{x} - \boldsymbol{\mu}$.
  $\boldsymbol{y}_I = \boldsymbol{u} + \boldsymbol{B}\boldsymbol{y}_R$, $\boldsymbol{u} \sim N(\boldsymbol{0}, \boldsymbol{C})$.

Conditional Distribution

- $P(\boldsymbol{x}_I|\boldsymbol{x}_R)$: What does it mean? Factorization
  1. Sample $\boldsymbol{x}_R \sim N(\boldsymbol{\mu}_R, \Sigma_R)$
  2. Sample $\boldsymbol{x}_I$ from Gaussian depending on $\boldsymbol{x}_R$
- $\mathrm{E}[\boldsymbol{x}] = \boldsymbol{\mu}$, $\mathrm{Cov}[\boldsymbol{x}] = \Sigma$ afterwards?
  $\Rightarrow$ Rule (2) must be $P(\boldsymbol{x}_I|\boldsymbol{x}_R)$!

For meticulous: We already know that $P(\boldsymbol{x}_I|\boldsymbol{x}_R)$ is Gaussian (by inspection)

- Ansatz: $\boldsymbol{y} = \boldsymbol{x} - \boldsymbol{\mu}$.
  $\boldsymbol{y}_I = \boldsymbol{u} + \boldsymbol{B}\boldsymbol{y}_R$, $\boldsymbol{u} \sim N(\boldsymbol{0}, \boldsymbol{C})$.
  Schur complement: $\boldsymbol{C} = \Sigma/\Sigma_R := \Sigma_I - \Sigma_{I,R}\Sigma_R^{-1}\Sigma_{R,I}$

$$\mathrm{E}[\boldsymbol{x}_I|\boldsymbol{x}_R] = \boldsymbol{\mu}_I + \Sigma_{I,R}\Sigma_R^{-1}(\boldsymbol{x}_R - \boldsymbol{\mu}_R),$$
$$\mathrm{Cov}[\boldsymbol{x}_I|\boldsymbol{x}_R] = \Sigma/\Sigma_R = \Sigma_I - \Sigma_{I,R}\Sigma_R^{-1}\Sigma_{R,I}$$

- $\mathrm{E}[\boldsymbol{x}_I|\boldsymbol{x}_R]$ linear in $\boldsymbol{x}_R$. $\mathrm{Cov}[\boldsymbol{x}_I|\boldsymbol{x}_R]$ independent of $\boldsymbol{x}_R$

The Schur Complement

$$\underbrace{P(\boldsymbol{x}_I, \boldsymbol{x}_R)}_{\mathrm{Cov}[\cdot]=\boldsymbol{\Sigma}} = \underbrace{P(\boldsymbol{x}_I | \boldsymbol{x}_R)}_{\mathrm{Cov}[\cdot]=\boldsymbol{\Sigma}/\boldsymbol{\Sigma}_R} \times \underbrace{P(\boldsymbol{x}_R)}_{\mathrm{Cov}[\cdot]=\boldsymbol{\Sigma}_R}$$

The Schur Complement

$$\underbrace{P(\boldsymbol{x}_I, \boldsymbol{x}_R)}_{\mathrm{Cov}[\cdot]=\boldsymbol{\Sigma}} = \underbrace{P(\boldsymbol{x}_I | \boldsymbol{x}_R)}_{\mathrm{Cov}[\cdot]=\boldsymbol{\Sigma}/\boldsymbol{\Sigma}_R} \times \underbrace{P(\boldsymbol{x}_R)}_{\mathrm{Cov}[\cdot]=\boldsymbol{\Sigma}_R}$$

- Holds more generally, whenever $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}_R$ nonsingular. Not just for symmetric $\boldsymbol{\Sigma}$
- Determinant identity

$$|\boldsymbol{\Sigma}| = |\boldsymbol{\Sigma}/\boldsymbol{\Sigma}_R| \cdot |\boldsymbol{\Sigma}_R|$$

The Schur Complement

$$\underbrace{P(\boldsymbol{x}_I, \boldsymbol{x}_R)}_{\mathrm{Cov}[\cdot]=\boldsymbol{\Sigma}} = \underbrace{P(\boldsymbol{x}_I | \boldsymbol{x}_R)}_{\mathrm{Cov}[\cdot]=\boldsymbol{\Sigma}/\boldsymbol{\Sigma}_R} \times \underbrace{P(\boldsymbol{x}_R)}_{\mathrm{Cov}[\cdot]=\boldsymbol{\Sigma}_R}$$

- Holds more generally, whenever $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}_R$ nonsingular. Not just for symmetric $\boldsymbol{\Sigma}$
- Determinant identity

$$|\boldsymbol{\Sigma}| = |\boldsymbol{\Sigma}/\boldsymbol{\Sigma}_R| \cdot |\boldsymbol{\Sigma}_R|$$

Useful special case:

$$|\boldsymbol{I} + \boldsymbol{U}\boldsymbol{V}| = |\boldsymbol{I} + \boldsymbol{V}\boldsymbol{U}|$$

Partitioned Inverse Equations

$$\Sigma^{-1} = \begin{bmatrix} \boldsymbol{A}_I & \boldsymbol{A}_{I,R} \\ \boldsymbol{A}_{R,I} & \boldsymbol{A}_R \end{bmatrix} = \begin{bmatrix} (\Sigma/\Sigma_R)^{-1} & -(\Sigma/\Sigma_R)^{-1}\boldsymbol{B} \\ -\boldsymbol{B}^T(\Sigma/\Sigma_R)^{-1} & \Sigma_R^{-1} + \boldsymbol{B}^T(\Sigma/\Sigma_R)^{-1}\boldsymbol{B} \end{bmatrix}$$

$$\boldsymbol{B} = \Sigma_{I,R}\Sigma_R^{-1}$$

- Very useful if $|I|$, $|R|$ different
  $\Rightarrow$ Do inverses in smaller of them only!

Partitioned Inverse Equations

$$\Sigma^{-1} = \begin{bmatrix} \boldsymbol{A}_I & \boldsymbol{A}_{I,R} \\ \boldsymbol{A}_{R,I} & \boldsymbol{A}_R \end{bmatrix} = \begin{bmatrix} (\Sigma/\Sigma_R)^{-1} & -(\Sigma/\Sigma_R)^{-1}\boldsymbol{B} \\ -\boldsymbol{B}^T(\Sigma/\Sigma_R)^{-1} & \Sigma_R^{-1} + \boldsymbol{B}^T(\Sigma/\Sigma_R)^{-1}\boldsymbol{B} \end{bmatrix}$$

$$\boldsymbol{B} = \Sigma_{I,R}\Sigma_R^{-1}$$

- Very useful if $|I|$, $|R|$ different
  $\Rightarrow$ Do inverses in smaller of them only!
- Could have conditioned on $\boldsymbol{x}_I$ just as well:
  Woodbury formula: $(\Sigma/\Sigma_I)^{-1} = \Sigma_R^{-1} + \boldsymbol{B}^T(\Sigma/\Sigma_R)^{-1}\boldsymbol{B}$

$$(\boldsymbol{E} + \boldsymbol{F}\boldsymbol{G}^{-1}\boldsymbol{H})^{-1} = \boldsymbol{E}^{-1} - \boldsymbol{E}^{-1}\boldsymbol{F}(\boldsymbol{G} + \boldsymbol{H}\boldsymbol{E}^{-1}\boldsymbol{F})^{-1}\boldsymbol{H}\boldsymbol{E}^{-1}$$

$\Rightarrow$ Not least formula to learn by heart

Product of Gaussians

$$N(\boldsymbol{x}|\boldsymbol{\mu}_1, \Sigma_1)N(\boldsymbol{x}|\boldsymbol{\mu}_2, \Sigma_2) = N(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma)C$$

- Product: Combination of messages / information

Product of Gaussians

$$N(\boldsymbol{x}|\mu_1, \Sigma_1)N(\boldsymbol{x}|\mu_2, \Sigma_2) = N(\boldsymbol{x}|\mu, \Sigma)C$$

- Product: Combination of messages / information
- Easy in natural parameters:

$$e^{\boldsymbol{r}_1^T\boldsymbol{x} - \frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}_1\boldsymbol{x}} \times e^{\boldsymbol{r}_2^T\boldsymbol{x} - \frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}_2\boldsymbol{x}} = e^{(\boldsymbol{r}_1+\boldsymbol{r}_2)^T\boldsymbol{x} - \frac{1}{2}\boldsymbol{x}^T(\boldsymbol{A}_1+\boldsymbol{A}_2)\boldsymbol{x}}$$

$\Rightarrow$ Sum of natural parameters

$$\boldsymbol{A} = \boldsymbol{A}_1 + \boldsymbol{A}_2, \quad \boldsymbol{r} = \boldsymbol{r}_1 + \boldsymbol{r}_2$$

Product of Gaussians

$$N(\boldsymbol{x}|\mu_1, \Sigma_1)N(\boldsymbol{x}|\mu_2, \Sigma_2) = N(\boldsymbol{x}|\mu, \Sigma)C$$

- Product: Combination of messages / information
- Easy in natural parameters:

$$e^{\boldsymbol{r}_1^T\boldsymbol{x}-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}_1\boldsymbol{x}} \times e^{\boldsymbol{r}_2^T\boldsymbol{x}-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}_2\boldsymbol{x}} = e^{(\boldsymbol{r}_1+\boldsymbol{r}_2)^T\boldsymbol{x}-\frac{1}{2}\boldsymbol{x}^T(\boldsymbol{A}_1+\boldsymbol{A}_2)\boldsymbol{x}}$$

$\Rightarrow$ Sum of natural parameters

$$\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\Sigma}_2^{-1}, \quad \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} = \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2$$

Product of Gaussians

$$N(\boldsymbol{x}|\mu_1, \Sigma_1)N(\boldsymbol{x}|\mu_2, \Sigma_2) = N(\boldsymbol{x}|\mu, \Sigma)C$$

- Product: Combination of messages / information
- Easy in natural parameters:

$$e^{\boldsymbol{r}_1^T \boldsymbol{x} - \frac{1}{2}\boldsymbol{x}^T \boldsymbol{A}_1 \boldsymbol{x}} \times e^{\boldsymbol{r}_2^T \boldsymbol{x} - \frac{1}{2}\boldsymbol{x}^T \boldsymbol{A}_2 \boldsymbol{x}} = e^{(\boldsymbol{r}_1 + \boldsymbol{r}_2)^T \boldsymbol{x} - \frac{1}{2}\boldsymbol{x}^T (\boldsymbol{A}_1 + \boldsymbol{A}_2)\boldsymbol{x}}$$

$\Rightarrow$ Sum of natural parameters

$$\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\Sigma}_2^{-1})^{-1}, \quad \boldsymbol{\mu} = \underbrace{\boldsymbol{\Sigma}(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2)}_{\text{"weighted avg."}}$$

Product of Gaussians

$$N(\boldsymbol{x}|\boldsymbol{\mu}_1, \Sigma_1)N(\boldsymbol{x}|\boldsymbol{\mu}_2, \Sigma_2) = N(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma)C$$

- Product: Combination of messages / information
- Easy in natural parameters:

$$e^{\boldsymbol{r}_1^T\boldsymbol{x}-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}_1\boldsymbol{x}} \times e^{\boldsymbol{r}_2^T\boldsymbol{x}-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}_2\boldsymbol{x}} = e^{(\boldsymbol{r}_1+\boldsymbol{r}_2)^T\boldsymbol{x}-\frac{1}{2}\boldsymbol{x}^T(\boldsymbol{A}_1+\boldsymbol{A}_2)\boldsymbol{x}}$$

$\Rightarrow$ Sum of natural parameters

$$\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\Sigma}_2^{-1})^{-1}, \quad \boldsymbol{\mu} = \underbrace{\boldsymbol{\Sigma}(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2)}_{\text{"weighted avg."}}$$

- And C? Often not needed. If you need it: Sampling argument (saves pages of algebra)

Linear-Gaussian Model

$$\begin{aligned} \boldsymbol{u} &\sim N(\mu_0, \Sigma_0) && \text{Prior} \\ \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{u} + \varepsilon,\ \varepsilon \sim N(\boldsymbol{0}, \Psi) && \text{Likelihood} \end{aligned}$$

1. Joint / marginal distribution: Tower formulae

$$\mathrm{E}[\boldsymbol{y}] = \mathrm{E}[\mathrm{E}[\boldsymbol{y}|\boldsymbol{u}]], \quad \mathrm{Cov}[\boldsymbol{y}] = \mathrm{Cov}[\mathrm{E}[\boldsymbol{y}|\boldsymbol{u}]] + \mathrm{E}[\mathrm{Cov}[\boldsymbol{y}|\boldsymbol{u}]]$$

$$\mathrm{Cov}[\boldsymbol{u}, \boldsymbol{y}] = \mathrm{Cov}[\boldsymbol{u}, \mathrm{E}[\boldsymbol{y}|\boldsymbol{u}]] + \mathrm{E}[\mathrm{Cov}[\boldsymbol{u}, \boldsymbol{y}|\boldsymbol{u}]]$$

Linear-Gaussian Model

$$\begin{aligned} \boldsymbol{u} &\sim N(\mu_0, \Sigma_0) && \text{Prior} \\ \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{u} + \varepsilon,\ \varepsilon \sim N(\mathbf{0}, \Psi) && \text{Likelihood} \end{aligned}$$

1. Joint / marginal distribution: Tower formulae

$$\mathrm{E}[\boldsymbol{y}] = \mathrm{E}[\mathrm{E}[\boldsymbol{y}|\boldsymbol{u}]], \quad \mathrm{Cov}[\boldsymbol{y}] = \mathrm{Cov}[\mathrm{E}[\boldsymbol{y}|\boldsymbol{u}]] + \mathrm{E}[\mathrm{Cov}[\boldsymbol{y}|\boldsymbol{u}]]$$
$$\mathrm{Cov}[\boldsymbol{u}, \boldsymbol{y}] = \mathrm{Cov}[\boldsymbol{u}, \mathrm{E}[\boldsymbol{y}|\boldsymbol{u}]] + \mathrm{E}[\mathrm{Cov}[\boldsymbol{u}, \boldsymbol{y}|\boldsymbol{u}]]$$

2. Posterior: Product Prior $\times$ Likelihood

$$\exp\left(-\frac{1}{2}[(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{u})^T \Psi^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{u}) + (\boldsymbol{u} - \mu_0)^T \Sigma_0^{-1} (\boldsymbol{u} - \mu_0)]\right)$$

Linear-Gaussian Model

$$\begin{aligned} \boldsymbol{u} &\sim N(\mu_0, \Sigma_0) && \text{Prior} \\ \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{u} + \varepsilon,\ \varepsilon \sim N(\mathbf{0}, \Psi) && \text{Likelihood} \end{aligned}$$

1. Joint / marginal distribution: Tower formulae

$$\mathrm{E}[\boldsymbol{y}] = \mathrm{E}[\mathrm{E}[\boldsymbol{y}|\boldsymbol{u}]], \quad \mathrm{Cov}[\boldsymbol{y}] = \mathrm{Cov}[\mathrm{E}[\boldsymbol{y}|\boldsymbol{u}]] + \mathrm{E}[\mathrm{Cov}[\boldsymbol{y}|\boldsymbol{u}]]$$
$$\mathrm{Cov}[\boldsymbol{u}, \boldsymbol{y}] = \mathrm{Cov}[\boldsymbol{u}, \mathrm{E}[\boldsymbol{y}|\boldsymbol{u}]] + \mathrm{E}[\mathrm{Cov}[\boldsymbol{u}, \boldsymbol{y}|\boldsymbol{u}]]$$

2. Posterior: Product Prior $\times$ Likelihood

$$\exp\left(-\frac{1}{2}[\boldsymbol{u}^T \underbrace{(\boldsymbol{X}^T\Psi^{-1}\boldsymbol{X} + \Sigma_0^{-1})}_{\mathrm{Cov}[\boldsymbol{u}|\boldsymbol{y}]^{-1}} \boldsymbol{u} - 2\boldsymbol{u}^T \underbrace{(\boldsymbol{X}^T\Psi^{-1}\boldsymbol{y} + \Sigma_0^{-1}\mu_0)}_{\mathrm{Cov}[\boldsymbol{u}|\boldsymbol{y}]^{-1}\mathrm{E}[\boldsymbol{u}|\boldsymbol{y}]} + \dots]\right)$$

Normal equations:
$$\mathrm{E}[\boldsymbol{u}|\boldsymbol{y}] = (\boldsymbol{X}^T\Psi^{-1}\boldsymbol{X} + \Sigma_0^{-1})^{-1}(\boldsymbol{X}^T\Psi^{-1}\boldsymbol{y} + \Sigma_0^{-1}\mu_0)$$

Linear-Gaussian Model

$$\begin{aligned} \boldsymbol{u} &\sim N(\mu_0, \Sigma_0) && \text{Prior} \\ \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{u} + \varepsilon,\ \varepsilon \sim N(\mathbf{0}, \Psi) && \text{Likelihood} \end{aligned}$$

2. Posterior: Product Prior × Likelihood

$$\exp\left(-\frac{1}{2}[\boldsymbol{u}^T \underbrace{(\boldsymbol{X}^T\Psi^{-1}\boldsymbol{X} + \Sigma_0^{-1})}_{\mathrm{Cov}[\boldsymbol{u}|\boldsymbol{y}]^{-1}} \boldsymbol{u} - 2\boldsymbol{u}^T \underbrace{(\boldsymbol{X}^T\Psi^{-1}\boldsymbol{y} + \Sigma_0^{-1}\mu_0)}_{\mathrm{Cov}[\boldsymbol{u}|\boldsymbol{y}]^{-1}\mathrm{E}[\boldsymbol{u}|\boldsymbol{y}]} + \dots]\right)$$

What if $\boldsymbol{y}$ less coefficients than $\boldsymbol{u}$?

Linear-Gaussian Model

$$\boldsymbol{u} \sim N(\mu_0, \Sigma_0) \quad \text{Prior}$$
$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{u} + \varepsilon,\ \varepsilon \sim N(\boldsymbol{0}, \Psi) \quad \text{Likelihood}$$

2. Posterior: Product Prior $\times$ Likelihood

$$\exp\Big(-\frac{1}{2}[\boldsymbol{u}^T \underbrace{(\boldsymbol{X}^T\Psi^{-1}\boldsymbol{X} + \Sigma_0^{-1})}_{\mathrm{Cov}[\boldsymbol{u}|\boldsymbol{y}]^{-1}} \boldsymbol{u} - 2\boldsymbol{u}^T \underbrace{(\boldsymbol{X}^T\Psi^{-1}\boldsymbol{y} + \Sigma_0^{-1}\mu_0)}_{\mathrm{Cov}[\boldsymbol{u}|\boldsymbol{y}]^{-1}\mathrm{E}[\boldsymbol{u}|\boldsymbol{y}]} + \dots]\Big)$$

What if $\boldsymbol{y}$ less coefficients than $\boldsymbol{u}$?

$$\mathrm{Cov}[\boldsymbol{u}|\boldsymbol{y}] = \mathrm{Cov}[(\boldsymbol{u}\,\boldsymbol{y})]/\mathrm{Cov}[\boldsymbol{y}] = \Sigma_0 - \Sigma_0\boldsymbol{X}^T(\Psi + \boldsymbol{X}\Sigma_0\boldsymbol{X}^T)^{-1}\boldsymbol{X}\Sigma_0,$$
$$\mathrm{E}[\boldsymbol{u}|\boldsymbol{y}] = \mathrm{E}[\boldsymbol{u}] + \mathrm{Cov}[\boldsymbol{u},\boldsymbol{y}]\mathrm{Cov}[\boldsymbol{y}]^{-1}(\boldsymbol{y} - \mathrm{E}[\boldsymbol{y}])$$
$$= \mu_0 + \Sigma_0\boldsymbol{X}^T(\Psi + \boldsymbol{X}\Sigma_0\boldsymbol{X}^T)^{-1}(\boldsymbol{y} - \boldsymbol{X}\mu_0)$$

Wrap-Up

Practice those Gaussian calculations

- They come back at you all the time
- They look messy only as long as you don't understand them
- Short derivations take much less time (waste it with funnier things)
- Short derivations contain fewer mistakes
- Short derivations are just so much cooler!